University of Alberta

Responsiveness of Generic Health Status Measures in Stroke

By

Alan Simon Pickard ©

A thesis submitted to the Faculty of Graduate Studies and Research in the partial fulfillment of the requirements for the degree of Doctor of Philosophy

In

Pharmaceutical Sciences

Faculty of Pharmacy and Pharmaceutical Sciences

Edmonton, Alberta

Spring 2002

0-612-68612-4

University of Alberta

Library Release Form

Name of Author: Alan Simon Pickard

Title of Thesis: Responsiveness of Generic Health Status Measures in Stroke

Degree: Doctor of Philosophy

Year this Degree Granted: 2002

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Alan Simon Pickard
13362 – 140 Street
Edmonton, Alberta
T5L 2E3

Date: December 10, 2001

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled Responsiveness of Generic Health Status Measures in Stroke submitted by Alan Simon Pickard in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Pharmaceutical Sciences

Dr. Jeffrey A. Johnson

Dr. David Feeny

K. C. Carriere

Dr. K. C. Carriere

Dr. Ashfaq Shuaib

Dennis Revicki

Dr. Dennis Revicki

Date: December 10, 2001

Dedication

To the outcome, a thesis that contributes but a drop of knowledge into humankind's
oceanic library
To the process, neural activity that brought me closer to the energy flow that animates
the universe
To the invisible structures for which I am truly
grateful:
The love of my wife, Suzanne; friends and family that make living
meaningful;
and the unique chemistry of light, time, matter and spirit that makes all life possible.
May the thrill of living
never fade

Abstract

The objectives of this study were to evaluate the responsiveness of the summary scores of several generic health status measures in stroke by self- and proxy-assessment and to examine the substitutability of the two assessment perspectives. The summary scores included the EQ-5D Index (EQ-Index), the EQ-5D VAS (EQ-VAS), the Health Utilities Index Mark 2 and 3 Overall Utility Scores (HUI2 OUS and HUI3 OUS, respectively), and the SF-36's Physical and Mental Component Summary (PCS-36 and MCS-36) scores.

Stroke patients and their caregivers were recruited. Ninety-seven patient-proxy pairs completing a 6-month follow-up by January 2001 were analyzed. Four different approaches to categorizing patient global health as 'improved', 'no change', and 'declined' formed the basis for calculations of responsiveness: patient rates self; patient and proxy agree on patient change; clinician rates patient; change in Barthel Index score-based category. Responsiveness was compared using median rank of each score based on: effect size, standardized response mean, and Guyatt's responsiveness statistic. Ability of proxy-assessment to substitute for patient self-assessment was studied by comparing systematic differences between mean scores, and by using Pearson's r and intraclass correlation coefficients (ICCs).

All external criteria indicated that the majority of patients 'improved'. For the 'improved' patients, proxy- and self-assessed scores of the HUI2 OUS, HUI3 OUS, EQ-Index, and PCS-36 (plus the EQ-VAS for proxy-assessed scores) captured statistically significant and meaningful change in stroke patients. Effect sizes were generally medium to large for each of the external criteria. Self- and proxy-assessed HUI2 OUS, HUI3

OUS and EQ-Index scores (and proxy-assessed EQ-VAS) were associated with larger magnitudes of change. Relative validity ratios indicated that the HUI2 OUS and HUI3 OUS would require a smaller sample size to detect a known group difference than the other summary scores when self-assessments are elicited; this advantage was not as evident for proxy-assessed scores. The EQ-VAS demonstrated inconsistencies for patient self-assessment. Proxy assessments greater than 1 month post-stroke may reliably substitute for self-assessment in cross-sectional studies incorporating the EQ-Index, HUI2 OUS and PCS-36. Consecutive imputation for the same patient is not recommended due to the generally poor reliability of the change scores (all ICCs<0.55).

Acknowledgements

A debt of gratitude to my supervisor *Jeff Johnson*, whose mentorship is invaluable to me.

Sincere thanks to the members of my committee: *David Feeny*, for mentorship and counsel that has benefited me; *KC Carriere* for challenging me; and *Ashfaq Shuaib* for graciously facilitating this study.

I could not have completed the study without the help of research assistants *Abdul Nasser*, *Aldis Hunt* and *Alison Weingardt*, all of whom dedicated much more of themselves than was asked.

Thanks to the people who shared their time with me including *Karen Farris*, *Carlyn Volume*, *Lisa Schapansky* and everyone at the *Institute of Health Economics*, especially *Devidas Menon*, who has been instrumental in supporting this project and my interest in research.

Best wishes and heartfelt thanks to all of the *study participants*.

This research was generously supported by a grant from *AstraZeneca Pharma*. Thanks to the Alberta Heritage Foundation for Medical Research and CIHR/HRF (Rx&D) for their scholarship support.

Table of Contents

List of Tables

List of Tables (in Appendices)

List of Figures

List of Abbreviations

Respondent	
R1	Patient Self-Assessment
R2	Proxy Assessment
Time of data collection	
W0 or T_0	Baseline (wave 0)
W1 or T_1	Month 1
W3 or T_3	Month 3
W6 or T_6	Month 6
SF-36 Scores	
PF	Physical functioning
RP	Role Physical
BP	Bodily Pain
GH	General Health
VT	Vitality
SF	Social Functioning
RE	Role Emotional;
MH	Mental Health
PCS	Physical Component Summary
MCS	Mental Component Summary
HUI Mark 3	
Saus	Single Attribute Utility Score
Sav3	Vision
Sah3	Hearing
Sas3	Speech
Saa3	Ambulation
Sad3	Dexterity
Sae3	Emotion
Sac3	Cognition
Sap3	Pain
OUS3	Overall Utility Score for HUI 3
HUI Mark 2	
Sas2	Sensation
Sam2	Mobility
Sae2	Emotion
Sac2	Cognition
Sat2	Self-Care
Sap2	Pain
OUS2	Overall Utility Score for HUI 2
EQ-5D	
eqmo	Mobility
eqsc	Self-Care
equa	Usual Activities
eqpd	Pain/Discomfort
eqad	Anxiety/Depression
EQ-VAS	EQ-5D Visual Analog Scale Score
EQ-Index/EQ-idx	EQ-5D Index-based Score
MRS	Modified Rankin Handicap Scale
NIHSS	National Institutes of Health Stroke Scale
SSS-48	Scandinavian Stroke Scale (48 score)
Bart Indx, BI	Barthel Index

CHAPTER 1: INTRODUCTION

1.1.0 Statement of the Problem

Perceptions of health and the expression of those perceptions present many philosophical and methodological challenges to researchers who wish to quantify health in a systematic and meaningful way. Descriptors of health such as rates of mortality (e.g. infant mortality rate) and morbidity (e.g. percentage of population with type 2 diabetes) convey limited information about the health of individuals, patient subgroups, and the population. Measurement of health status and health-related quality of life (HRQL) further refines the manner in which health may be evaluated and described. Rising health care costs, the need to evaluate the quality of medical care, and advances in research methods have further driven the health outcomes movement and health status measurement.

Health-related quality of life (HRQL) measures are gaining status as one of the benchmarks of health outcomes. HRQL and health status measures have found numerous applications: to describe population health status; to predict intensity of future resource utilization; to facilitate clinical decision-making; and to assist in determining the effectiveness of health care services and medical interventions.

Several commentaries have suggested that the emphasis of research should shift from the development of HRQL measures to generating a greater body of evidence on the validity of existing instruments (de Haan, 1993; Feeny 1999; McHorney, 1998). Head-to head-comparisons of responsiveness across HRQL instruments, the use of proxy respondents to assess HRQL, and further testing of generic measures in specific patient subgroups are some of the areas requiring further research. Much of the validation of generic HRQL instruments has focused upon discrimination, such as known groups comparisons, which tests the ability to distinguish between different degrees of health status. Evidence on discriminative ability can be obtained by administering the instrument in a single cross-sectional survey. In contrast, a dearth of literature is available on the longitudinal construct validity of HRQL measures that study the ability of instrument to be responsive or sensitive to meaningful change on successive occasions in the same group of patients.

Elicitation of patient self-assessment is not always possible. Some conditions may preclude self-assessment, such as pediatric patients who are not old or mature enough comprehend and adequately respond to a questionnaire. Conditions that impair cognition, such as stroke and Alzheimer's disease, also necessitate the use of proxy-assessments if any information on such patients is to be obtained. The validity and reliability of using proxy assessments to elicit HRQL by generic HRQL measures has been studied to a limited extent.

Investigation into the responsiveness generic HRQL measures and the agreement between proxy- and self-assessment is particularly well-suited to a longitudinal study of a cohort of stroke patients. Stroke is an often debilitating or fatal age-related neurological disorder that occurs in approximately 50,000 Canadians each year, and is a major cause of death and disability in Canada. Despite the considerable impact of stroke on society, little attention had been focused on HRQL in stroke outcome research (de Haan et al, 1993) prior to the mid-1990s. The initial months of post-stroke recovery are often characterized by considerable improvement in a patient's clinical status, thereby providing suitable circumstances for the testing of responsiveness of generic HRQL measures. Furthermore, stroke is a condition where the potential need to perform a proxy assessment of HRQL may arise. Such findings may have direct implications for missing data problems in clinical trials involving stroke patients. Thus, an investigation into responsiveness and aspects of proxy-assessment based upon a cohort of stroke patients and their caregivers (proxies) will potentially contribute both to the general literature, and have direct relevance to the condition being studied.

1.2.0 Research Objectives

The overall purpose of this study was to evaluate and compare the responsiveness of five generic health status instruments in stroke.

The research objectives were to evaluate the responsiveness of self-assessed and proxy-assessed summary scores of the selected HRQL measures, compare and contrast indices of responsiveness between patient self- and proxy generated summary scores, and examine the extent of agreement between self assessments and proxy assessments. The summary scores being evaluated include: the overall utility scores from the Health Utilities Index Mark 2 (HUI 2 OUS) and Mark 3 (HUI 3 OUS) (Feeny et al, 1996); the mental and physical component summary scores of the SF-36 (MCS-36, PCS-36) (Ware et al, 1994); the EQ-5D visual analogue scale (EQ-VAS; a 'feeling thermometer' rated from 0 to 100) and EQ-5D index-based scoring system from York (the EQ-Index) (Dolan, 1997). Thus, there are 5 generic HRQL measures in the study (SF-36, HUI 2, HUI 3, EQ-VAS, EQ-Index) generating 6 summary scores (PCS-36, MCS-36, HUI 2 OUS, HUI 3 OUS, EQ-VAS, EQ-Index). Specifically, the primary objectives of the study were:

1. to compare the responsiveness of the summary scores of selected generic HRQL measures (SF-36, HUI 2 OUS, HUI 3 OUS, EQ-VAS, EQ-Index) obtained by patient self-assessment during a longitudinal study of the post-stroke recovery process;
2. to compare the responsiveness of the summary scores of selected generic HRQL measures (SF-36, HUI 2 OUS, HUI 3 OUS, EQ-VAS, EQ-Index) for stroke patients obtained by proxy-assessment;
3. to compare and contrast the ability of self- and proxy-assessed summary scores to capture meaningful change according to several anchor-based criteria;
4. to evaluate the extent of agreement between stroke patient and proxy HRQL assessment, in relation to cross-sectional and change score agreement.

A secondary objective was to examine the interrelationships (bivariate correlations) between the baseline scores on the various clinical and HRQL measures (domain and summary scores) in order to study the construct validity of the measures for self-assessment and proxy assessment in stroke.

To investigate the issues of responsiveness and agreement, a natural history study of stroke survivors paired with family caregivers was designed. All of the summary scores were expected to capture statistically significant and meaningful improvement (i.e. to be sensitive and responsive, respectively) between the investigated time periods (baseline and 1 month; baseline and 6 months). These expectations would be consistent with clinical reports in the literature on physical and mental changes typically experienced by stroke survivors during their recovery (AHCPR, 1998; Jorgensen et al, 1995; Kelly-Haynes et al, 1989; Dombovy et al, 1987).

Criteria for determining change were defined *a priori*, and several responsiveness indices were calculated to evaluate each of the summary scores. In general, the mean change scores were expected to be in the direction consistent with the category of the change group (i.e. for patients whose global health change was categorized as 'improved', a positive change score was expected). Specific hypotheses were developed that considered the external anchor-based criteria used to group the overall health of patient as improved, no change, or declined. The magnitudes of change scores, as measured by the responsiveness indices, were expected to be lower for the no change group than the improved or declined group.

A priori hypotheses about the extent of agreement between self- and proxy-assessed summary scores was hypothesized. First, there would be greater agreement between cross-sectional scores as more time elapsed and the patients clinically stabilize. Second, greater agreement would occur on the more observable domains (i.e. PCS scores will agree more than MCS scores) for both cross-sectional and change scores. Third, poorer agreement would be observed for the EQ-5D VAS than the other summary scores, because the EQ-VAS scores involve both the patient's assessment and the valuation of their health status. The other summary scores reflect the patient's assessment of health status and community preferences or norms. Fourth, there would be greater agreement between self- and proxy assessed change scores for time periods when important clinical change is expected to take place (i.e. between baseline and 1 month, and 1 month and 3 months), than between 3 and 6 months (i.e. as the health status of the patient stabilizes). Fifth, proxy assessments of HRQL were generally expected to be lower than assessments by patient self-report (Sprangers and Aaronson, 1992; Sneeuw et al, 1997a). Last, cross-sectional scores were expected to demonstrate greater agreement between self- and proxy-assessments than change scores.

Construct validation of baseline scores involved construction of a correlation matrix and the generation of *a priori* hypotheses about the expected strength of association between the clinical measures and domains of the various HRQL instruments.

1.3.0 Significance of the Research

The instruments selected for this study are among the most commonly used generic health status/HRQL measures in clinical trials. This study will generate evidence on the responsiveness of the summary scores of commonly used generic HRQL instruments for the assessment of patients in longitudinal studies, both by patient self-assessment and by proxy-assessment. By using multiple external anchor-based approaches and several indices of responsiveness, evidence on the strengths and weaknesses of these methods will be accumulated. The clinical and generic HRQL measurements will be informative about the post-stroke recovery process. Utility scores generated by indirect preference-based measures can be summarized in relation to level of functioning according to the Barthel Index, Bamford classification of stroke, and Rankin Disability Score, and used to inform medical decision-making. By comparing the agreement between self- and proxy assessed HRQL scores, insight into the potential for substituting proxy assessments for self-assessments in clinical trials at the group level is obtainable. Subsequent analysis of the data collected in this study can be used to compare the ability different statistical modeling techniques used to impute missing data to proxy-assessments of patient HRQL, and evaluated against the patient self-assessments.

Thus, the findings of this study are anticipated to assist clinicians and researchers in selecting HRQL outcome measures for use in clinical practice and research settings such as clinical trials.

CHAPTER 2: LITERATURE REVIEW

2.1.0 Overview of Stroke

2.1.1 The Burden of Illness

Stroke is an often debilitating or fatal age-related neurological disorder that occurs in approximately 50,000 Canadians each year (Hakim et al, 1998). Over 200,000 Canadians are estimated to be survivors of stroke (Hodgson, 1998). In addition to being a major cause of death and disability in Canada, stroke has a significant economic impact on the families of stroke victims and the health care system. The total cost of stroke in Ontario for 1994/95 was estimated to be between $719 and $964 million, with direct costs accounting for 60% of the total costs (Chan and Hayes, 1998).

A stroke is a clinically defined syndrome of rapidly developing symptoms or signs of focal loss of cerebral function with no apparent cause other than that of vascular origin, but the loss of function can at times be global (Warlow, 1998). The syndrome varies from recovery in a day, to severe disability or death (Warlow, 1998). A stroke can cause extensive physical disability and social maladjustment, and the rehabilitation process is frequently long and affects all aspects of a person's life (Bendz, 2000).

The extent of recovery from stroke varies considerably among stroke patients. The majority of neurological and functional recovery from stroke occurs mainly within the first 6 months of the event (Jorgensen et al, 1995), and most rapidly in the first 1 to 3 months after a stroke (Kelly-Haynes et al, 1989). A long-term follow-up of stroke survivors in Minnesota found the proportion of stroke patients with moderately severe to severe disability decreased from 58% at the time of stroke to 26% at the time of institutional discharge, and 17% after 6 months (Dombovy et al, 1987). The level of disability remained relatively constant from 6 months through 5 years of observation. A British study of stroke patients found that 106 were alive (42%) after 5 years; 29% were moderately to severely disabled, 37% were mildly disabled, and 34% were functionally independent (Wilkinson et al, 1997). A 6-year follow-up of stroke survivors from the Auckland Stroke Study reported 639 of 1761 cases (36%) still alive, and health-related quality of life (HRQL) to be relatively good for the majority of survivors, despite significant ongoing physical disability (Hackett et al, 2000).

The burden of illness caused by stroke affects not only the patient, but can also have a tremendous impact on family caregivers. Caregivers of stroke patients may have elevated levels of depression at both the acute and chronic phases of stroke (Han and Haley, 1999). Stroke caregivers have higher levels of anxiety compared with norms (Evans et al, 1989), and caregivers are more likely to be depressed if the patient is more physically impaired, if the caregiver reports disharmony in the family, and if they have lesser perceptions of hope (Thompson et al, 1990). Life satisfaction, a component of quality of life, has been found to be lower among caregivers than in the general population and is directly associated with level of caregiving stress (Segal and Schall, 1996). Han and Haley (1999) indicated that much of the literature on stroke caregivers has focused on depression, with few studies addressing such areas as the physical health of caregivers and evaluation of caregiver-centered interventions.

2.1.2 Treatment and Rehabilitation

Treatment and rehabilitation of the stroke patient depend upon the time elapsed since the acute event, the part of the brain in which the stroke occurred, and the degree and type of functional impairment resulting from the stroke. In addition, practical limitations, such as the availability of specialists in neurology, specialized stroke units, and thromolytic agents, may dictate treatment and rehabilitation options. Patient treatment and rehabilitation is also dictated by the numerous possible problems that have been associated stroke: hemiplegia or hemiparesis, compromised skin integrity, bladder and bowel dysfunction, motor dysfunction, communication disorders such as aphasia and dysarthria, visual-spatial dysfunction, depression, decreased social functioning, loss of sexual functioning (Granger et al, 1987), cognitive impairment (Kwa et al, 1996), and central post-stroke (thalamic) pain (Segatore, 1996).

Until recently, health care has focused on support care and rehabilitation, but stroke is now recognized as being treatable and preventable. While considerable mortality is associated with stroke, a slight trend toward improved mortality rates appears to be emerging with the advent of new therapeutic advances; the 30-day mortality rate for acute stroke patients in Ontario was 19.5% in 1992 and 19.2% in 1996 (Tu and Porter, 1999). Studies on the use of specialized stroke units and stroke teams have reported substantially reduced mortality and morbidity and improved patient outcomes (Jorgensen et al, 1999; Indredavik et al, 1998; Mayo et al, 2000a; Ronning and Guldvog, 1998; Ontario Ministry of Health Report of the Joint Stroke Strategy Working Group, June 2000, p 42). However, there is little consistency in the measurement of outcome in acute stroke trials at the present time, and this may complicate interpretation of the results and reduce the likelihood of detecting worthwhile effects (Duncan et al, 2000).

The prospect of technological advancements that reduce stroke-related mortality is likely to stimulate more research into the burden of stroke and quality of life among stroke survivors and their caregivers. The prevalence of stroke is such that many Canadians have been impacted directly or indirectly by stroke. The majority of patients who experience stroke will survive longer than a year, but the condition is associated with a considerable burden of illness that impacts the patient, their family, and companions. A substantial literature has evolved around quality of life issues that arise as a result of the compromised health status of many stroke survivors. The concept of health-related quality of life (HRQL) and the study of HRQL in stroke will be explored next.

2.2.0 The Concept of Health-Related Quality of Life

The World Health Organization (WHO) (1958) defines health as "not merely the absence of disease or infirmity, but a state of complete physical, mental, and social well-being" (as cited by Berzon, 1998). Health status, quality of life, and functional status are three concepts often used interchangeably to refer to 'health' (Guyatt et al, 1993). The WHO definition of health provides the framework for many definitions of HRQL. HRQL represents those aspects of quality of life that directly relate to an individual's health, which conceptually include the domains of physical, psychological, social, spiritual, and role functioning, as well as general well-being (Spilker and Revicki, 1996). HRQL may be considered synonymous with subjective health status assessment, and involves the aspects of a person's experience that are affected only by health care interventions (Berzon, 1998). Another perspective defines HRQL as being "the value assigned to duration of life as modified by the impairments, functional states, perceptions, and social opportunities that are influenced by disease, injury, treatment, or policy" (Patrick and Erickson, 1993, pg 22). Note that this definition further qualifies

HRQL as involving a valuation. As a result of the multidisciplinary appeal of HRQL research, the operationalization of HRQL constructs often vary according to the background of the researcher. However, most agree that health is complex and multi-dimensional, and measurement of HRQL involves the assessment of subjective and objective attributes.

2.2.1 Considerations in Selecting a HRQL Instrument

Considerable theoretical and application-driven development of HRQL measures has occurred in recent years. This development is warranted and is likely to continue for many reasons, not the least being the need to quantify and evaluate health as an input and output of health care systems around the world. Population health status monitoring, treatment effectiveness evaluations, prediction of future resource utilization, patient monitoring, and economic evaluations to derive 'value for money' estimates for the allocation of health care resources are all potential applications of health status and HRQL instruments. Choice of a suitable HRQL measure will depend upon the goals of the investigation, as well as practical considerations such as mode of administration, respondent burden, and availability of resources.

HRQL measures can be broadly categorized as generic instruments (including single indicators, health profiles, and preference-based measures) and specific instruments (specific to disease, population, function, or condition) (Guyatt et al, 1993). Generic instruments permit broad comparisons across a variety of conditions, but they may be insensitive or unresponsive to changes in specific conditions. Profiles can detect differential effects on different domains or attributes of HRQL (Guyatt et al, 1993; MacKeigan and Pathak, 1992). Health status profiles are not preference weighted and may not provide summary scores, thus leaving an array of scores to be interpreted. Preference-based measures reflect preferences for treatment process and outcome, and summarize HRQL as a single number that can be used for cost-utility analysis (Guyatt et al, 1993). An argument can be made for the inclusion of a specific measure, a generic measure, and a preference-based measure when the major focus of a study is patient benefit (Guyatt et al, 1996; Canadian Coordinating Office for Health Technology Assessment, 1997).

2.2.2 Applications and Measurement Properties

Guyatt, Feeny and Patrick (1993) indicate that HRQL measures have a number of applications: description, discrimination, prediction and evaluation. The purpose of a discriminative measure is to distinguish between individuals or groups at a point in time with respect to an underlying dimension. An instrument with predictive properties is able to classify individuals into a set of predefined measurement categories when a gold standard is available, which is subsequently used to determine whether the individuals have been classified correctly. Prediction can refer to the ability of a short form to predict a long-form or it can be longitudinal. A measure used for evaluative purposes assesses the magnitude of within-person change over time.

The validity of a measure should be considered in the context of its potential application when choosing a HRQL instrument. Much of the literature on validity of HRQL measures has focused on the discriminative abilities of measures. Fewer studies have been devoted to examining the evaluative properties of generic HRQL measures, which are of particular concern to investigators who wish to employ HRQL measures in longitudinal studies.

The validity and reliability of a HRQL instrument cannot be established through a single study. The concept of reliability is a fundamental way to reflect the amount of error, both random and systematic, inherent in any measurement (Streiner and Norman, 1995, pg 104). Aspects of reliability of a scale are examined by studying how reproducible the results of a scale are under different conditions. The essence of reliability is based on the amount of error that is present in a set of scores; a measurement is considered more reliable if a greater proportion of the total observed variance is represented by the true score variance (Portney and Watkins, 2000, pg 558). Many aspects of reliability can be studied: examination of scores from the same observer on two viewings of the same stimulus (intra-observer agreement); different observations of the same stimulus (inter-observer agreement); different occasions separated by a short time interval (test-retest); different items (internal consistency); and different forms of the scale (parallel forms) (Streiner and Norman, 1995, pg 128).

Construct validity may be defined as "the extent to which a measure reflects the concept it is supposed to measure and does not reflect concepts that it is not supposed to measure" (Glossary [Medical Care], 2000). Our knowledge of the determinants of human behavior is imperfect, and hypothesized relationships between an observation and what it reflects have to be validated against actual performance (Streiner and Norman, 1995, pg 145). Thus, validation is a process of hypothesis testing.

Streiner and Norman (1995) describe three categories of validity: content, criterion, and construct. The Glossary [Medical Care] (2000) from the health outcomes methodology symposium proceedings provides definitions for each of those categories. Content validity is the extent to which a measure or battery represents all aspects of a defined concept. Criterion validity is the extent to which a measure corresponds to an accurate or previously validated measure of the same concept or to an external criterion established by the investigators. Construct validity is a process by which theory-based associations that are hypothesized among measures are confirmed through empirical testing.

Further elaborating on definitions from the Glossary [Medical Care] (2000), construct validity can be empirically tested by examining convergent, discriminant, concurrent and predictive validity. Convergent validity tests the strength of association between two measures of a similar construct, while discriminant validity is a test of the extent to which measures are not associated with other measures that are hypothesized to not be associated. Predictive validity is a form of construct validity in which the hypothesis being tested is whether the measure can forecast the likelihood of another event or state. Finally, concurrent validity tests an association of measure that are both assessed at the same point in time.

The 1999 edition of Standards for Educational and Psychological Tests cautions against the separation of validity into subcategories such as content validity, predictive validity and criterion validity (as cited in Lenert and Kaplan, 2000). Instead, the standards propose two new concepts: construct-irrelevant variance and construct underrepresentation. Construct-irrelevant variation occurs when scores are influenced by factors irrelevant to the construct. Construct underrepresentation describes the failure to capture important facets of the construct. These new definitions may eventually displace the historic terminology.

2.3.0 Interpretation of Differences and Changes in HRQL

The ability of measures to detect change over time, and how best to present and interpret observed changes are key areas of concern with respect to measurement validity (Patrick and Chiang, 2000a). In defining the evaluative properties of a measure, sensitivity can be differentiated from responsiveness. Sensitivity has been defined as the ability of an instrument to measure change in a state, irrespective of whether it is relevant or meaningful (Glossary [Medical Care], 2000). Responsiveness refers to the extent to which an instrument can detect changes in scores over time that are important or meaningful, even if those changes are small.

Responsiveness is conceptually the same as longitudinal construct validity, although responsiveness is sometimes categorized as a measurement property distinct from validity and reliability. However, responsiveness is not an intrinsic characteristic of a measure, as the process of deciding whether change is minimally important or what the change means can be considered interpretation (Patrick and Chiang, 2000a).

The threshold that defines an instrument as responsive is the minimally important difference. The minimally important difference is that difference in score on a health-related quality of life instrument that corresponds to the smallest change in status that stakeholders (persons, patients, significant others, or clinicians) consider important (Glossary [Medical Care], 2000). Similarly defined is the minimal clinically important difference (MCID), "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management" (Jaeschke et al, 1989). Although clinically important differences and interpretability are crucial to the widespread adoption of HRQL measures into clinical practice, there is no consensus on the most appropriate method to assessing the ability of an instrument to capture change.

Approaches to interpretation of differences in HRQL can be framed in terms of the individual, or from a population perspective. Anchor-based interpretations of clinical meaningfulness anchor or compare changes seen in an individual's HRQL to an external criterion, such as a clinical measure. The clinical significance of change may also be based on population benchmarks such as population attributable risk, resource utilization and the cost per unit of HRQL moved (Lydick and Epstein, 1993). Anchored-based HRQL changes initially framed in terms of the individual are often further analyzed as a group using distribution-based methods.

2.3.1 Anchor-based Interpretation

Because there is no 'gold standard', the evaluation of responsiveness must define criteria thought to represent meaningful change. Anchored-based approaches include within-patient differences, between-patient differences, and clinician judgment (Neymark et al, 1998).

A seminal study that investigated within-patient differences to estimate the minimal clinically important difference began with a discussion among those with extensive experience with HRQL questionnaires (Jaeschke et al, 1989). The hypothesis that the MCID of 0.5 per item on a 7-point scale was tested by asking patients to provide a global rating of change in their health status, expressed on a 15-point scale (no change and a 7-point scale in two directions). Comparable study designs have since been implemented across numerous disease states using patient's judgments to investigate within-patient differences, and a consistent pattern appears to have emerged (Neymark et al, 1998): a change of 0.5 (on the 7 point scale) per disease specific question represents a minimal difference; 0.75 to 1.25 can be considered a moderate difference; and greater than 1.5 can be considered as large. The global rating of change question has been criticized on the basis that patients systematically underestimate their initial state, provide retrospective estimates of change that are highly correlated with their present state, and lack data on test-retest reliability of the subjective change in health status question (Norman et al, 1997; Wyrwich and Wolinsky, 2000). Because retrospective ratings of change are only weakly related to the size of treatment effects, they are of limited value in the assessing the impact of therapeutic interventions (Norman et al, 1997).

Another approach focuses on between-patient differences, where patients with a particular condition rate themselves relative to another person. When applied in a cohort of patients diagnosed with rheumatoid arthritis, participants generally rated themselves as less disabled than others (Wells et al, 1993). A major limitation of this method is the noise and potential for measurement error introduced by each patient's perception of their own and their paired partner's HRQL (Wyrwich and Wolinsky, 2000).

Responsiveness can be defined and therefore evaluated from perspectives other than the patient, such as using his or her proxy, society, or the health care professional (Liang, 2000; Juniper et al, 1996). Clinicians' judgments of changes in scores of measures that are well known to them have been used to determine clinically important change. Groups who have experienced change may be defined using criteria whereby independent patient and clinician assessments indicate that the patient has changed (Deyo et al, 1991). Other anchor-based interpretations of clinical meaningfulness have been based on life events, the threshold effect, and predictive ability using receiver operator curves (Lydick and Epstein, 1993).

2.3.2 Distribution-based Interpretation

Numerous standardized or 'distribution-based' approaches have been applied to examine clinically important differences in health status measures. Statistical approaches include individual effect size, effect size index (Liang et al, 1985), relative efficiency, standard error of measurement, mean square error (MSE), standardized response means, Guyatt's responsiveness statistic, receiver-operator characteristic curves, comparison of F ratios, responsiveness coefficient, reliable change index, and SEM (as cited in Liang, 2000), and smallest real difference (Pfenning et al, 1999). These methods can be used to determine sensitivity. The approach is simplified in that only statistical significance need be demonstrated, rather than also requiring a clinically important difference. Effect size statistics have limited appeal when comparing instruments across studies because the magnitude of the test statistic is heavily influenced by the characteristics (e.g. degree of heterogeneity) and size of the sample (i.e. the paired t-test) (Norman et al, 1997). Nonetheless, effect size variants may be useful as relative measurements of sensitivity and responsiveness when comparing instruments in studies based on the same sample.

Effect size statistics are derived from variants on signal-to-noise ratios generated by the sample. One formula for effect size is a ratio of the mean change in raw scores to the standard deviation of the baseline scores (Kazis et al, 1989). The standardized response mean is similar, except the denominator is the standard deviation of the change scores (Hays et al, 1998).

The responsiveness statistic (Guyatt et al, 1987) attempts to improve on the definition of the signal to noise ratio by modifying the effect size statistic, designating the denominator (noise) as the standard deviation of change score among stable subjects. To generate this statistic, the 'change' group must be differentiated from the 'no change' group. The change group is defined using an external criterion of change, such as a clinical measure, clinician judgment, or a global health change question rated by the patient or a proxy. Some studies have used a 15-point global health change question, subsequently grouping patients into 3 categories: if patients scored -1, 0 , or +1 they are considered to have stayed the same and if they scored between -7 and -2, or +2 and +7 they are considered to have clinically changed (Juniper et al, 1996).

Standard error of the measurement (SEM) is defined as the standard error in observed scores that obscures the true score, and is the product of the standard deviation at baseline and the square root of 1 minus the reliability of the HRQL measure (Wyrwich and Wolinsky, 2000). Based on a SEM of 1, 2 and 2.77, the SEM was studied for for its correspondence to a minimal clinically important difference (MCID) established for several previous studies (Wyrwich et al, 1999; Wyrwich and Wolinsky, 2000). In a study of patients with coronary artery disease and congestive heart failure, a SEM of 1 corresponded well to the patient-driven MCID standards on all Chronic Heart Failure Questionnaire dimensions (Wyrwich et al, 1999). However, the SEM approach suffers from the same limitations as the anchor-based methods, both being limited by the validity and reliability of a global health change question or whatever criteria is used to form the basis for establishing a MCID.

Effect size has an intuitive appeal as a standard for representing responsiveness. Samsa et al (1999) advanced the idea of basing clinically important difference benchmarks upon the effect size because: (1) it is efficient (requires relatively few resources); (2) it has been applied outside the field of HRQL; (3) a tradition already exists for presenting HRQL using effect size; (4) it is based upon external standards; and (5) they are similar to benchmarks obtained from more explicit anchor-based approaches. Effect size statistics are problematic in that standards for HRQL assessments may differ from Cohen's traditional benchmarks (Kazis et al, 1989), and estimates of effect size can vary widely among samples taken from the same population (Samsa et al, 1999). Thus, the development of standardized ES-based CID benchmarks advocated by Samsa et al (1999) requires large representative samples to establish group and individual benchmarks (Wyrwich and Wolinsky, 2000).

2.3.3 Issues in Interpretation of Responsiveness and Important Change

Some authors believe that defining absolute thresholds for MCIDs for HRQL measures is fraught with conceptual and practical problems. Hays and Woolley (2000) have described some conceptual problems with defining a MCID: if an improvement in group-based scores lies below the MCID threshold, is it worthless? Any threshold for a MCID is debatable, in the same way that a cutoff for categorizing patients as mildly hypertensive will always be arbitrary to an extent. MCIDs are typically derived from average change in HRQL for a group, so is it meaningful to infer the amount of change that is detectable or important to individuals based on a group average? Other practical problems with estimation of a single threshold to establish MCIDs include a dependency on the distributional index and external standard or anchor; the direction of change; and the starting the baseline value. Perhaps the solution lies in solely focusing on individual patient scores.

Despite the problems inherent in estimating MCIDs for HRQL measures, it is worth exploring some estimates of MCID for generic HRQL instruments. Samsa et al (1999)'s literature review indicated that the MCIDs for the Sickness Impact Profile (SIP) and SF-36 were in the range of 3 to 5 points, regardless of whether the study was longitudinal or cross-sectional. On the EQ-5D, the smallest coefficient representing transition between levels within a domain is 0.03 on the York scoring system (Dolan, 1997), and if movement between domain levels on the EQ-5D is considered important, 0.03 might be interpreted as the smallest meaningful difference. A CID for the overall multi-attribute utility score on the HUI 2 is 0.03 (Feeny, personal communication, 1998). Developers of HRQL scales have avoided publishing MCID thresholds for generic instructions.

There is no well-established methodology that solves the problem of clinically meaningful changes with a single strategy, so a variety of approaches must be implemented to increase the meaning of HRQL measures (Guyatt in Neymark et al (1998)). Identification of a clinically meaningful difference is part of a more general goal of providing familiar anchors that aid in the interpretation of unfamiliar units (Hays and Woolley, 2000), such as assisting clinicians and researchers in interpreting HRQL research. Researchers have begun to engage in comparisons between effect size and anchor-based techniques. Norman et al (2001) demonstrated that treatment benefit can be directly estimated from the effect size, and that effect size and anchor-based approaches provide equivalent information about the proportion of patients who will benefit from treatment. Proportion benefiting from treatment was found to be independent of the choice of minimally important difference.

Other issues other those of statistical approach and interpretation may impose limitations on studies of responsiveness. Evidence of validity and generalizability are limited by the patient sample evaluated. Instrumentation bias, arising from ceiling or floor effects, limits the responsiveness of an instrument. The precarious process of recalling whether or not one has changed, and retrospective modification of previous valuations further confounds the ability to easily measure change.

HRQL changes have been described as originating from 4 sources: (1) a catalyst, referring to changes in the respondent's health status; (2) antecedents, pertaining to stable or dispositional characteristics of the individual (e.g. personality); (3) mechanisms, encompassing behavioral, cognitive, or affective processes to accommodate the changes in health status (e.g. initiating social comparisons, reordering goals); and (4) response shift, defined as changes in the meaning of one's self-evaluation of HRQL resulting from changes in internal standards, values, or conceptualization (Sprangers and Schwartz, 1999). Response shift resulting from adaptation to chronic conditions (Postulart and Adang, 2000; Groot, 2000) may serve to attenuate or to exaggerate estimates of change in longitudinal evaluations, further justifying criterion measures of change be included in studies of HRQL (Schwartz and Sprangers, 1999). The meaning of responsiveness statistics has been questioned, particularly when different statistics result in different conclusions when comparing the performances of different HRQL measures (Hays et al, 1998). Recall bias can also compromise the measurement of responsiveness, such as in the use of a global health change question (Norman et al, 1997). These issues emphasize the need to use several external criteria of change for a more robust evaluation of responsiveness. In general terms, longitudinal evaluations involving HRQL outcome measures must ensure that concepts are correctly defined and measured, that the validity of measures used for different applications and in different populations is well-documented; and that observed effects can be clearly interpreted (Patrick and Chiang, 2000b).

2.4.0 Modes of Administration and Proxy Assessment of HRQL

Alternative forms have been developed and validated for a number of HRQL measures in order to accommodate variations in modes of administration, cultural backgrounds, language and the perspective of the respondent and/or assessor.

2.4.1 Modes of Administration

HRQL measures are administered by written or computer-based self-completion, or by trained interviewers. Face-to-face interviews tend to maximize response rates and minimize missing items as well as the misunderstanding of items, but are resource intensive and may reduce the willingness of respondents to acknowledge problems (Guyatt et al, 1996). Telephone-administered interviews have similar advantages while requiring fewer resources, but this mode limits the format of the instrument. Self-completion requires the least amount of resources, but is associated with a greater likelihood of low response rates, missing items and misunderstanding. Separate studies investigating the effect of mode of administration on the SF-36 (Weinberger et al, 1996) and the Medical Outcomes Study HIV Health Survey (MOS-HIV) and EuroQol (Wu et al, 1997) were not able to demonstrate that discrepancies in responses were attributable to different modes of administration.

2.4.2 Proxy Respondents

The measurement of HRQL is challenging in certain patient populations, such as in pediatric patients, or in very sick or elderly patients. Under such circumstances, a proxy respondent may be used to assess the health status of patient. Surrogate responders reduce stress on the patient and allow for the inclusion/assessment of a patient, but perceptions of the proxy may differ from those of the patient (Guyatt et al, 1996). Although it could be argued that the use of proxy respondents runs contrary to the conceptual basis for self-assessment of HRQL, which is predicated upon patient perception, the use of proxies represents an approach to capturing otherwise missing data. Further, in some cases, the perspective of the proxy may provide important complementary data to self-reported information (e.g. child and parent reports of health status).

Missing data present difficulties in the design and analysis of longitudinal studies for two reasons: loss of power to detect change over time or differences between groups; and potential for bias of the estimates as a result of non-randomly missing data (Curran et al, 1998). The use of proxy respondents may represent a preferable method of dealing with missing data when compared to the alternatives, such as imputation-based approaches to the analysis of missing data or the total exclusion of patients who are unable to respond to HRQL questionnaires. In addition, the use of proxy respondent for missing data can serve the dual purpose of presenting an alternative perspective on the patient's HRQL.

Literature on the use of proxy respondents in health services research contains ongoing discussion about the validity and reliability of proxy respondents in assessing patient HRQL using specific and generic HRQL measures. One of the central issues to the use of proxy respondents is that the very ability to assess one's own quality of life and communicate an assessment may be affected by the disorder (Coen, 1999). A number of conditions that exemplify this point have been studied using proxy respondents to assess health status, including: the elderly disabled (Pierre et al, 1998), Alzheimer's disease (Neumann et al, 1999), epilepsy (Hays et al, 1995) and stroke (Mathias et al, 1997; Sneeuw et al, 1997b; Pickard et al, 1999).

Methodological inquiries into the reliability of proxy assessments that focus on stroke have been performed using several different generic health status instruments, including the Health Utilities Index (HUI) (Mathias et al, 1997), the EuroQol (Dorman et al, 1997b), the Sickness Impact Profile (SIP) (Sneeuw et al, 1997b), and the SF-36 (Segal and Schall, 1994). The general approach of this growing body of literature has been to examine the agreement between patient self-report and a proxy respondent, such as a family caregiver or health care provider, in a cross-sectional study. The conclusions of these investigations have been mixed. Mathias et al (1997) reported moderate to high agreement between stroke patients and proxies on a modified, combined version of the HUI2/3, suggesting that family caregivers can complete the HUI reliably when patients are unable to do so. Dorman et al (1997b) concluded that the HRQL information obtained on stroke patients by proxy for the more observable domains of the EQ-5D may be sufficiently valid and unbiased to be useable in most types of trials and surveys, but found poor agreement for the domain that assessed psychological function. Sneeuw et al

(1997b) studied the validity of proxy assessments using the SIP, and found moderate to high intra-class correlation coefficients (ICC) for most of the SIP subscales (average ICC 0.63). On the other hand, Segal and Schall indicated that proxy agreement for the HSQ (SF-36) scales was poor, with a median ICC of 0.32 for the eight dimensions. Agreement was highest on the physical functioning dimension (ICC 0.67), but was otherwise poor for the other dimensions that largely consisted of more subjective items. The authors postulated that poorly educated respondents had more difficulty with comprehension of the HSQ items, further detracting from interrater agreement.

Characteristics of the proxy assessor have been found to influence the extent of agreement with patient self-assessment. A comprehensive review by Sprangers and Aaronson (1992) identified a number of trends: health care providers and significant others tend, in general, to underestimate patients' HRQL (relative to the patients' self-assessment); health care providers and significant others appear to evaluate patients' HRQL with a comparable degree of (in)accuracy; health care providers tend to underrate the pain intensity of their patients; proxy ratings appear to agree with patient ratings when the information sought is concrete and observable; and while significant others' ratings tend to be more accurate (greater agreement) when they live in close proximity to the patient, they can also be affected (poorer agreement) by the caregiving function of the rater.

Since the review by Sprangers and Aaronson (1992), numerous studies have been published with findings on patient/proxy characteristics that do not lend themselves to generalizations. An evaluation of HRQL assessments by cancer patients, significant others, and physicians and nurses found that disagreement was not dependent on the type of proxy rater, or on raters' background characteristics, but was influenced by the HRQL dimension being considered and by the clinical status of the patient (Sneeuw et al, 1999). A study of cancer patients reported that although several characteristics of the patients and their significant others were associated with level of agreement, the characteristics explained less than 15% of the variance in patient-proxy differences (Sneeuw et al, 1997b). Extent of disagreement has been attributed to greater caregiver strain (Knapp and Hewison, 1999). A study of concurrence between subject and proxy ratings of HRQL for people with and without intellectual disabilities reported a high degree of concurrence overall, with no factors (living arrangement, patient or proxy gender, empathy) directly affecting agreement when proxies were selected on the basis of close and regular contact (McVilly et al, 2000).

A common implication of the findings on agreement between patient and proxy assessments respects the substitutability of the proxy assessment for the patient assessment. Generally, conclusions about the substitutability are cautiously worded and describe specific circumstances and/or conditions for which proxy assessments are appropriate. Hays et al (1995) indicated that for group level comparisons, proxy respondents can be substituted for adults with epilepsy having low to moderate seizure frequency. However, the authors cautioned that individual level assessments by proxy should be used with caution. A study of cancer patients with brain metastases using the Spitzer QL-Index suggested that substituting proxy ratings for patient ratings in cancer clinical trials could lead to different conclusions concerning radiation therapy's effect on quality of life (Moinpour et al, 2000). The results of 130 paired proxy-patient assessments of the EQ-5D were liberally generalized by the authors to be sufficiently valid and unbiased to be useable in most types of trials and surveys (Dorman et al, 1997b). In an unrelated study of stroke, Sneeuw et al (1997a) suggested that the benefits of using proxy ratings for non-communicative stroke patients outweigh the limitations.

A study of the use of significant others as proxy raters of the quality of life of patients with brain cancer discussed the validity of using proxy assessments rather than patient assessments despite lack of agreement (Sneeuw et al, 1997a). Although the patient's rating is generally considered central and taken as the gold standard to which the proxy rating should conform, the authors found that the reliability (both test-retest and internal consistency) of the proxy-generated data was slightly higher than that of the patient.

The validity of multiple perspectives is an important consideration in the analysis and interpretation of proxy assessments. A subtle yet methodologically important aspect of research into the use of proxy assessments is the viewpoint of the proxy. Proxy respondents may be asked to complete health status assessments from the perspective of the patient (i.e. how do you think the patient would respond if he or she could; the proxy-patient perspective). Alternatively, the proxy may be asked about their view of the health status of the subject (proxy-proxy perspective). The proxy-patient perspective is an example of substituted judgment, which requires the proxy to "stand in the shoes" of the patient (Coen, 1999), and is an attempt to provide a perspective closely aligned with that of the patient. The proxy-proxy perspective represents the intentional elicitation of a different viewpoint that may or may not correspond with the views of the patient, if it were possible for the patient to express their point of view. For example, studies of HRQL in pediatric cancer patients have found information provided by patients, parents and healthcare professionals is often complementary and each has a valid and important perspective (Feeny, 1999; Pickard et al, 2000).

The explicit statement of proxy perspective is important because the two perspectives serve slightly different purposes. The proxy-patient perspective was studied by Epstein et al (1989), who found strong correlations between proxy and patient assessments of overall health, functional status, social activity and emotional health. This viewpoint was also examined in the disabled elderly, but findings indicated only poor to moderate agreement between proxy and patient pairs (Pierre et al, 1998).

The alternative viewpoint, from the proxy-proxy perspective, is also commonplace in the literature. This perspective is the explicit standard for proxy versions of the HUI2 and HUI3 questionnaires. Examples of this perspective have been employed in studies of childhood cancer (Varni et al, 1998), epilepsy (Hays et al, 1995), stroke (Mathias et al, 1996) and pediatric asthma (Guyatt et al, 1997). An advantage of the proxy-proxy perspective is that it may serve as an external criterion for validation purposes, such as in the examination of self-reported changes scores.

2.4.3 Statistics of Reliability and Agreement

Agreement between patient and proxy assessments is typically approached with the traditional methods for analyzing agreement between pairs of continuously scaled responses, which include: (1) a paired t-test on the difference score to detect the presence of a bias in the proxy sample mean relative to the patient sample mean; (2) the Pearson product moment correlation coefficient r; and (3), the intraclass correlation coefficient (ICC). Pearson's r is a statistic of association, while the ICC is a statistic of agreement, which contains information on systematic differences in scale location in addition to the extent of association between responses. Structural equation modeling (SEM) has also been described as an approach to evaluating agreement between clinical assessment methods (Marshall et al, 1994)

The intraclass correlation coefficient (ICC) has been employed to test for agreement between patient and proxy respondents. Six forms of ICC have been described by Shrout and Fleiss (1979) based on ANOVA models: a one-way random effects model, two-way random effects model, and two-way mixed effects model (ICC cases 1, 2, and 3, respectively). The one-way model separates the targets being rated (e.g. the patients) as the between-group variance (the between group mean squares (BMS)) and the remaining variance is assigned to the within-error term (within group mean squares (WMS)) (Shrout and Fleiss, 1979). The one-way model is expressed as:

$$\mathrm{ICC}\ (1,1) = \mathrm{BMS} - \mathrm{WMS} / (\mathrm{BMS} + (k - 1)\ \mathrm{WMS}$$

where k is the number of judges. If the same k judges rate all n targets (i.e. if the same raters rate all patients), it is possible to separate the effects of the judge and the judge x target interaction from the error variance. Thus, the 2-way random effects ICC is:

$$\mathrm{ICC}\ (2,1) = \mathrm{BMS} - \mathrm{EMS} / [\mathrm{BMS} + (k-1)\mathrm{EMS} + k\ (\mathrm{JMS} - \mathrm{EMS})/n],$$

where EMS is the residual mean squares and JMS is the mean squares of judges. Because the effect of judges is the same for all targets under cases 2 and 3, interjudge variability does not affect the expectation of BMS (Shrout and Fleiss, 1979). Case 3 differs from case 2 in the assumption that the judges are fixed, and is expressed as:

$$\mathrm{ICC}\ (3,1) = \mathrm{BMS} - \mathrm{EMS} / [\mathrm{BMS} + (k-1)\mathrm{EMS}].$$

To differentiate when the use of case 2 or 3 is appropriate, Shrout and Fleiss (1979) provide the following example. "Suppose a reliability study (G study) precedes a substantive study (the decision study) in which each of the k judges is responsible for rating his or her own separate random sample of targets. If all the data in the final study are combined for analysis, the judges' effects will contribute to the variability of the ratings, and the random model with its associated ICC (2,1) is appropriate. If, on the other hand, each judges' ratings are analyzed separately, and the separate results pooled, then interjudge varilability will not have any effect on the final results, and the model of fixed effects with its associate ICC (3,1) is appropriate."

Case 2 or 3 will not be appropriate if different judges rate different targets, because without repeated ratings by the same judges, an effect due to judges cannot be separately estimated. This point is crucial because the paper by Shrout and Fleiss is frequently cited as the primary reference for study designs that are inappropriate for the application of the ICC (2,1) and (3,1). However, the one-way model described by Bartko (1966), which is case 1 from Shrout and Fleiss (1979), is appropriate for unmatched data (where different rater rate each subject) when estimating the reliability of a single rating of each subject (Algina, 1978). For agreement between paired data, the relationship between the Pearsonian correlation and the intraclass correlation has been presented by Robinson (1957) as follows:

ICC (Robinson) = $\{[(s_1^2 + s_2^2) - (s_1^2 - s_2^2)]r - (X_{M1} - X_{M2})^2/2\} / [(s_1^2 + s_2^2) + (X_1 - X_2)^2/2]$,

where X_{M1} and X_{M2} denote the means of X_1 and X_2, s_1 and s_2 the standard deviations, and r the Pearsonian correlation between X_1 and X_2.

For cases 1 to 3 as described by Shrout and Fleiss (1979), two forms have been developed for each case, depending on whether an investigation is interested in the reliability for the average of several ratings (inter-rater reliability) or for the reliability of a single rater (intraclass correlation) (Hays et al, 1998, pp. 172). The average measure ICC determines the reliability based, for example, on the average score for 4 raters, and is higher than the reliability of each rater individually (the single rater ICC). Because the ICCs are derived from ANOVA models (with the exception of the ICC described by Robinson (1957)), data should be evaluated to ensure that the assumptions underlying ANOVA are not violated (Portney and Watkins, 2000).

Although there exists a variety of ICCs, as described above, the type of ICC used is sometimes not described by investigators (Mathias et al, 1997; Dorman et al, 1997; Segal and Schall, 1994). Among studies that have been explicit about the type of ICC employed, strangely, a 2-way random effects model was used in to assess proxy-patient agreement in the elderly disabled (Pierre et al, 1998) and cancer patients (Sneeuw et al, 1997b). The appropriate one-way random effects model ICC was employed by Hays et al (1995) to capture absolute differences not represented by Pearson's r in a study of agreement between proxy and self-report of quality of life in epilepsy.

2.5.0 Patient Outcome Measurement in Stroke

2.5.1 Clinical Measures

Many tests and measures have been developed to assess the functional abilities of the stroke patient (see section 2.5.4 for a discussion of specific measures of HRQL in stroke). Stroke scales and classifications help to deal with the intrinsic difficulties in studying stroke by standardizing its study (D'Olhaberriague, 1996). These measures are often used as part of the criteria to determine the need for health care support services such as physical rehabilitation, homecare support, as well as to clinically characterize the patient in stroke trials and for epidemiological purposes. The Barthel Index (Mahoney and Barthel, 1965), a modified Barthel Index (Granger et al, 1987), the Rankin Handicap Scale (Rankin, 1957), and modified Rankin Scale (MRS) (van Swieten et al, 1988) are all commonly used scales that measure disability or dependence in activities of daily living in stroke patients (Sulter et al, 1999). Barthel Index scores between 50 and 95 have been used as cutoff scores to define favorable outcome. Similarly, favorable outcomes on the MRS have been defined as either less than or equal to 1 or 2 (Sulter et al, 1999).

Clinical trials and research involving acute stroke patients have facilitated the development of stroke specific measures and classification systems. The National Institute of Health Stroke Scale (NIHSS) (Brott et al, 1989) is a 15-item neurological examination designed for use in acute stroke therapy trials. The NIHSS has become one of the most widely used measures for assessing stroke severity. For interpretation of the NIHSS, higher scores are associated with greater stroke severity. Baseline NIHSS scores have been found to predict strongly the likelihood of patient recovery after stroke: a score of greater than or equal to 16 forecasts a high probability of death or severe disability, whereas a score of less than or equal to 6 is associated with a good recovery (Adams et al, 1999).

The Scandinavian Stroke Scale (SSS) (Scandinavian Stroke Study Group, 1985) was developed with intended applications similar to those of the NIHSS. The scale can generate a prognostic score (out of 22) or/and a long-term score (out of 48) (SSS-48). The 9 domains include status of consciousness, eye movement, arm motor power, hand motor power, leg motor power, orientation, speech, facial palsy and gait. A cutoff point of less than 42 points (out of a maximum 48) has been used to dichotomize patients into mild/moderate symptoms and severe symptoms (Kotila et al, 1998).

Several stroke classification systems have been used to describe patients in enrolled in large clinical trials. A widely used classification system for clinically identifiable subtypes of cerebral infarction is known as the Oxford (Bamford) stroke classification (Bamford et al, 1991). This system classifies cerebral infarction into 4 subtypes: lacunar infarcts (LACI), total anterior circulation infarcts (TACI), partial anterior circulation infarcts (PACI), and posterior circulation infarcts (POCI). The Trial of Org 10172 in Acute Stroke Treatment (TOAST) derived five subtype classifications: large-artery atherosclerosis, cardioembolism, small-artery occlusion (lacunar infarction), other etiology, and undetermined etiology (Adams et al, 1999). The American Heart Association Stroke Outcome Classification (AHA.SOC), has 3 components and was developed to measure the full range of domains affected by stroke: the number of affected neurological domains, the severity of impairments, and classifies post-stroke functional disabilities and handicap (Lai and Duncan, 1999).

An appraisal of the evidence of reliability and validity studies in stroke examined the literature on stroke classifications and stroke scales up to 1995 (D'Olhaberriague et al, 1996). The authors concluded the NIHSS was among the scales with the highest reliability, while the Barthel Index was the most reliable disability scale.

2.5.2 Measures of Depression

Depression occurs in almost half of stroke survivors (Kotila et al, 1998). Abnormal emotion affect and depression is particularly common within the first 3 months of stroke (Kotila et al, 1984). Depression scales are among the best-established of health measurements, and some of the best measures are over 20 years old (McDowell and Newell, 1996). Widely-studied scales that are self-administered include the Beck-Depression Inventory (BDI) (Beck, 1961), the Self-rating Depression Scale (SDS) (Zung, 1965), the Center for Epidemiologic Studies Depression Scale (CES-D) (Locke and Putman, 1971), and the Geriatric Depression Scale (GDS) (Yesavage and Brink, 1982). McDowell and Newell (1996) describe the SDS as one of the most widely used scales, but indicate mixed evidence on its validity and reliability. The BDI is one of the best depression screening tools available, but must be purchased on a per test basis. The 32-item GDS and 20-item CES-D have comparable evidence of validity and reliability. However, as the name suggests, the GDS was developed for application in elderly populations.

The CES-D has been extensively used and norms are available. The CES-D is scored out of 60, and consists of 20 items, each with 4 possible response options, scored from 0 to 3 (Weissman et al, 1977). For all except 4 items, the higher score indicates more impairment; for those four items, the scoring is reversed. By summing all items for each patient, a total score for the scale is obtained, and the total score is used as an estimate of the degree of depressive symptomatology. The CES-D has been studied as a screening instrument for depression, with a cutoff score (at or above) 16 being used as the basis to refer patients for diagnostic assessment of depression (Radloff and Locke, 1986).

2.5.3 Generic HRQL Measures

Generic HRQL are designed to be applicable across all diseases or conditions, across different medical interventions and across a wide range of adult populations (Patrick and Deyo, 1989). The SF-36, EQ-VAS, EQ-Index, HUI 2 and HUI 3 are five of the most commonly used generic HRQL measures. These instruments were all included in a review comparing that the major generic HRQL instruments (Coons et al, 2000). The SF-36 health survey was the mostly highly rated profile measure, with extensive evidence on the reliability, validity, conceptual and measurement model, respondent and administrative burden, alternative forms, and cultural and language adaptations. Three preference-based families of measures were reviewed; the Quality of Well-Being Scale (QWB), the HUI and the EQ-5D. The conceptual and measurement model and alternative forms of the HUI were deemed 'extensive', and a rating of 'adequate' for all other criteria. The EQ-5D was rated as having extensive cultural and language adaptations, and was otherwise adequate, with the exception of 'limited' evidence of alternative forms (e.g. proxy-designed forms). The SF-36, the HUI and the EQ-5D are discussed in terms of their content, format, scoring and applicability to stroke below.

2.5.3.1 The Short Form-36 (SF-36) Health Survey

Perhaps the most well-known generic health status measure is the Medical Outcomes Trust SF-36. The SF-36 has been described as a profile-based measure, composed of 36 items grouped into 8 domains that include: physical functioning (PF), role physical (RP), bodily pain (BP), general health (GH), vitality (VT), social functioning (SF), role emotional (RE), and mental health (MH) (Ware et al, 1994). It is available as an acute version (1 week recall) and chronic version (4 week recall), and is designed to be self completed by people 14 years of age or older, or administered by trained interviewers either in person or by telephone.

A vast amount of evidence has been published on the SF-36, including a bibliography referencing over 1000 articles on development, psychometric properties, and applications of the SF-36 (Manocchia et al, 1998). Original authorship is attributed to John E. Ware, Jr., Cathy D. Sherbourne, Ron D. Hays, Anita Stewart, Sandy Berry, and Barbara Gandek (February 13, 2001; http://www.outcomes-trust.org/instruments/catalog.html#9). Permission to use the instrument is required from the Medical Outcomes Trust, but the instrument is obtained from Qmetric (www.qmetric.com). An abbreviated version of the SF-36, known as the SF-12, has been available since 1995 (Ware et al, 1995).

The authors who originally developed the 36 item short form diverged on their conceptual approaches to scoring the instrument. As a result, two approaches to scoring the original 36 items are available. The RAND-36 Health Status Inventory (RAND-36) is promoted by Hays (1998), while the SF-36 scoring system (version 1) has been promoted by Ware et al (1994).

The measurement model of the SF-36 has 3 levels: items, scales that aggregate items, and summary measures that aggregate scales (Ware et al, 1994). The 8 SF-36 scales form 2 distinct clusters with four scales (MH, RE, SF, VT) correlating highest with the MCS and lowest with the PCS, and a second cluster (PF, RF, BP, GH) correlating highest with the PCS and lowest with the MCS. The SF-36 mental and physical component summary scales, MCS and PCS, respectively, are scored using orthogonal factor rotation under the assumption there is no correlation between physical and mental health. The scoring algorithm to achieve this orthogonality gives positive scoring coefficients to the 4 scales more highly correlated and negative coefficients to the 4 scales with lower correlations.

Unlike the SF-36, the mental and physical summary scores of the RAND-36 are non-orthogonal, allowing for correlation between the scores. The RAND-36 scoring procedure is based upon Item-Response Theory (IRT), which allows for the empirical weighting of responses to questions of differing difficulty along a single continuum of health (Hays, 1998). The RAND-36 includes a mental health composite, a physical health composite, and a global health composite (Hays, 1998).

In the measurement of the HRQL of stroke patients, there are potential problems with the content validity of both the domains and the items comprising the domains on the SF-36. A floor effect is likely to be encountered on some items, such as those regarding mobility (Williams, 1998). The limited number of response options on some of the SF-36 items may also hinder the ability of the SF-36 to detect changes in health status.

2.5.3.2 The Health Utilities Index Mark 2 and 3

The HUI Mark 2 and 3 systems are generic preference-based measures. A substantial body of literature has evolved on the HUI systems (http://www.healthutilities.com/references.htm). The HUI2 and HUI3 consist of 2 complementary components: a multi-attribute health status classification system used to describe health status, and a multi-attribute utility function that is used to value the health status described by the classification system (Feeny et al, 1996). The HUI2 was originally developed for application to childhood cancer. The HUI3 was initially developed for population health surveys.

The HUI2 and HUI3 were developed with the intention of capturing 'within the skin' attributes of health status (Feeny et al, 1996). The HUI2 consists of 7 attributes:

sensation (vision, hearing, speech), mobility, emotion, cognition, self-care, pain, and fertility. The 8 attributes of the HUI3 were deliberately selected to be structurally independent and include: vision, hearing, speech, ambulation, dexterity, emotion, cognition, pain.

The HUI Mark2 and Mark 3 are often presented together and have been formatted for interviewer-administration (in person and via telephone) and self-completion. Standardized forms have been developed for both proxy and self-assessment, with recall periods of 1 week, 2 weeks, 4 weeks, and 'usual', a format often used for long-term follow-up or population health surveys (Coons et al, 2000).

Both the HUI2 and HUI3 scoring systems are based upon multiplicative multi-attribute utility functions (Furlong et al, 1998; Feeny et al, 1996). This facilitates calculation of HRQL scores, where dead has a utility of 0 and healthy has a utility of 1.0. Single attribute utility scores can also be calculated for each attribute in the HUI 2 and HUI 3. The utility scores have interval scale properties, whereas the attribute levels do not have interval scale properties.

A number of the attributes of the HUI2 and HUI 3 are specifically relevant to the study of stroke, including speech, dexterity, vision, cognition, ambulation, mood, emotion, and self-care. The HUI2 and HUI3 do not contain items that explicitly inquire about social roles, family roles, energy, work/productivity, and personality. These dimensions, which may be considered 'outside the skin', are important to stroke patients (Williams, 1998). The scoring functions of the HUI2 and HUI3 may indirectly capture these 'outside the skin' attributes, in the same way that the EQ-5D scoring may capture cognitive impairment without including cognition among its domains. Studies of stroke patients have reported evidence of the construct validity of the HUI2 and HUI3. Analyses of a stroke patient sub-sample from a population health survey using the HUI3 found the greatest burden of morbidity in cognition, pain, ambulation and dexterity, compared to a reference group without stroke or arthritis (Grootendorst et al, 2000). Mean overall utility scores for stroke patients have been reported: 0.54 on the HUI3 (Grootendorst et al, 2000), and 0.72 for the HUI2 (Samsa et al, 1999). This is consistent with the observation that HUI3 overall utility scores are usually lower than those on the HUI2 for the same sample for groups with moderate or severe burdens of morbidity.

2.5.3.3 The EQ-5D

The EQ-5D was designed as a cardinal index of health for describing and valuing HRQL (Brooks et al, 1996). The instrument consists of a descriptive health state classification system and a visual analog scale 'health thermometer' (the VAS component). The descriptive health state classification system consists of 5 domains (mobility, self-care, usual activities, anxiety/depression, and pain/discomfort), each with 3 response levels (no problems, some problems, extreme problems). The health 'thermometer' represents a subjective, global evaluation of the respondent's health status on a scale between 0 and 100, where 0 represents worst imaginable health state and 100 represents best imaginable health.

Three types of data are produced for each patient: a health state vector or profile describing the extent of problems on each of the 5 domains, a population-weighted health-index based on the health state vector (the EQ-5D index score), and a VAS-based self-rated assessment of HRQL (Coons et al, 2000). The EQ-5D index score reflects the respondent's rating of his or her function or behavior for each of the 5 domains. The EQ-

5D was intended for self-completion and the recall period refers to the present (today). No alternative forms are reported in the literature, but are anticipated in the near future via a book to be published in late 2001 by the EuroQol group.

The scoring algorithm typically applied to the descriptive system is United Kingdom-based York scoring system (Dolan, 1997). The scoring system was generated from an study where a sample of the general UK population was interviewed, and asked to rank and then value hypothetical EQ-5D health states using the Time Trade-Off approach (Dolan, 1997). Although no Canadian-based scoring 'tariff' has been developed for the EQ-5D, a scoring model has been generated for VAS-based valuations in an adult US sample (Johnson et al, 1998). A study of differences between a European and Canadian-based sample of EQ-5D valuations found VAS valuations for EQ-5D health states were comparable for domains other than Usual Activities (Pickard et al, 2001).

A perceived strength of using the EQ-5D to study HRQL in stroke patients is its brevity. Unfortunately, brevity detracts from the informational content of the measure in the study of stroke. The EQ-5D lacks dimensions of HRQL that may be impacted by stroke such as vision, speech and cognition. The ubiquitous dimension 'usual activities' has the potential to elicit some information on personality, family roles, or productivity, but the bundling of so many potential aspects of HRQL obscures interpretation. A general concern about the EQ-5D relates to the three levels of response options, which would appear to limit responsiveness and sensitivity. Floor and ceiling effects have also been observed to a greater degree in the EQ-5D than in the SF-12, an abbreviated version of the SF-36 (Johnson and Pickard, 2000).

2.5.3.4 Use of Generic HRQL Measures in Stroke

The value of using generic health status and HRQL measures in stroke patients may be demonstrated through their multi-dimensionality, or ability to capture health deficits in key components of well-being, such as physical, mental, emotional, and social functioning. While uni-dimensional measures of capacity for activities and physical functioning such as the Barthel Index are useful for assessing home care needs, the complementary use of generic instruments designed to measure the multi-dimensional aspects of health status can provide more extensive information on post-stroke recovery.

Unlike a comprehensive condition-specific battery of tests, generic instruments can facilitate comparisons of health status and HRQL between different patient subgroups and populations. Preference-based generic measures integrate morbidity and mortality, a characteristic that is particularly useful for capturing the diversity of patient outcomes in stroke. Standardized assessment of persons with stroke must evaluate across the entire continuum of health related-function (Duncan et al, 1997), and inclusion of generic measures may also capture the impact of unforeseen side effects of therapy or symptoms, such as pain, that otherwise may not be recognized by clinicians or researchers. One of the first surveys of patient preferences for stroke outcomes revealed the importance of directly eliciting patient preferences for the consequences of stroke, finding that severe strokes may be viewed as tantamount to or worse than dead (Solomon et al, 1994).

De Haan et al. (1993) performed a review of quality of life measurement in stroke, concluding that the emphasis of future research should be placed on further psychometric evaluation of existing quality of life measures rather than generating new instruments. Since 1993, literature has emerged on the psychometric properties of several generic HRQL instruments in stroke patients (Buck et al, 2000), mainly through cross-sectional studies of validity, feasibility, and reliability of HRQL (Anderson et al, 1996; Dorman et al, 1998; Dorman et al, 1997a; Dorman et al, 1997b; O'Mahony et al, 1998; Grootendorst et al, 2000). Because stroke may result in neurological deficits, the reliability of proxy assessments of the health status of stroke survivors has been examined using similar study designs for the HUI 2 (Mathias et al, 1997), the EQ-5D (Dorman et al, 1997b), the Health Status Questionnaire (SF-36) (Segal and Schall, 1994), and the Sickness Impact Profile (SIP) (Sneeuw et al, 1997b).

Analysis of responsiveness was not cited as an objective of the handful of studies that have involved repeated administration of generic HRQL measures to the same cohort (i.e. a panel design) of stroke patients. The analysis of treatment effect was restricted to tests of statistical significance in intervention-based studies (Indredavik et al, 1998; Anderson et al, 2000; Mayo, 2000). The Nottingham Health Profile (NHP) was used as an outcome measure in a randomized control trial in Norway comparing patients in a rehabilitation stroke unit to those in a general ward, but was given only at the end of the study (Indredavik, 1998). A study by Mayo et al (2000a) evaluated early supported discharge for stroke patients using the SF-36 at 1 month and 3 month found a significantly higher score for the home intervention group on the SF-36 Physical Health Component than the usual care group.

A study comparing different patient subgroups using the EQ-5D was given at baseline, 4 weeks and 3 months to elderly acute care patients included 17 stroke patients (Coast et al, 1998). The authors reported a very large standard deviation in relation to a small mean difference in EQ-5D scores, resulting in effect sizes less than 0.10 for each the measurements.

Finally, the SIP was administered to patients at 3, 6, and 12 months after stroke concluded that HRQL improved 'somewhat' over the period (Jonkman et al, 1998). Scores were 'poor' relative to controls matched for age, last occupation and educational level after 1 year. The decrease in HRQL was correlated with depression and neurological deficit.

Studies have begun to include HRQL instruments among their outcome measures, but there continues to be a need for evidence of the responsiveness of generic HRQL instruments in stroke. Of generic measures are available, the SF-36, EQ-5D, and HUI 2 and HUI3 represent 3 of the instruments most relevant to researchers and clinicians who wish to incorporate generic measures into their research. All have been previously applied in the study of stroke patients. In addition to the attention these measures have received in stroke research and in the general literature as outcome measures in clinical trials and describing patient populations, normative Canadian data is available for all of these instruments. Population norms can be used to compare the relative levels of health status of recovering stroke survivors to the general population.

2.5.4 Stroke-specific HRQL Measures

A review of stroke-specific HRQL conducted in 2000 (Buck et al, 2000) cited evidence of reliability and validity for the Frenchay Activities Index, Niemi QOL scale, Ferrans and Powers QOL Index-Stroke Version, and Stroke Adapted Sickness Impact Profile (SA-SIP30) (Buck et al, 2000). The SAS-SIP30 is a shortened version of the SIP, which reduced the respondent burden from 136 items to 30 items (Van Straten et al, 1997). However, Buck et al (2000) observed that none of the stroke-specific instruments were developed with patient-centered approaches to ensure all HRQL-related issues were covered by the measure.

Since that review, evidence of the validity, reliability, and sensitivity to change has been reported for the Stroke Impact-Scale Version 2.0, a stroke-specific measure (Duncan et al, 1999). The SIS is a self-report measure that includes 64 items and assesses 8 domains (strength, hand function, activities of daily living, mobility, communication, emotion, memory and thinking, and participation). The value of the SIS 2.0 will become more evident as further studies of its psychometric properties are conducted.

Several other stroke-specific measures are currently being developed. A stroke-specific utility-based module to the EQ-5D, elaborates upon the tri-level structure of the EQ-5D by adding domains such as driving (Mayo et al, 2000b). Another measure, the stroke-specific quality of life measure (SSQOL), has 49 items on 12 domains that include energy, family roles, language, mobility, mood, personality, self-care, social roles, thinking, upper extremity function, vision, and work/productivity (Williams et al, 1999). Preliminary results on the reliability, validity, responsiveness of the SSQOL are encouraging.

Because stroke affects many aspects of HRQL and lifestyle, patient-centered stroke-specific measures contain many items and cover numerous domains. A paradoxical situation arises for stroke-specific instrument developers: in order to be comprehensive, a potentially prohibitive respondent burden is created. Caution should be exercised in describing some of the stroke-specific measures as HRQL measures, as they may not be measuring HRQL constructs. A recent systematic review of all stroke specific measures of HRQL concluded no existing measure comprehensively covers all relevant domains or fully addresses the issues of obtaining and combining HRQL assessment in patients and proxies in many stroke populations (Golomb et al, 2001).

2.5.5 Missing Data

Missing data are inevitable in longitudinal studies of stroke patients. In HRQL research, there are two main types of missing: item non-response, when at least one question has not been answered on a questionnaire; and unit non-response, when the whole questionnaire is missing for patient (Curran et al, 1998). Unit non-response may result from intermittent missing forms, drop-out from the study, or late entry into the study.

There are three classes of missing data: missing completely at random (MCAR), where the reasons for missing data are assumed to be completely unrelated to the patient's HRQL; missing at random (MAR), where the missing HRQL data are dependent on previous assessments but are independent of current and future HRQL assessments; and not missing at random (NMAR), where absence of an observation is associated with the current and future HRQL outcomes (Fairclough, 1998; Revicki et al, 2001). Strong assumptions are required to treat missing data as MCAR.

Missing data (NMAR) can make interpretation of treatment of effects difficult, and can introduce significant bias in treatment comparisons. Unit non-response at any observation period will prevent the inclusion of a patient in methods of analysis such as repeated measures ANOVA, which results in the loss of valuable data and introduces informative censoring (particularly when the data are NMAR).

There is no currently accepted technique for imputing missing HRQL scores in clinical trials (Revicki et al, 2001). When individual items that constitute scale scores are missing, the scale score can be treated as missing, simple mean imputation can be employed (where the scale score is estimated from the mean of those items available), or general imputation methods can be used, such as hot deck imputation. Hot deck imputation refers to selecting a score at random from patients with observed data and substituting it for the missing value (Curran et al, 1998). The results of each method can be contrasted to test the robustness of the findings as a form of sensitivity analysis.

When imputing data, the analytic goals to be kept in mind include the generation of unbiased estimates, appropriate estimation of variance, simplified analytic methods (ie, MANOVA), and simplified interpretation (Fairclough, 2000). Hot decking is a form of simple imputation, as are the methods of last value carried forward, simple mean imputation, and regression (predicted value) from covariates. All of these methods of simple imputation result in varying degrees of underestimation of the true standard deviation and the standard error of the underlying population. To compensate, imputation models that add random error for the uncertainty of the estimates are recommended. However, relatively large datasets are required, and there are no completely satisfactory solutions.

CHAPTER 3: METHODS

3.1.0 Research Goals and Objectives

3.1.1 Purpose

The overall purpose of this study was to evaluate and compare the responsiveness of the summary scores of five generic health status instruments (HUI2, HUI3, SF-36, EQ-VAS, and EQ-Index) in stroke.

3.1.2 Objectives

The research objectives were to evaluate the responsiveness of self-assessed and proxy-assessed summary scores of the selected HRQL measures, compare and contrast indices of responsiveness between patient self- and proxy generated summary scores, and examine the extent of agreement between self-assessments and proxy-assessments. The summary scores being evaluated include: the multi-attribute utility scores from the Health Utilities Index Mark 2 (HUI 2 OUS) and Mark 3 (HUI 3 OUS) (Feeny et al. 1996); the mental and physical component summary scores of the SF-36 (MCS-36, PCS-36) (Ware et al. 1994); and the EQ-5D visual analogue scale (EQ-VAS; a 'feeling thermometer' rated from 0 to 100) and index-based scoring system from York (the EQ-Index) (Dolan, 1997). Thus, there are 5 generic HRQL measures in the study (SF-36, HUI 2, HUI 3, EQ-VAS, EQ-Index) generating 6 summary scores (PCS-36, MCS-36, HUI 2 OUS, HUI 3 OUS, EQ-VAS, EQ-Index). Specifically, the primary objectives of the study were to:

1. to compare the responsiveness of the summary scores of selected generic HRQL measures (SF-36, HUI 2 OUS, HUI 3 OUS, EQ-VAS, EQ-Index) obtained by

patient self-assessment during a longitudinal study of the post-stroke recovery process;

2. to compare the responsiveness of the summary scores of selected generic HRQL measures (SF-36, HUI 2 OUS, HUI 3 OUS, EQ-VAS, EQ-Index) for stroke patients obtained by proxy-assessment;
3. to compare and contrast the ability of self- and proxy-assessed summary scores to capture meaningful change according to several anchor-based criteria;
4. to evaluate the extent of agreement between stroke patient and proxy HRQL assessment, in relation to cross-sectional and change score agreement.

A secondary objective was to examine the interrelationships (bivariate correlations) between the baseline scores on the various clinical and HRQL measures (domain and summary scores) in order to study the construct validity of the measures for self-assessment and proxy assessment in stroke.

3.2.0 Study Design

The study design was a longitudinal natural history study of a cohort panel of stroke survivors and their proxies (preferably family caregivers). The study compared responsiveness of HRQL scores assessed by patient and proxy pairs for 6 months following the stroke event.

3.2.1 Subjects

The individuals for this study included the first 97 consecutive patients who were recruited as part of a cohort of 124 stroke patients. Participants had to have a confirmed eligible stroke, defined by the World Health Organization criteria as of "rapid onset and of vascular origin reflecting a focal disturbance of cerebral function, excluding isolated impairments of higher function and persisting longer than 24 hours" (WHO, 1983). The stroke was confirmed by clinical examination and by one or more of the following: CT (computerized tomography) brain scan, Doppler , electrokardiogram (EKG), echocardiogram, and MRI (magnetic resonance imaging) scan. Further inclusion and exclusion criteria were designed with the intention of enhancing the generalizability of the study sample while minimizing attrition.

Inclusion criteria:

1. Patient must have caregiver who is also willing and able to consent to participate as a proxy respondent in the study, and both patient and caregiver must live within approximately 150 kilometres of Edmonton, Alberta.
2. Patient and caregiver can comprehend the English language in the judgement of the nurse and clinical assessor who recruited the potential participants.
3. Patient and caregiver who are able to, and do, consent.
4. Patient and caregiver are older than 18 years of age.

Exclusion criteria:

1. Patient has life expectancy of less than 6 months for any medical reason, in the judgment of the clinical assessor/recruiter.
2. Patient has history of previous degenerative or space occupying brain disorder.
3. Hemorrhagic or lower brain stem stroke
4. Subarachnoid hemorrhage or transient ischemic attack.
5. Patient is in coma, or with global or Wernicke's aphasia.
6. Patient has history of dementia prior to stroke.

7. Patient and/or caregiver live more than 150 kilometres from Edmonton, Alberta
8. Patient or caregiver are cognitively impaired in judgement of clinical assessor/recruiter.

3.2.2 Recruitment/Consent

Post-acute phase stroke patient-proxy pairs were recruited into the study, typically within 2 weeks of the stroke and no longer than 3 weeks post-stroke. The stroke patients were recruited from one of two hospitals in Edmonton, Alberta: the University of Alberta Hospital (UAH) and the Royal Alexandra Hospital (RAH). Hospitals in the city of Edmonton admit approximately 2000 stroke patients annually. Of these patients, the UAH admits approximately 1050, the RAH admits 300 to 400, and the remainder go to other hospitals or return home (Shuaib, 2001). Operational approval to recruit from the RAH was not obtained until March 10th, 2000; only 6 patients were recruited from the RAH for the study.

The proxy was preferably a family caregiver such as a spouse, significant other, sibling or offspring. If no family were available, a close friend was enlisted to serve as proxy. Professional health care providers with no relation to the patient did not qualify as proxies.

A research assistant with a clinical background (hereafter called the clinical assessor) recruited participants and performed a clinical assessment at baseline and 6 months. The tests included in the clinical assessment were chosen upon review of the validity, reliability, respondent burden, appropriateness to the study sample and in consideration of the objectives of the research. The clinical assessment consisted of the Scandinavian Stroke Scale (SSS-48) (Scandinavian Stroke Study Group, 1985), the modified Rankin Scale (MRS) (van Swieten et al, 1988), the modified Barthel Index (BI) (Granger et al, 1987), and the National Institute of Health Stroke Scale (NIHSS) (Brott et al, 1989). The clinical assessor reviewed the medical charts of the previous day's stroke admissions at both hospitals and screened for possible candidates. A consecutive sampling technique was used to enroll patients meeting the selection criteria. Reasons for exclusion and non-participation were documented.

Upon identification of a potential candidate for inclusion in the study, the clinical assessor conferred with the nursing staff and physicians to better ascertain whether the patient was able to participate in the study. If so, the premise of the study was explained to the patient and caregiver, and participant information was presented to the patient and caregiver at that time if they were interested (see Appendix 1: Study Information and Consent Form). Both the patient and caregiver were asked for consent to participate. If the patient and or caregiver chose not to participate or did not meet the selection criteria, it was documented on standardized form (Appendix 2). Further information on the clinical status of the patient and demographic background on the patient and caregiver were collected on standardized data forms (Appendix 2). Enrollment of 124 patient-proxy pairs took approximately one year.

3.2.3 Measures

The clinical tests include the CES-D, BI, MRS, NIHSS, and the SSS-48 (Appendix 3). The clinical measures were useful for descriptive purposes as well as for construct validity tests. The survey versions of the SF-36, HUI2, HUI3, and EQ-5D used for patient self-assessment (Appendix 4) and proxy assessment (Appendix 5) are described below, as well as approaches to scoring and treatment of missing data.

3.2.3.1 SF-36

The self-completed one-week recall version of the SF-36 was administered to the patient and proxy at baseline, 1 month, 3 month and 6 months. At inception of the study, no official proxy version was endorsed, so a proxy form was derived from the self-completed version of the SF-36 (Appendix 4) to provide a consistent perspective with that of the standard HUI proxy version, which was the proxy's view of the patient's health status (Appendix 5).

The SF-36 was scored with the SAS scoring program from the Medical Outcomes Trust (Ware et al, 1994). The multiple items on most domains of the SF-36 allowed for mean imputation, a feature of the data management and scoring algorithm. When several missing items on a domain prevent use of the standard SF-36 algorithm for imputing a domain score, the domain score was imputed with hot decking. Relationships between the domain interest (with the missing score) and similar items or/and domains from other measures were examined using bivariate correlations, using Pearson's product moment correlation coefficient (on continuous data) or Spearman's rho (ordinal data), depending on the underlying scaling properties of the measures involved. The item/attribute with the strongest correlation was subsequently selected as the matching covariate, and all cases with the same value for the selected covariate comprised the pool from which a randomly selected value was generated for the missing domain score.

3.2.3.2 EQ-5D

The standard Canadian version of the EQ-5D, which is for self-completion with present day recall (i.e. how are you feeling today?), was administered to the patient and proxy at baseline, 1 month, 3 month and 6 months. No official proxy version was available at the inception of the study, so a proxy version was developed for the EQ-5D, with the items were modified to elicit the proxy's view of the patient's health status (Appendix 5). The EQ-5D was scored using the Time Trade Off (TTO) based scoring system derived from research in the United Kingdom (Dolan et al, 1997).

No official missing item algorithm has been recommended for the EQ-5D. Missing items were imputed with hot decking as described in the above section for the SF-36.

3.2.3.3 HUI Mark 2 and Mark 3

Standard 1-week recall versions of the HUI 2 and HUI 3 questionnaire (15 item) for self-completion were administered to the patient and proxy (HUI23S1.15Q/ HUI23P1.15Q, respectively) at baseline, 1 month, 3 month and 6 months. The HUI Mark 2 and Mark 3 were scored based upon the recommended algorithms for determining the HUI2 (Feeny et al, 1996) and HUI3 (Furlong et al, 1998) health status classification levels, health states, single-attribute level utility scores and overall health-related quality of life utility scores. Permission to use the SPSS syntax file that contains the code for generating the aforementioned scores was obtained and the code validated using a sample data set (Furlong, 2000).

No official missing item algorithm has been recommended for the HUI2 and HUI3. Again, missing items were imputed with hot decking as described in the above section on the SF-36. Missing values for hearing, speech, cognition, and vision were hot decked matched on the values of the attribute for a previous or successive wave of data collection for the same person. This was necessary because these attributes were not strongly correlated (r <0.35) with any other items.

3.2.3.4 CES-D

The CES-D is scored by simple summation. Simple mean imputation was used for item non-response on the CES-D because all items pertain to depression.

3.2.3.5 Clinical Measures

The NIHSS, Barthel Index and SSS-48 are all scored by simple summation. No imputation for these measures was conducted if the follow-up at 6 months was missed.

3.2.4 Measurement/Data Collection

The patients and caregivers were administered questionnaires upon enrollment, and at 1 month, 3 months, and 6 months post-baseline. The timing of data collection points was chosen based upon literature on the recovery of stroke survivors (Kelly-Hayes et al, 1989; Dombovy et al, 1987; Duncan et al, 1997, Lai and Duncan, 1999). Neurological and functional recovery reportedly occurs most rapidly in the first 1 to 3 months after a stroke, but some patients continue to progress after that time, especially with respect to language and visuospatial functions (Kelly-Hayes et al, 1989). The level of disability appears to remain relatively constant from 6 months through 5 years of observation (Dombovy et al, 1987). The patient and proxy surveys given at 1 month, 3 months and 6 months also contained a global health change question with 15 response options (Jaeschke et al, 1989).

Upon enrollment, the Bamford classification for the patient's stroke was extracted from the patient chart by the clinical assessor. Clinical tests (NIHSS, SSS-48, BI, MRS) were performed on the patient by the clinical assessor, followed by administration of the surveys. The order of the HRQL instruments in the survey was not randomized. The order of the families of instruments in the survey was as follows: the HUI2/3 questionnaire, EQ-5D, CES-D, SF-36, and finally, the global change question (after baseline). Clinical tests were performed at baseline enrollment and at 6 months. At the 6-month follow-up, the clinical assessor evaluated the patient's health status in relation to baseline using a similar format to the 15-response option global health change question posed to the patient and caregiver.

Two additional research assistants (RAs) conducted the follow-up visits at 1 month and 3 months. The role of the RAs was to ensure the surveys were completed at the designated times of follow-up and to reiterate the conditions under which the surveys were to be completed. Whenever possible, the RAs would contact the patient and proxy, arrange appointments and visit the patient and proxy to oversee the completion of the questionnaire. The patient and proxy were requested not to discuss the items with one and other. The RAs were trained so as to take advantage of the strengths associated with the presence of interviewers (e.g. minimizing missing data due to inability to read or physical impairment), but only intervening if necessary. Otherwise, their primary function was to ensure the patient and proxy understood and completed the survey. Thus, all surveys were completed by self-assessment, but the research assistants were permitted to assist in the completion of the surveys if the respondent was otherwise unable to do so. All assistance by research assistants was documented. The collection of self-reported data via research assistants who visited the patients face-to-face was intended both to minimize the number of voluntary drop-outs from the study, and to ensure the surveys were completed under the appropriate conditions.

If the patient or proxy lived more than 50 kilometres outside of Edmonton, they were asked if they felt comfortable completing the survey on their own for the 1-month and 3-month follow-ups, and mailing it back to the project co-ordinator. The RAs contacted the mailout participants to inform them that the survey was in the mail, and followed-up to ensure the surveys had been received, completed and in the mail if not received by the project co-ordinator within 2 weeks of the mailout. Due to personal time constraints such as shiftwork, several caregivers opted to complete the surveys themselves and mail them back in a stamped, self-addressed envelope provided by the study.

3.2.5 Sample Size Considerations

Scenarios for sample size requirements were calculated for each generic measure based upon MCIDs, using the conventional probability of error levels of alpha = 0.05 (two-tailed) and beta = 0.20. A cohort of 315 participants was theoretically required to detect the CID of 0.03 (Feeny, 1998) on the HUI 2 and HUI3, using a HUI Mark 2 multi-attribute utility score standard deviation of 0.19 from the stroke literature (Mathias et al, 1997).

For the EQ-5D, any change in level and therefore score for an individual could be considered potentially important (Dorman et al, 1998). Based on an CID of 0.036, the smallest co-efficient from the UK scoring system (Dolan, 1997), and a standard deviation of 0.18 for self-completing stroke patients (Dorman et al, 1998), the sample size required for the EQ-5D was 197.

The definition of important change on the SF-36 is more controversial, as some investigators consider a 5-point difference on any one of its domains to be of potential importance (Ruta et al, 1994). However, the developers do not commit themselves to a MCID.

The power of the study was determined for a series of mean score differences (e.g. between patient and proxy scores) using the actual standard deviations for the scores obtained in the study (see section 4.2.3) at an alpha level of 0.05. A statistical power analysis was conducted to assess the likelihood that a particular test of statistical significance would be sufficient to reject a false null hypothesis (e.g. a null hypothesis that there is no difference in mean scores when comparing self-assessed and proxy assessed scores) for several sample sizes. For the scenarios in Table 1, an independent samples t-test was used to test for a difference between means, which assumes the scores of each group of scores are normally distributed. Power A is the power for the present analysis, power B is the power of the full study sample, and power C is the power of the original study design.

Table 1: Sample Size Scenarios

Instrument	Std Dev	Difference in Mean Score	Power A N = 97	Power B n =124	Power C n =160
EQ-VAS	18	10	0.97	0.99	0.99
		7	0.77	0.86	0.94
		5*	0.49	0.59	0.79
EQ-Index	0.36	0.1	0.49	0.59	0.70
		0.07	0.27	0.33	0.41
		0.05	0.16	0.19	0.24
		0.036*	0.07	0.08	0.09
HUI 2	0.18	0.1	0.97	0.99	0.99
		0.07	0.77	0.86	0.94
		0.05	0.49	0.59	0.79
		0.03*	0.21	0.26	0.32
HUI 3	0.30	0.1	0.63	0.74	0.84
		0.07	0.36	0.45	0.55
		0.05	0.21	0.26	0.32
		0.03*	0.11	0.12	0.14
SF-36	12	10	0.99	0.99	0.99
		7	0.98	0.99	0.99
		5*	0.82	0.90	0.96

*assumed to be CID

3.2.6 Ethical Considerations

The Health Research Ethics Review Board at the University of Alberta approved this study. Every effort was made to ensure that participants were informed about the study, that they were able to withdraw from the study at any time without consequence, and that participant confidentiality was protected. Upon entry, the participants received a Study Information Sheet outlining the nature, procedures, risks and benefits of participating in research of this nature, which involved the completion of questionnaires, and a brief clinical assessment of the patient at baseline and a 6-month follow-up (Appendix 1). Written consent was obtained from each participant. Patient confidentiality was maintained by assigning a patient id number. All documents identifying the participants are stored in a locked cabinet and will be destroyed after 7 years following the completion of the study. Results were only presented in an anonymous, aggregate format.

3.3.0 Data Analysis

3.3.1 Data Preparation

Study data were entered into SPSS Version 10.0.7 and re-entry (verification) was done by the Population Research Laboratory at the University of Alberta. The analysis of the responsiveness of HRQL measures was performed on two versions of the data. To evaluate longitudinal validity, a data set with imputed missing values on non-item responses was created using the hot-decking approach. An unmodified version of the data (the data set without hot-decked items) was used to evaluate agreement between self- and proxy-assessments. Analysis of agreement between patient self- and proxy-assessed scores was performed on the data prior to non-item response imputation because some missing items were hot-decked using the paired respondent's value for certain variables as a best match.

Non-item response imputation was accomplished by creating bivariate correlational matrices for self- and proxy-assessed scores at each time period, similar to the matrices for construct validity. The item/attribute/domain was identified as missing in the data, and a logical covariate with the strongest correlation was selected as a match. The specific missing items that were hot decked and the matching covariates were documented (Appendix 10) for each measure.

3.3.2 Descriptive Statistics

The characteristics of the patients and proxy in the cohort were described in terms of sex, age, relationship to patient and type of stroke (Bamford classification). Presence of statistically significant differences between clinical measure scores at baseline and 6 months were tested using paired t-tests (NIHSS, SSS-48, BI) and Chi-squared tests for the ordinal ratings of the MRS. Participant retention and time required to complete the questionnaires were summarized for each time point. The characteristics of patients for with completed data (for all four data collection time points) were compared to those patients who did not have complete data (those with one or more non-unit responses). An ANCOVA, adjusting for baseline scores, was used to test the null hypothesis of no systematic differences between the research assistants who administered the survey.

3.3.3 Longitudinal Construct Validity

Longitudinal construct validity of the summary scores of each HRQL measure was evaluated by examining sensitivity and responsiveness of the scores. Analyses of longitudinal construct validity were performed on data hot-decked for item non-response. Participants who died were excluded from the analysis; MCS-36, PCS-36 and EQ-VAS scores could not be determined for those who died.

Four different criteria were developed that categorized the health status of patients as having declined, improved, or not changed. Criteria A and B focused on change between baseline and 1 month; criteria C and D categorized the health status of patients as having changed between baseline and 6 months. Each criterion used was independent of the summary scores (i.e. was an external anchor) and each classified the patients differently.

Criterion A was based on the patients' self-assessed responses global health transition question (has there been any change in your health since the last survey?). A score of -1, 0, or +1 on the global health transition question indicated no change; scores of -2 or less were equated with a decline in health, and a score of +2 or more was considered as improved (Juniper et al, 1996; Juniper et al, 1997). Criterion B required that both the patient and proxy assessments agree that change had occurred and on the direction of change, again based on the above criteria for of change groups.

For criterion C, the clinical assessor was asked to evaluate the extent of change in overall patient health between baseline and the 6-month follow-up, using the 15-response option global health change question. The same cutoffs applied for criteria A and B to categorize patients as improved, declined or no change were used.

Criterion D categorized patients into change groups based on a transition in severity of disability between baseline and month 6 using categories derived from Barthel Index scores. Barthel index scores group patients by stroke severity as follows: mild ($\geq$ 85), moderate ($\geq$ 60 but <85), and severe (< 60). A score of 85 or better has been demonstrated to correspond with independence requiring minimal assistance in studies, although several trials have used a more arbitrary Barthel Index-based score cutoff of 95 (Sulter et al, 1999).

These 4 sets of criteria formed the basis for evaluating sensitivity and responsiveness (effect size (ES), standardized response mean (SRM), and Guyatt's responsiveness statistic (GRS)) of the summary scores of the HRQL measure. Criteria A, B and C were based on methods similar to the approaches used by Juniper et al (1996) and Juniper et al (1997). Criterion D used an independent clinical measure of functional ability to benchmark clinically important change.

3.3.3.1 Sensitivity

Sensitivity was examined by performing tests of statistical significance on the pre- and post-scores of patients who were categorized as changed (improved or declined) according to criteria A, B, C and D. Because responsiveness refers to the ability to detect meaningful change, both improvement and deterioration, absolute values of the change scores of patients were used. Paired t-tests were used to detect statistically significant differences between the time periods (for criteria A and B, change between baseline and 1 month; for criteria C and D, change between baseline and 6 months). The relative sensitivity of the summary scores of the HRQL measures was examined by the ratio of squared t-statistics (Liang et al, 1985), a method comparable to F-statistic ratios used in other studies of responsiveness (Sneeuw et al, 1997b; Birback et al, 2000). While the magnitude of the t-statistic cannot be equated with magnitude of effect, it provides some insight into the power of each summary score to detect statistically significant differences: "the relative validities are equivalent to the ratio of sample sizes that would be required to detect the known group difference using one measure versus the other" (Hays et al in Staquet, 1998, pg. 177). The EQ-VAS was arbitrarily chosen as the reference comparator set equal to one.

3.3.3.2 Responsiveness – Patient Self-Assessment

To compare the ability of each HRQL summary scores to detect change, several statistics were used: effect size, standardized response mean (SRM), and Guyatt's responsiveness statistic (GRS) (Birbeck et al, 2000). As described by Hays et al in Staquet (1998), effect size was calculated as the ratio between mean change scores (D) and the standard deviation of baseline scores (SDbl), expressed as ES = D/SDbl. The SRM was calculated as a ratio of mean change scores to the SD of the change scores (SDch) across the time period of interest (SRM = D/SDch). GRS was calculated as a ratio of the mean change scores on a measure to the standard deviation of the change score among stable subjects (SDst) (GRS = D/SDst). The denominator for GRS in this study was derived from the standard deviation of the change scores of patients were identified as 'no change' according to both patient and proxy global change ratings between months 3 and 6.

To compare responsiveness across the summary scores, the magnitude of each responsiveness statistic was ranked. The 6 summary scores were ranked relative to each other, and a median ranking (from the 3 indices of responsiveness) was generated for each summary score. This was performed for each of the anchor-based criterion.

Responsiveness was first evaluated on a more aggregated level, using the 4 criteria to categorize patients as 'changed' or 'not changed'. The responsiveness statistics focused on patients who were categorized as 'improved or declined'. Independent samples t-tests were performed to compare mean change scores between patients categorized as 'changed' to those categorized as 'not changed' according to each criterion. The 'changed' group was expected to have significantly greater change scores than the 'no change' group.

Responsiveness was further examined by subdividing the patients categorized as changed into 'improved' or 'declined', so as to determine the magnitude of change differed by change sub-group. A minimum of 10 patients was arbitrarily chosen for this sub-analysis.

3.3.3.3 Responsiveness - Proxy Assessments

An approach to estimating responsiveness for the different indices described in the previous section was similarly applied in the evaluation of responsiveness for summary scores from proxy assessments. The same four criteria were used to categorize the health status of patients as having declined, improved, or not changed.

3.3.4 Differences between Self- and Proxy-Assessments Over Time

Statistically significant differences between patient self-assessed and proxy-assessed change scores for patients grouped according to each criterion were analyzed using paired t-tests. This analysis was performed for the purpose of determining whether the differences in change scores were dictated by the external criterion. Differences between self- and proxy-assessment were also evaluated using repeated measures analysis of variance (RM ANOVA). A statistically significant interaction effect between time and type of assessment (self- or proxy-assessment) would reject the null hypothesis. RM ANOVA must first satisfy the assumptions underlying ANOVA. ANOVA is based the assumptions that independent random samples have been taken from each population; that the distributions are normal; and that the population variances are all equal (Norusis, 2000).

In a repeated measures design following one group, the homogeneity of variance assumption is called the assumption of sphericity, which states that the variances of change scores will be relatively equal and correlated with each other (Portney and Watkins, 2000). Because the repeated measures test examines correlated scores across treatment conditions, it is especially sensitive to variance differences, biasing the test in the direction of type 1 error (wrongly rejecting the null hypothesis). Mauchly's test of sphericity is performed to determine if an adjustment to the value of p is needed to account for possible violations of sphericity. The correction is made by decreasing the degrees of freedom, thereby increasing the critical value of the F statistic, thus compensating for bias towards a type 1 error (Portney and Watkins, 2000). If Mauchly's test of sphericity is not significant, sphericity can be assumed. Otherwise, a correction factor is applied ("epsilon"); versions of epsilon include the "Greenhouse-Geisser" and the more conservative "lower-bound".

3.3.5 Patient Self-Report and Proxy Agreement

The relationship between patient self-assessment and proxy assessment was examined using (1) Pearson's r; (2) one-way random effects model ICC based on

ANOVA (case 1 from Shrout and Fleiss, 1979); (3) the ICC described by Robinson (1957); (4) a paired t-test on mean difference scores between self- and proxy assessment (upon testing for equality of variance). These statistics were performed on the cross-sectional and change scores for each of the HRQL summary scores (HUI2 OUS, HUI3 OUS, MCS-36, PCS-36, EQ-5D VAS, and EQ-5D Index). The single rater reliability for the one-way random effects model is more applicable for the purposes of this study, because only one proxy assessment is likely to be sought when patient self-assessment is not possible. Analysis of agreement was performed only for patient and proxy who had completed responses for a measure; those respondents who required imputation for non-item response were not included in the analysis of agreement.

The statistics were calculated for cross-sectional scores at baseline, month 1, month 3, and month 6, and for change scores between baseline and 1 month (t_{0-1}), 1 month and 3 months (t_{1-3}), 3 months and 6 months (t_{3-6}), and baseline and 6 months (t_{0-6}).

The findings were interpreted in the context of previously literature on acceptable levels of reliability. An acceptable standard of reliability for measures at the group level in clinical trials is 0.70 (Hays et al, 1993). A guideline for strength of agreement to interpret levels of clinical or practical significance for interrater reliability generalizability coefficients such as the ICC (Cicchetti and Sparrow, 1982), which draws heavily from previously published work (Landis and Koch, 1977), is: poor (less than 0.40); fair (0.41 to 0.59); good (0.60 to 0.74); and excellent (0.75 or over).

3.3.6 Cross-Sectional Construct Validity

The construct validity of the responses by self-report and proxy respondents was evaluated by testing hypothesized relationships between the instruments. The intention of this exercise was to confirm that the results make sense. Correlational strength between intra- and inter-instrument domains and the between the summary scores each of measure was tested and presented in a matrix. Substantial deviance from the hypothesized correlations between measures and domains may indicate potential problems with the study design, sample, or data collection that would require further scrutiny. The clinical measures serve as an external source of validation criteria. The construct validity was examined for self-assessed scores at baseline, and for proxy-assessed scores at baseline.

When the variables are continuous and normally distributed, it is appropriate to use Pearson product-moment correlation. For analysis of ordered categorical variables such as the domains of the EQ-5D, Spearman's rho is an appropriate statistical technique (Hulley and Cummings, 1988). The criteria for interpretation of strength of correlation between the scores are: absent (less than 0.2), weak (0.2 to 0.34), moderate (0.35-0.5) and strong (greater than 0.5) (Juniper et al, 1996; Hillers et al, 1994). These criteria have previously been applied by researchers who have extensive experience with HRQL instruments who have found that in validating HRQL measures where there is no criterion standard, correlation with related measures much greater than 0.5 are very seldom observed (Hillers et al, 1994).

In general terms, items and attributes from each HRQL measure that relates to physical functioning should be more strongly correlated with each other than with items/attributes that pertain to mental health, and vice versa. Specific relationships are stated in the hypotheses (section 3.1.3). Comparisons of the HRQL scores of stroke patients to those of the general population are used for evidence of construct validity. Canadian normative data for the SF-36 health survey is available (Hopman et al, 2000).

Overall and single attribute scores for the HUI 3 system have been reported for large-scale Canadian population health surveys (e.g. the 1996-97 National Population Health Survey). Using the 1990 Ontario Health Survey, Grootendorst et al (2000) reported scores for stroke survivors living in the community. Furthermore, the National Population Health survey (NPHS) pre-tested the HUI3 and EQ-5D in a sample of 1,477 Canadian respondents in 1998 and reported the distribution of responses for each generic measure (Statistics Canada, 2000). While the pre-test sample selected was not designed to be representative of the Canadian population, it may provide a reasonable approximation of Canadian population-based norms for the EQ-5D.

3.4.0 Hypotheses

3.4.1 Hypotheses: Sensitivity

All of the summary scores are expected show statistically significant change (i.e. demonstrate sensitivity) because important clinical change is generally recognized to occur during the stroke recovery process, both mentally and physically (AHCPR, 1998; Jorgensen et al, 1995; Kelly-Haynes et al, 1989; Dombovy et al, 1987). Statistically significant change is expected for mean change scores between baseline and 1 month (patients categorized as changed according to criteria A and B) and for mean change scores between baseline and 6 months (patients categorized as changed according to criteria C and D).

3.4.2 Hypotheses: Responsiveness

The following hypotheses pertained to both self- and proxy-assessed scores, and are designated as general responsiveness hypotheses. Aspects of responsiveness for which the self-assessed scores are expected to diverge from the proxy-assessed scores are explicitly addressed (subsections 3.4.2.1 and 3.4.2.2).

For both the self- and proxy assessed scores, mean change scores are hypothesized to be in the direction that corresponds with its category of change. Patients classified as 'improved' are expected to have a positive change score was expected. Patients classified as 'declined' are expected to have a negative change score. Patients in the 'no change' group are expected to have smaller mean change scores than patients the 'improved' or 'declined' group.

Specific aspects of responsiveness for the 4 criteria that may differ for self- and proxy-assessed summary scores are described in the subsections that follow.

3.4.2.1 Hypotheses: Responsiveness of Patient Self-Assessments

Hypotheses relating to the four different criteria for patient self-assessed scores are as follows:

Criterion A (patient rates self t_0/t_1): responsiveness indices derived from change scores based on self-assessed summary scores for criterion A should not deviate from the general hypotheses, because the patients classified themselves. Inconsistencies are cause to question the reliability of criterion A as a basis for categorizing patients as improved, no change, or declined in overall health status.

Criterion B (patient and proxy agree patient changed t_0/t_1): this criterion is more stringent than criterion A, because both patient and proxy must agreement whether the patient changed or not. Fewer patients will be exclusively categorized as improved, no change, or declined. Criterion B is anticipated to generate findings more consistent with the general hypotheses than criterion A. Thus, larger magnitudes of change are expected for the self-assessed summary scores using criterion B than for criterion A.

Criterion C (clinical assessor rates patient t_0/t_6): the direction and magnitude of change is expected to be consistent with the general hypotheses. The MCS-36 and EQ-VAS are the most probable to deviate from these expectations due to the limited contact the clinical assessor had with the patients.

Criterion D (Barthel Index category change t_0/t_6): the PCS-36 scores are expected to be more responsive than the other summary scores, especially the MCS-36, for criterion D. This is because the criterion D anchor, the Barthel Index, is a measure of physical independence.

The responsiveness statistics examined patients who were categorized as 'improved or declined' to the exclusion of the patients categorized as 'no change' based on the 4 criteria. Mean difference summary scores are expected to be greater in the 'change' group than in the 'no change' for each of the criteria.

3.4.2.2 Hypotheses: Responsiveness of Proxy-Assessments

Hypotheses relating to the four different criteria used to evaluate the responsiveness of proxy-assessed summary scores using the 4 criteria were as follows:

Criterion A (patient rates self t_0/t_1): Because this criterion is based on patient self-classification, the responsiveness indices derived from change scores based on proxy-assessed summary scores may deviate from the general hypotheses on magnitude of change and even direction of change (if the change group is small), because the proxy perspective may disagree with the patients' response to global health change question.

Criterion B (patient and proxy agree patient changed t_0/t_1): because the proxy's assessment of patient global health change formed part of the criteria, the proxy-assessed summary score were expected to be consistent with the direction of change. The responsiveness indices for this criterion for proxy assessments should be consistent with the general responsiveness hypotheses, and larger magnitudes of change are expected relative to criterion A.

Criterion C (clinical assessor rates patient t_0/t_6): while the direction and magnitude of change is expected to be consistent with the general responsiveness hypotheses, deviations may occur from these expectations because of the limited contact the clinical assessor had with the patient. The proxy-assessed scores may not concur with the clinical assessor's directional assessment of global health change.

Criterion D (Barthel Index category change t_0/t_6): The PCS-36 scores were expected to be more responsive than the other summary scores, especially the MCS-36, for criterion D. The responsiveness indices for the proxy-assessed scores are expected to be comparable to those for patient self-assessment, and because the criterion is based on observable functionality, and may even be greater for proxy assessment scores.

In comparing the mean change scores for the 'change' group to the 'no change' group by each criteria, change scores assessed by proxy are expected to be greater in the 'change' group than in the 'no change' for each of the criteria. Possible exceptions include criterion A (patient rates self), and criterion C, because the proxy-assessed scores may not be consistent with the patient assessment and/or clinician assessment.

3.4.3 Hypotheses: Comparison of Change by Patient and Proxy

The sensitivity of self- and proxy-assessed HRQL summary scores are expected to differ depending on the criteria used to categorize patients as changed.

Criterion A (patient rates self t_0/t_1): patient self-assessed scores are expected to be more sensitive (larger change scores) because criterion is based upon patient self-rated global scale of change.

Criterion B (patient and proxy agree patient changed t_0/t_1): because both patient and proxy agree that change has taken place, no significant differences between self- and proxy assessed change scores are expected.

Criterion C (clinical assessor rates patient t_0/t_6): proxy-assessed change scores may be larger than patient-assessed change scores because the clinician's perspective may have more in common with the proxy perspective.

Criterion D (Barthel Index category change t_0/t_6): proxy-assessed change scores may be larger than patient-assessed change scores because the ability to function independent may be more obvious to the proxy than the patient, particularly in cases where patient denies of impairment.

In addition to the differences in self- and proxy assessed scores as related in subsections 3.4.2.1 and 3.4.2.2, the overall mean summary scores assessed by proxy are generally expected to be slightly lower than assessments by patient self-report (Sprangers and Aaronson, 1992; Sneeuw et al, 1997a). The null hypothesis that scores do not differ across time by type of assessor (patient self or proxy) is tested for each HRQL summary score using RM ANOVA.

3.4.4 Hypotheses: Patient Self-Report and Proxy Agreement

In terms of agreement between self and proxy assessment, it is hypothesized that:

(1) greater agreement between cross-sectional scores is observed as more time elapses and the patient stabilizes;

(2) greater agreement between self- and proxy assessed scores occurs for cross-sectional summary scores based on more observable domains (PCS scores will agree more than MCS scores);

(3) agreement statistics are poorer for the EQ-VAS, because the EQ-VAS score reflects both the patient's assessment and valuation of health status, whereas the other summary scores reflect the patient's assessment of health status using standard scoring algorithms;

(4) proxy-assessed scores are systematically higher than patient self-assessed scores

(5) greater agreement is observed for the change scores where important change typically occurs according to the clinical literature; this is between baseline and 1 month, and 1 month and 3 months.

3.4.5 Hypotheses: Cross Sectional Construct Validity

The *a priori* hypothesized relationships between clinical tests and HRQL instruments are presented in a correlational matrix (Table 2). Strength of correlation between the scores was interpreted as follows: absent (less than 0.2), weak (0.2 to 0.34), moderate (0.35-0.5) and strong (greater than 0.5) (Juniper et al, 1996; Hillers et al, 1994).

The clinical measures of neurological functioning (NIHSS and SSS-48) are indicators of the severity of the stroke and are expected to be moderate to strongly correlated with the overall summary scores for each HRQL measure. Moreover, the clinical measures designed to evaluate physical functioning (BI, MRS) are predicted to strongly correlate with the physical domain and summary scores of the HRQL measures (e.g. PCS-36). Strong correlations are expected between the CES-D score and MCS-36, as well as between the CES-D and emotion-based domains of the HRQL measures. Domains on the HUI2/3, SF-36, and EQ-5D related to physical functioning are expected to correlate more strongly with each other than with items/attributes relating to mental health, and vice versa. Intervariate correlations were not expected to differ by assessment perspective.

A report by Statistics Canada of the HUI Mark 3 and EQ-5D forms the basis for some of the predicted correlations (Statistics Canada, 2000). Correlations between EQ-5D and HUI3 attributes (level) were absent or weak between most domains, with the exception of the following: HUI3 ambulation was moderately correlated with EQ-5D mobility, self-care, and usual activities; EQ-5D anxiety/depression moderately correlated with HUI3 emotion; and the HUI3 attribute of pain was strongly correlated with the EQ-5D domains of mobility, usual activities, and pain/discomfort, and moderately correlated with self-care. Similar results were observed for Pearson correlations based on scores. Strong Pearson correlations were reported between the overall scores of the HUI3, EQ-5D Index, and EQ-VAS scores ($0.60 < r < 0.70$).

Further evidence of construct validity can be observed by comparing the HRQL scores of stroke patients to those of the general population. Summary scores reported for stroke patients by self-report and proxy were expected to be lower than those of the general population. This hypotheses is tested using Canadian normative data for the SF-36 health survey (Hopman et al, 2000), the HUI 3 system (the 1996-97 National

Population Health Survey); and for the EQ-5D, compared to the distribution of responses in a sample of 1,477 Canadian respondents in 1998 (Statistics Canada, 2000). The overall and single attribute scores for the HUI 3 are compared to scores of stroke survivors reported in an Ontario-based study of community dwelling stroke survivors (Grootendorst et al, 2000).

In general, stroke-related morbidity manifests itself through motor dysfunction, problems with language and communication, visual-spatial dysfunction, depression, decreased social functioning, cognitive impairment (Kwa et al, 1996), and pain (Grootendorst et al, 2000). Therefore, attributes on the various measures that relate to ambulation, mobility, language, cognition, emotion, pain, sensation, dexterity, social functioning and variations on these domains were expected to be lower among stroke patients than for Canadian population-based norms.

Table 2: Hypothesized Strength of Correlation (Baseline Scores)

		PF	RP	BP	GH	VT	SF	RE	MH	PCS	MCS	SAV3	SAH3	SAS3	SAA3	SAD3	SAE3	SAC3	SAP3	OUS3	SAS2	SAM2	SAE2	SAC2	SAT2	SAP2	OUS2	VAS	INDX	MRS	NIHSS	SSS	BI	CESD
S	PF																																	
F-	RP	S																																
3	BP	W	M																															
6	GH	W	W	W																														
	VT	M	W	M	W																													
	SF	M	W	W	M	W																												
	RE	W	W	S	M	M	S																											
	MH	A	A	W	W	M	M	S																										
	PCS																																	
	MCS									A																								
H	SAV3	A	W	A	A	A	A	A	A																									
U	SAH3	A	A	A	A	A	A	A	A			A																						
I	SAS3	A	A	A	A	A	A	A	A			A	A																					
3	SAA3	S	M	A	W	W	M	A	A			A	A	A																				
	SAD3	M	A	A	A	A	A	A	A			A	A	A	W																			
	SAE3	W	W	W	W	M	M	S	W			A	A	A	A	A																		
	SAC3	A	A	A	A	A	A	A	M			A	A	A	A	A	W																	
	SAP3	A	A	S	W	W	A	M	W			A	A	A	A	A	A	A																
	OUS3									S	M																							
H	SAS2	A	A	A	A	A	A	A	A			S	S	S	A	A	A	A	A															
U	SAM2	S	S	W	M	W	W	W	W			A	A	A	S	S	A	A	A		A													
I	SAE2	A	A	W	W	M	M	S	M			A	A	A	A	A	S	A	M		A	A												
2	SAC2	A	A	A	A	A	A	A	A			A	A	A	A	A	A	S	A		A	A	A											
	SAT2	S	S	W	M	W	M	W	W			A	A	A	M	M	A	A	A		A	M	A	A										
	SAP2	A	A	S	A	A	A	W	A			A	A	A	A	A	A	A	S		A	A	W	A	A									
	OUS2									S	M									S														
E	VAS									S	M									S							S							
Q	INDX									S	M									S							S	S						
C	MRS	S	S	A	A	A	W	A	A	S	W	W	A	A	S	W	A	W	A	S	A	S	W	A	S	A	S	S	S					
L	NIHSS	S	S	A	A	A	W	A	M	S	W	W	A	A	M	W	A	W	A	S	A	S	W	W	S	A	S	S	S	S				
I	SSS	S	S	A	A	A	W	A	A	S	W	W	A	A	M	W	A	W	A	S	A	S	W	W	S	A	S	S	S	S	S			
N	BI	S	S	A	A	A	W	A	A	S	W	A	A	A	S	W	A	W	A	S	A	S	W	A	S	A	S	S	S	S	S	S		
	CESD	A	A	W	A	A	S	S	S	W	S	A	A	A	A	A	S	W	W	S	A	W	S	A	M	W	S	S	S	S	S	S	S	

Correlation:
A: less than 0.2;
W: weak (0.2 to 0.34), M: moderate (0.35-0.5); S: *strong* (> 0.5)

Hypothesized correlations between the domains of the EQ-5D and the domain and summary scores of the other HRQL measures are separately postulated for patient assessments at baseline (Table 3).

Table 3: Hypothesized Correlations for EQ-5D Domains (Baseline Scores)

	Mobility	Self Care	Usual Activities	Pain/ Discomfort	Anxiety/ Depression
SF-36					
PF	S	S	S	A	A
RP	M	M	W	A	A
BP	A	A	W	S	A
GH	A	A	W	W	A
VT	A	A	W	W	A
SF	M	M	M	W	M
RE	A	A	A	W	M
MH	A	A	A	A	A
PCS	S	S	S	W	A
MCS	W	W	A	A	S
HUI 3					
SAV3	A	A	A	A	A
SAH3	A	A	A	A	A
SAS3	A	A	A	A	A
SAA3	S	S	S	A	A
SAD3	S	S	S	A	A
SAE3	A	A	A	A	S
SAC3	A	A	A	A	A
SAP3	A	A	A	S	A
OUS3	S	S	M	M	M
HUI 2					
SAS2	A	A	A	A	A
SAM2	S	S	M	A	A
SAE2	A	A	A	A	S
SAC2	A	A	A	A	A
SAT2	S	S	S	A	A
SAP2	A	A	A	S	W
OUS2	S	S	S	M	M
EQ-5D					
EQ-VAS	S	S	S	M	M
EQ-INDX	S	S	S	M	M
Clinical					
MRS	S	S	S	A	A
NIHSS	S	S	S	A	A
SSS	S	S	S	A	A
BI	S	S	S	A	A
CESD	W	W	W	W	S

Correlation: A=less than 0.2; W=weak (0.2 to 0.34); M=moderate (0.35-0.5); S=strong (> 0.5)

CHAPTER 4: RESULTS

This chapter presents the results of a longitudinal natural history study of self-reported and proxy assessed HRQL. First, characteristics of the participants are described. Patient and proxy-assessed clinical and HRQL summary scores for each time period are then described. Statistics for sensitivity and responsiveness of stroke patient self-assessment are presented, followed by the findings for proxy assessment. Finally, extent of agreement between self- and proxy-assessment and construct validity are presented.

4.1.0 Study Sample

Ninety-seven patient and proxy pairs were recruited into the study between October 1999 and June 2000.

4.1.1 Sample Recruitment

A total of 556 patients were reviewed as potential candidates for recruitment into the study between October 15th 1999 and September 20th 2000. Of these, 356 patients did not meet the selection criteria (64.0%), including 29 patients who did not have a suitable caregiver (family or friend). Of the 200 patients who met the selection criteria, 54 of these patients were not interested or unwilling to participate (27.0%) and 22 patients did not participate due to caregiver reluctance (on behalf of the patient or themselves) (11.0%). A total of 124 patients and proxy pairs (62.0% of those eligible) were recruited into the study between October 15th 1999 and September 20th 2000. Analysis was performed on the 97 patient-proxy pairs who completed the 6-month follow-up by January 15th, 2001.

Of the 97 patient-proxy pairs, 76 patients and 77 proxy respondents completed the 6-month assessment (78.4% and 79.4% retention, respectively) (Table 4). Six patients died during the 6-month follow-up period. If the patient or proxy was unable to complete the survey due to limitations imposed by their own health, they were classified as a not missing at random (NMAR) unit non-response. These classifications were decided via discussion within the project team on a case-by-case basis. If their health did not prevent them from completing the survey but they chose to withdraw from the study or simply not complete the survey for a specific data collection point, it was assumed to be missing at random unit non-response (MAR). The time required to complete the surveys was similar across time periods for patients, requiring approximately 30 to 35 minutes. The time required by the proxy respondents to complete the surveys was also relatively invariant across the data collection periods, requiring approximately 20 to 25 minutes per survey.

Table 4: Respondent Retention and Time Required for Survey

Respondent	**Respondent Retention (number)**			
	Baseline	**Month 1**	**Month 3**	**Month 6**
Patient				
Completed	97	86	79	77
Missing at Random (MAR)	0	7	8	12
Not Missing at Random (NMAR)	0	3	6	3
Dead	0	1	4	6
Time Required for Survey (in Minutes: Mean, SD)	33.9 (11.6)	35.4 (17.3)	32.1 (16.9)	32.4 (11.2)
Proxy				
Completed	97	84	79	76
Missing at Random (MAR)	0	9	10	12
Not Missing at Random (NMAR)	0	3	4	2
Dead	0	1	4	6
Time Required for Survey (in Minutes: Mean, SD)	25.2 (9.7)	25.2 (11.8)	22.0 (8.5)	23.9 (9.5)

4.1.2 Demographics

Of the 97 stroke patients, 51 were male and 46 female (Table 5). Patients were generally older (69.2 years of age (SD 14.5)) than the proxy caregiver (56.5 years of age (SD 13.7)) (Tables 5 and 6). Almost 50% of proxy caregivers were the patient's spouse, and the more than 2/3 of proxy caregivers were female.

Table 5: Characteristics of Entire Patient Sample at Baseline

Characteristic	[N (%)]
Patient median; mean age (SD), years	73.0; 69.2 (14.5)
Patient sex	
Female [N (%)]	46 (47.4)
Male [N (%)]	51 (52.6)
Bamford Classification [N (%)]	
TACI	9 (9.3)
PACI	52 (53.6)
POCI	26 (26.8)
LACI	9 (9.3)
Number of Previous Strokes [N (%)]	
0	82 (84.5)
1	4 (4.2)
2	9 (9.3)
Don't know	2 (2.1)

Table 6: Proxy Characteristics (Baseline)

Characteristic	[N (%)]
Relationship of Proxy to Patient [N (%)]	
Spouse	46 (47.4)
Daughter	20 (20.6)
Son	14 (14.4)
Sister	5 (5.2)
Brother	2 (2.1)
Friend	10 (10.3)
Proxy median; mean age (SD), years	57.5; 56.5 (13.7)
Resides with Patient	
Yes [N (%)]	44 (45.4)
No [N (%)]	53 (54.6)
Proxy sex	
Female [N (%)]	31 (32.0)
Male [N (%)]	66 (68.0)

Of the four clinically identifiable subgroups of cerebral infarction: 54% were cortical infarcts (partial anterior circulation infarcts, PACI); 27% of infarcts were associated with the vertebrobasilar arterial territory (posterior circulation infarcts, POCI); 9% had anterior circulation infarcts with both cortical and subcortical involvement (total anterior circulation infarcts, TACI); and 9% had infarcts confined to the territory of the deep perforating arteries (lacunar infarcts, LACI). Similar to Bamford's initial classification study (Bamford et al, 1991), the PACI subtype occupied the highest incidence. A smaller proportion of TACI and LACI were observed. The TACI subtype is associated with the highest mortality rate of the 4 types (Bamford et al, 1991), and a poor prognosis may have prevented a greater proportion of this stroke subtype from meeting the selection criteria of this study. Also, a smaller proportion of LACI-type stroke patients participated in this study relative to the Bamford et al study. Although it is erroneous to describe LACI subtypes as 'mild' strokes, they are associated with low case mortality rate (Bamford et al, 1991). LACI type stroke patients are often discharged several hours after presentation to the ER, without admission to the neurology ward where recruitment into the study would take place. Thus, the stroke population in this study could be described as having proportionally fewer LACI and TACI strokes in relation to the total incidence of stroke subtypes.

4.2.0 Descriptive Scores: Clinical and HRQL measures

4.2.1 Clinician Assessed Measures

Mean and median scores improved between baseline and 6 months on the stroke specific measures administered by the clinical assessor (Tables 7 and 8). Lower scores at 6 months versus baseline for the SSS-48 and NIHSS and higher scores on the BI all represented statistically significant improvement in terms of physical functionality and stroke-specific neurological recovery (all paired t-tests: p< 0.001) (Table 7). The proportion of patients categorized as having signs and symptoms of disability according to the MRS also decreased between baseline and month 6 (Wilcoxon signed ranks test, p-value <0.001) (Table 8).

Table 7: Clinical Scores for Stroke Patients

Clinical Measure	Mean	SD	Median	25th percent ile	75th percent ile	Mini-mum	Maxi-mum
Baseline (n=97)							
NIHSS	9.20	4.26	8.00	6.00	12.00	2.00	25.00
SSS-48	30.91	8.96	33.00	25.75	37.00	8.00	46.00
BI	50.63	25.62	45.00	28.00	64.75	3.00	100.00
Month 6 (n=76)							
NIHSS*	4.11	3.93	3.50	1.00	6.00	.00	19.00
SSS-48*	41.13	8.51	44.00	39.00	48.00	16.00	48.00
BI*	84.80	23.01	95.00	76.75	100.00	3.00	100.00

*paired t-test, p-value <0.001

Table 8: Sample by Modified Rankin Handicap Scale Grade

	Baseline (n = 97)		Month 6 (n = 76)*	
MRS Category	Frequency	%	Frequency	%
No symptoms	0	0	15	15.5
No significant symptoms	3	3.1	16	16.5
Slight disability	10	10.3	16	16.5
Moderate disability	20	20.6	14	14.4
Moderately severe disability	53	54.6	12	12.4
Severe disability	11	11.3	3	3.1

*Wilcoxon signed ranks test between baseline and month 6, p-value <0.001

4.2.2 CES-D Scores

The mean average CES-D scores for the cohort of patients who were retained over the 6 months of the study improved in the view of both patients and proxies (Table

9). Self-assessed CES-D scores ranged from 0 (no depressive symptoms) to 46; proxy assessed scores ranged from 0 to 59. The mean and median scores followed a similar pattern: both patient and proxy assessments indicate that fewer depressive symptoms are present at 1 month post-stroke than at baseline, but the burden of depressive morbidity does not change much thereafter. Proxy assessed CES-D mean and median scores were constantly higher than patient self-assessed scores, implying that proxies perceived more depressive symptoms in patients than the patients themselves. Baseline median CES-D scores were above 16 for both patient and proxy-assessment, a cutoff used in the literature to describe potentially depressed subjects. Thus, more than half of the sample had symptoms indicative of clinical depression at baseline.

Table 9: CES-D Assessments by Patient and Proxy

	*CES-D Scores: Patient Assessed				†CES-D Scores: Proxy Assessed			
Statisitic	Base-line	Month 1	Month 3	Month 6	Base-line	Month 1	Month 3	Month 6
n	97	87	79	78	97	83	77	76
Mean	16.77	13.77	12.90	12.25	20.27	15.29	14.83	14.59
Median	17.00	11.00	10.00	11.00	19.00	15.00	12.00	14.00
Std Dev	7.91	10.13	11.02	9.58	11.48	10.38	10.77	11.23
Minimum	.00	.00	.00	.00	.00	.00	.00	.00
Maximum	39.00	46.00	46.00	43.00	52.00	43.16	47.00	59.00

*p <0.01 (F-statistic 6.53 in RM ANOVA) n=71; †p < 0.01 (F-statistic=12.16 in RM ANOVA) n=69

4.2.3 Descriptive Statistics of HRQL Measure Summary Scores

Because parametric statistical tests, such as ANOVA, are predicated on assumption that the distribution of scores resembles a normal distribution, the frequency distributions (histograms) of the summary scores were visually inspected for each HRQL measure (Appendix 6). Of all of the HRQL summary scores, the distributions of EQ-VAS scores most closely resemble a normal distribution. Several distributions of the EQ-Index scores appeared bimodal, and others were slightly skewed to the left. The PCS-36 scores of both patient self-assessment and proxy assessment appeared slightly skewed to the right at baseline, indicative of severe physical morbidity associated with most of the patient sample. MCS-36 scores became more skewed to the left as time progressed, with similar distributions for both patient self-assessment and proxy-assessment. Several of the HUI2 OUS distributions were close to normal; the HUI3 OUS distributions were less so.

The summary scores for each HRQL measure for all respondents have been summarized in tabular (Table 10) and graphical form (Appendix 7). Scatterplots of the relationship between self- and proxy assessed scores were created for both cross-sectional scores (Appendix 8) and change scores (Appendix 9). These plots are further discussed in the context of agreement (Section 4.5.0). Non-item responses were imputed for this group using a hot-decking approach (Appendix 10). Descriptive statistics were summarized for all of the domains and attributes of the SF-36, HUI2/3 and EQ-5D (Appendix 11). An in-depth analysis of the change scores for each of the HRQL measures and the comparability of patient self-assessed and proxy-assessed scores appears in the next sections.

Table 10: Descriptive Statistics for Self- and Proxy-Assessed Summary Scores

	Patient						Proxy					
Time	**n**	**Mn**	**Md**	**SD**	**Min**	**Max**	**n**	**Mn**	**Md**	**SD**	**Min**	**Max**
EQ-VAS												
Baseline	97	61	60	16	61	95	97	51	50	20	10	99
Month 1	83	62	65	19	5	95	83	64	68	19	20	98
Month 3	79	68	70	17	20	100	79	66	70	20	10	95
Month 6	76	70	70	19	20	100	76	70	72	20	0	98
EQ-Index												
Baseline	97	0.23	0.15	0.36	-0.74	0.81	97	0.18	0.14	0.40	-0.74	1.00
Month 1	83	0.51	0.64	0.38	-0.74	1.00	83	0.43	0.46	0.36	-0.48	1.00
Month 3	79	0.57	0.64	0.32	-0.24	1.00	79	0.50	0.59	0.37	-0.48	1.00
Month 6	76	0.59	0.69	0.34	-0.22	1.00	76	0.54	0.61	0.39	-0.74	1.00
PCS-36												
Baseline	97	29	27	9	10	61	97	26	24	8	12	53

Month 1	83	32	31	11	13	60	83	29	27	11	11	58
Month 3	79	34	33	11	15	58	79	33	32	12	13	58
Month 6	76	35	35	12	12	63	76	33	31	13	11	58
MCS-36												
Baseline	97	47	46	11	19	67	97	47	47	12	17	74
Month 1	83	51	53	13	20	74	83	50	53	12	23	67
Month 3	79	52	54	13	19	77	79	49	50	13	16	69
Month 6	76	53	55	10	24	74	76	52	56	12	17	73
HUI2 OUS												
Baseline	97	0.52	0.50	0.19	0.16	1.00	97	0.50	0.47	0.20	0.08	0.95
Month 1	83	0.63	0.67	0.21	0.06	1.00	83	0.59	0.58	0.21	0.08	1.00
Month 3	79	0.64	0.63	0.21	0.16	1.00	79	0.62	0.67	0.24	0.11	1.00
Month 6	76	0.64	0.67	0.23	0.18	1.00	76	0.65	0.64	0.23	-0.02	1.00
HUI3 OUS												
Baseline	97	0.22	0.15	0.30	-0.22	1.00	97	0.18	0.08	0.33	-0.29	0.97
Month 1	83	0.40	0.40	0.35	-0.29	0.97	83	0.33	0.28	0.32	-0.29	1.00
Month 3	79	0.42	0.43	0.33	-0.27	1.00	79	0.37	0.43	0.36	-0.27	1.00
Month 6	76	0.45	0.46	0.35	-0.19	1.00	76	0.42	0.43	0.36	-0.36	1.00

Mn= mean; Md=median

4.2.3.1 Research Assistant Bias

ANCOVA was used to test the null hypothesis there was no systematic differences between the research assistants who administered the survey. The clinical assessor administered the surveys at baseline and 6 months. ANCOVA tests were performed on the summary scores at 1 month and 3 months on the self-assessment and proxy assessment scores, entering baseline scores as a covariate to adjust for baseline differences (Table 11).

ANCOVA requires the same assumptions as ANOVA: independent random samples taken from each population; the populations are normal; and the population variances are all equal (Norusis, 2000, pp. 263). The samples grouped according to research assistant were independent, but the groups were unequal in size. Groups of patient-proxy pairs were followed up (at 1 month, 3 months) by the clinical assessor (n=6; n=4), research assistant AH (n=38; n=35), research assistant AW (n=23; n=19); and mailouts (n=19; n=21). Levene's test for equality of variance was used to examine the equality of variance assumption. When the test failed (p-value <0.05 for Levene's statistic), the non-parametric Kruskal-Wallis test that requires fewer assumptions was used to test the null hypothesis.

Table 11: Statistical Tests on Summary Scores to Detect RA Bias

Summary Score	Levene's Statistic	P-value (sig)	F-statistic	P-value (sig)	†Kruskal Wallis Chi Square	P-value (sig)
Month 1						
Patient						
EQ-VAS	2.45	0.07	0.24	0.87		
EQ-Index	3.53	0.02†	1.60	0.20	5.85	0.12
PCS-36	0.80	0.50	1.27	0.29		
MCS-36	2.43	0.07	3.31	0.02*		
HUI2 OUS	2.87	0.04†	0.81	0.50	3.57	0.31
HUI3 OUS	1.13	0.34	1.05	0.38		
Proxy						
EQ-VAS	2.18	0.10	0.32	0.81		
EQ-Index	0.08	0.97	1.18	0.32		
PCS-36	1.97	0.13	0.22	0.88		
MCS-36	1.32	0.28	1.00	0.40		
HUI2 OUS	0.25	0.87	1.35	0.27		
HUI3 OUS	0.25	0.87	0.58	0.63		
Month 3						
Patient						
EQ-VAS	0.90	0.44	2.70	0.05*		
EQ-Index	5.78	0.001†	1.99	0.12	2.57	0.46
PCS-36	1.31	0.27	1.46	0.23		
MCS-36	2.52	0.06	1.06	0.37		
HUI2 OUS	2.42	0.07	0.25	0.86		
HUI3 OUS	0.63	0.60	0.08	0.97		
Proxy						
EQ-VAS	2.19	0.10	0.72	0.55		
EQ-Index	0.79	0.50	0.95	0.42		
PCS-36	0.81	0.49	0.99	0.40		
MCS-36	0.79	0.50	0.69	0.56		
HUI2 OUS	0.35	0.79	2.56	0.06		
HUI3 OUS	0.30	0.83	1.00	0.40		

†p <0.05 for Levene's statistic: Kruskal Wallis performed in place of ANOVA; *p < 0.05

Of the 24 ANOVA/Kruskal Wallis tests, only 2 tests detected a statistically significant difference at an alpha level of 0.05: the self-assessed MCS-36 at 1 month and the EQ-VAS at 3 months. It is possible that this significant result was due to chance alone as multiple tests were performed, increasing the likelihood that a type 1 error would occur. Applying a Bonferroni correction, no tests were significant. Stratification resulted in small sample sizes with little power to detect a difference between RAs.

4.2.3.2 Unit Non-Respondents

Baseline clinical and demographic differences were compared between patients who had complete HRQL summary scores (after imputation of item non-responses) for the 4 data collection waves to those patients who missed one or more data collection times. The number of respondents for whom item non-response occurred, and the number of missing items that required imputation is detailed in the Appendix 10.

Baseline clinical and demographic differences were compared between the 71 patients who had complete EQ VAS scores across the 4 data collection periods and the 26 patients who did not have a complete set of scores. In addition to the clinical and demographic data, baseline summary scores of each HRQL measure were contrasted for complete case patients versus patients missing 1 or more units on each summary score (Tables 12 to 16). No statistically significant differences were observed between these 2 groups on any of the patient-based demographic or clinical characteristics. No statistically significant differences were detected between completed cases and the group with missing units for the summary scores based on self-assessment as well as for the summary scores generated by proxy-assessment.

Table 12: Complete vs Incomplete Unit Respondents (EQ-VAS)

Characteristic	Complete [n=71]	Missing 1+ units [n=26]
Patient median; mean age (SD), years	73.0; 68.5 (14.6)	74.5; 71.1 (14.5)
Patient sex		
Female [N (%)]	32 (45.1)	14 (53.8)
Male [N (%)]	39 (54.9)	12 (46.2)
NIHSS; mean (SD)	9.3 (4.3)	8.9 (4.4)
SSS-48; mean (SD)	30.7 (8.9)	31.5 (9.3)
BI; mean (SD)	50.2 (26.3)	51.9 (24.1)
MRS [N (%)]		
No symptoms	0 (0)	0 (0)
No significant symptoms	2 (3)	1 (4)
Slight disability	9 (13)	1 (4)
Moderate disability	12 (17)	8 (31)
Moderately severe disability	40 (56)	13 (59)
Severe disability	8 (11)	3 (12)
Bamford Classification [N (%)]		
TACI	9 (13)	0 (0)
PACI	35 (50)	17 (65)
POCI	18 (26)	8 (31)
LACI	8 (11)	1 (4)
Baseline Patient EQ-5D VAS; mean (SD)	61.0 (16.9)	59.4 (14.6)
Baseline Proxy EQ-5D VAS; mean (SD)	50.4 (19.0)	53.5 (21.2)

*no statistically significant differences detected (all p-values>0.05)

Table 13: Complete vs Incomplete Unit Respondents (EQ-5D Classifier)

Characteristic	Complete [n=70]	Missing 1+ units [n=27]
Patient median; mean age (SD), years	72.5; 68.3 (14.6)	75.0; 71.6 (14.4)
Female [N (%)]	32 (45.7)	14 (51.9)
Male [N (%)]	38 (54.3)	13 (48.1)
NIHSS; mean (SD)	9.4 (4.3)	8.8 (4.3)
SSS-48; mean (SD)	30.7 (9.0)	31.4 (9.1)
BI; mean (SD)	50.4 (26.5)	51.3 (23.8)
Baseline Patient EQ index score; mean (SD)	0.22 (0.36)	0.24 (0.37)
Baseline Proxy EQ index score; mean (SD)	0.21 (0.41)	0.12 (0.38)
MRS [N (%)]		
No symptoms	0 (0)	0 (0)
No significant symptoms	2 (3)	1 (4)
Slight disability	9 (13)	1 (4)
Moderate disability	12 (17)	8 (31)
Moderately severe disability	39 (56)	14 (59)
Severe disability	8 (11)	3 (12)
Bamford Classification [N (%)]		
TACI	9 (13)	0 (0)
PACI	35 (50)	17 (63)
POCI	18 (26)	8 (30)
LACI	7 (10)	2 (8)

*no statistically significant differences detected (all p-values>0.05)

Table 14: Complete vs Incomplete Unit Respondents (SF-36 Surveys)

Characteristic	Complete [n=71]	Missing 1+ units [n=26]
Patient median; mean age (SD), years	73.0; 68.5 (14.6)	75.0; 71.1 (14.4)
Patient sex		
Female [N (%)]	32 (45.1)	14 (53.8)
Male [N (%)]	39 (54.9)	12 (46.2)
NIHSS; mean (SD)	9.3 (4.3)	8.9 (4.3)
SSS-48; mean (SD)	30.7 (9.0)	31.5 (9.3)
BI; mean (SD)	50.2 (26.3)	51.9 (24.1)
Baseline Patient PCS-36 score; mean (SD)	29.42 (9.54)	27.45 (7.41)
Baseline Proxy PCS-36 score; mean (SD)	25.63 (9.36)	25.74 (5.24)
Baseline Patient MCS-36 score; mean (SD)	46.79 (11.16)	47.97 (11.98)
Baseline Proxy MCS-36 score; mean (SD)	46.19 (11.72)	47.23 (13.26)
MRS [N (%)]		
No symptoms	0 (0)	0 (0)
No significant symptoms	2 (3)	1 (4)
Slight disability	9 (13)	1 (4)
Moderate disability	12 (17)	8 (31)
Moderately severe disability	39 (56)	13 (50)
Severe disability	8 (11)	3 (11)
Bamford Classification [N (%)]		
TACI	9 (13)	0 (0)
PACI	35 (50)	17 (65)
POCI	18 (26)	8 (31)
LACI	8 (11)	1 (4)

*no statistically significant differences detected (all p-values>0.05)

Table 15: Complete vs Incomplete Unit Respondents (HUI2 Surveys)

Characteristic	Complete [n=71]	Missing 1+ units [n=26]
Patient median; mean age (SD), years	73; 68.5 (14.6)	75.0; 71.1 (14.5)
Patient sex		
Female [N (%)]	32 (45.1)	14 (53.8)
Male [N (%)]	39 (54.9)	12 (46.2)
NIHSS; mean (SD)	9.3 (4.3)	8.9 (4.3)
SSS-48; mean (SD)	30.7 (9.0)	31.5 (9.3)
BI; mean (SD)	50.2 (26.3)	51.9 (24.1)
Baseline Patient HUI2 OUS; mean (SD)	0.525 (0.196)	0.515 (0.156)
Baseline Proxy HUI2 OUS; mean (SD)	0.494 (0.203)	0.500 (0.197)
Patient sex		
Female [N (%)]	32 (45.1)	14 (53.8)
Male [N (%)]	39 (54.9)	12 (46.2)
MRS [N (%)]		
No symptoms	0 (0)	0 (0)
No significant symptoms	2 (3)	1 (4)
Slight disability	9 (13)	1 (4)
Moderate disability	12 (17)	8 (31)
Moderately severe disability	40 (56)	13 (50)
Severe disability	8 (11)	3 (12)
Bamford Classification [N (%)]		
TACI	9 (13)	0 (0)
PACI	35 (50)	17 (65)
POCI	18 (26)	8 (30)
LACI	8 (11)	1 (4)

*no statistically significant differences detected (all p-values>0.05)

Table 16: Complete vs Incomplete Unit Respondents (HUI3 Surveys)

Characteristic	Complete [n=71]	Missing 1+ units [n=26]
Patient median; mean age (SD), years	73; 68.5 (14.6)	75.0; 71.1 (14.5)
Patient sex		
Female [N (%)]	32 (45.1)	14 (53.8)
Male [N (%)]	39 (54.9)	12 (46.2)
NIHSS; mean (SD)	9.3 (4.3)	8.9 (4.3)
SSS-48; mean (SD)	30.7 (9.0)	31.5 (9.3)
BI; mean (SD)	50.2 (26.3)	51.9 (24.1)
Baseline Patient HUI3 OUS; mean (SD)	0.221 (0.311)	0.219 (0.286)
Baseline Proxy HUI3 OUS; mean (SD)	0.167 (0.323)	0.200 (0.355)
Patient sex		
Female [N (%)]	32 (45.1)	14 (53.8)
Male [N (%)]	39 (54.9)	12 (46.2)
MRS [N (%)]		
No symptoms	0 (0)	0 (0)
No significant symptoms	2 (3)	1 (4)
Slight disability	9 (13)	1 (4)
Moderate disability	12 (17)	8 (31)
Moderately severe disability	40 (56)	13 (50)
Severe disability	8 (11)	3 (12)
Bamford Classification [N (%)]		
TACI	9 (13)	0 (0)
PACI	35 (50)	17 (65)
POCI	18 (26)	8 (30)
LACI	8 (11)	1 (4)

*no statistically significant differences detected (all p-values>0.05)

Further investigation into differences in HRQL between dropouts and study participants retained in each successive data cc llection wave was studied by graphing the mean summary scores, with patients grouped according to number of assessments completed (Figures A10.C to A10.H in Appendix 10). The graphs illustrate that those who dropped out immediately after baseline had generally lower baseline scores, with considered clinically important differences in relation to those retained for the entire study on the PCS-36, EQ-Index, HUI2 OUS and HUI3 OUS. Interestingly, patients who were retained for the first 2 data collection waves yet dropped out prior to the third wave had greater scores exceeding CIDs on the EQ-Index and HUI2 OUS, and HUI3 OUS. Hence, there is a possibility of selection bias by restricting the analysis of responsiveness and agreement to patients who responded to every survey unit.

4.3.0 Longitudinal Construct Validity of HRQL Scores

To reiterate the external anchors for evaluation of responsiveness, patients were categorized as experiencing global health change using 4 different criteria. For criterion A, patient self-rated their overall health as improved, did not change, or declined between baseline and 1 month. For criterion B, patient and proxy had to agree that the patient improved or declined to qualify for those categories, again between baseline and month 1. Criterion C was a clinician-based assessment of patient health change between baseline and 6 months. For criterion D, patient change groups were based on movement between BI score based categories between baseline and 6 months.

Table 17: Classification of Patients into Change Groups by Method

	Declined	No Change	Improved	Any Change	Valid N
Criterion A: Patient Self-Rated Change t_0/t_1	12	11	63	75	86
Criterion B: Patient Proxy agree on type of change between t_0/t_1	2	1 agreed; 24 no consensus	55	57	82
Criterion C: Clinician rated change between t_0/t_6	8	7	63	71	78
Criterion D: Stroke severity by BI Category t_0/t_6	0	25	51	51	76

The majority of patients experienced an improvement in overall health according to all 4 criteria (Table 17). For criterion B, there was agreement between patient and proxy that the patient declined in 2 cases; agreement that the patient improved in 55 cases; agreement that patient did not change in 1 case; and 24 cases in which the proxy and patient did not agree that change had occurred. No patients declined in health according to criterion D, which was based on BI score.

4.3.1 Longitudinal Construct Validity – Self Assessment of Health Status

4.3.1.1 Sensitivity – Self-Assessed Scores

Tests of statistical significance (paired t-tests) were performed on the pre- and post summary scores of patients who were categorized as changed (improved or declined) according to criteria A to D (Table 18).

As hypothesized, all self-assessed summary scores were sensitive (p-value < 0.001) using criteria A through D to define patients whose health changed. For all 4 criteria, the HUI2 OUS had the largest squared t-statistic ratio (i.e. was most sensitive) relative to the other summary scores, followed by the HUI3 OUS, (the EQ-VAS served as an arbitrary reference of 1.0). The rank order of magnitude of t-statistics for sensitivity of patient-assessed summary scores for criterion A and B was the same: HUI2 OUS > HUI 3 OUS > EQ-VAS > MCS-36 > EQ-Index > PCS-36. Although criteria A and B pertained to change between baseline and 1 month while criteria C and D related to change between baseline and 6 months, the relative magnitude of t-statistics were similar across the criteria.

Table 18: Comparison of Sensitivity for Self-Assessed Summary Scores

Summary Score	Mean Change Score (Absolute Values)	Std Dev	t-stat*	Squared t-stat ratio	Rank
Criterion A (n=75)					
EQ-VAS	20	14	12.2	1.00	3
EQ-Index	0.37	0.31	10.2	0.70	5
PCS-36	7	7	9.7	0.63	6
MCS-36	11	9	10.8	0.78	4
HUI2 OUS	0.20	0.13	13.1	1.15	1
HUI3 OUS	0.30	0.21	12.4	1.03	2
Criterion B (n=57)					
EQ-VAS	21	15	10.7	1.00	3
EQ-Index	0.38	0.32	8.9	0.69	5
PCS-36	8	7	8.8	0.68	6
MCS-36	12	9	9.8	0.84	4
HUI2 OUS	0.21	0.13	11.8	1.22	1
HUI3 OUS	0.33	0.22	11.3	1.12	2
Criterion C (n=70)					
EQ-VAS	19	15	10.5	1.00	4.5
EQ-Index	0.43	0.34	10.5	1.00	4.5
PCS-36	11	9	9.5	0.82	6
MCS-36	11	8	11.1	1.12	3
HUI2 OUS	0.22	0.14	13.0	1.53	1
HUI3 OUS	0.33	0.24	11.3	1.16	2
Criterion D (n=51)					
EQ-VAS	20	17	8.5	1.00	6
EQ-Index	0.51	0.35	10.2	1.44	3
PCS-36	11	9	8.6	1.02	5
MCS-36	12	10	8.8	1.07	4
HUI2 OUS	0.25	0.15	12.4	2.13	1
HUI3 OUS	0.37	0.23	11.4	1.80	2

*all t-statistic p-values < 0.001

4.3.1.2 Responsiveness of Self-Assessed Scores (Any Change)

Responsiveness statistics were calculated and compared across summary scores for the patients who experienced any change (decline or improvement) according to criteria A through D.

Table 19: Responsiveness of Self-Assessed Scores (Any Change)

Summary Score	Mean change score (A)	SDch (B)	SDst (C)	SDbl (D)	SRM (A/B) (1)	GRS (A/C) (2)	Effect Size (A/D) (3)	Median Rank (1,2,3)
Criterion A (n=75)								
EQ-VAS	19.60	13.90	19.05	16.20	1.40	1.03	1.21	3
EQ-Index	0.37	0.31	0.47	0.36	1.17	0.79	1.01	5
PCS-36	7.50	6.70	5.20	9.00	1.12	1.44	0.83	6
MCS-36	11.00	8.80	10.44	11.30	1.25	1.06	0.98	4
HUI2 OUS	0.20	0.13	0.19	0.19	1.51	1.07	1.08	2
HUI3 OUS	0.30	0.21	0.22	0.30	1.43	1.40	0.99	2
Criterion B (n=57)								
EQ-VAS	20.90	14.70	19.05	16.20	1.42	1.10	1.29	3
EQ-Index	0.38	0.32	0.47	0.36	1.18	0.81	1.04	5
PCS-36	7.90	6.80	5.20	9.00	1.16	1.52	0.88	6
MCS-36	11.50	8.80	10.44	11.30	1.30	1.10	1.02	4.5
HUI2 OUS	0.21	0.13	0.19	0.19	1.56	1.11	1.12	2
HUI3 OUS	0.33	0.22	0.22	0.30	1.50	1.51	1.07	2
Criterion C (n=70)								
EQ-VAS	18.70	14.80	19.05	16.20	1.26	0.98	1.15	4
EQ-Index	0.43	0.34	0.47	0.36	1.25	0.91	1.17	5
PCS-36	10.60	9.30	5.20	9.00	1.14	2.04	1.18	1.5
MCS-36	11.10	8.40	10.44	11.30	1.33	1.07	0.99	4
HUI2 OUS	0.22	0.14	0.19	0.19	1.55	1.17	1.18	1.5
HUI3 OUS	0.33	0.24	0.22	0.30	1.36	1.53	1.08	2
Criterion D (n=51)								
EQ-VAS	20.20	17.00	19.05	16.20	1.19	1.06	1.24	6
EQ-Index	0.51	0.35	0.47	0.36	1.43	1.09	1.39	3
PCS-36	11.20	9.30	5.20	9.00	1.20	2.15	1.24	3.5
MCS-36	11.80	9.60	10.44	11.30	1.23	1.13	1.04	4
HUI2 OUS	0.25	0.15	0.19	0.19	1.73	1.36	1.37	2
HUI3 OUS	0.37	0.23	0.22	0.30	1.60	1.72	1.22	2

All of the self-assessed HRQL summary scores were responsive based on the mean scores of patients categorized as 'changed', using absolute values for the change scores (Table 19). Important change between time points was captured by each of the HRQL summary scores, using criteria A to D as external anchors of important change.

The HUI2 OUS and HUI3 OUS had the highest median ranked responsiveness indices for each of the criteria except for criterion C, where they were displaced by the PCS-36 as the most responsiveness.

Both mental and physical components of health were observed to have changed to a large extent, as represented by the effect sizes for the PCS-36 and MCS-36. All self-assessed summary scores had large effect sizes for criteria A through D, ranging from 0.83 (criterion A: PCS-36) to 1.39 (criterion D: EQ-Index). The SRM and GRS were generally larger than the effect size statistic. The SRM ranged from 1.12 (criterion A: PCS-36) to 1.73 (criterion D: HUI2 OUS). The range of the GRS was from 0.79 (criterion A: EQ-Index) to 2.15 (criterion D: PCS-36).

4.3.1.3 Responsiveness of Subgroups of Change – Self-Assessment

Responsiveness statistics were calculated and compared across summary scores for the patient subgroups of change (declined, no change, or improved) according to criteria A through D (minimum n = 10). These tables provide insight into the direction and magnitude of self-assessed scores for patients who were categorized as improved, did not change, or declined based on the change group criteria.

The mean change scores by self-assessment for patients belonging to the 'improved' change group were all positive, as hypothesized (Table 20). The mean difference score for the EQ-VAS in the 'improved' group based on criteria A and B was very small, with an associated trivial effect size. Otherwise, small to medium effect sizes were observed on the MCS-36 and PCS-36 for criteria A and B, and medium to large effect sizes were demonstrated for the EQ-Index, HUI2 OUS and HUI3 OUS. Criterion B was more stringent than criterion A for categorizing patients as having improved in health, but was not associated with larger effect sizes. The magnitudes of effect for self-assessed scores were generally larger for criteria C and D than for criteria A and B.

Table 20: 'Improved' Group Responsiveness of Self-Assessed Scores

Summary Score	Mean change score (A)	SDch (B)	SDst (C)	SDbl (D)	SRM (A/B) (1)	GRS (A/C) (2)	Effect Size (A/D) (3)	Median Rank (1,2,3)
Criterion A (n=63)								
EQ-VAS	1.49	25.18	19.05	16.46	0.06	0.08	0.09	6
EQ-Index	0.28	0.41	0.47	0.36	0.70	0.61	0.78	1
PCS-36	3.30	10.21	5.20	9.03	0.32	0.64	0.37	4
MCS-36	4.10	14.14	10.44	11.59	0.29	0.39	0.36	5
HUI2 OUS	0.12	0.21	0.19	0.20	0.57	0.64	0.64	3
HUI3 OUS	0.20	0.32	0.22	0.31	0.63	0.93	0.66	2
Criterion B (n=55)								
EQ-VAS	2.44	25.70	19.05	16.20	0.09	0.13	0.15	6
EQ-Index	0.28	0.42	0.47	0.36	0.68	0.61	0.78	1
PCS-36	3.01	10.20	5.20	9.00	0.30	0.58	0.33	5
MCS-36	4.93	13.59	10.44	11.30	0.36	0.47	0.44	4
HUI2 OUS	0.14	0.21	0.19	0.19	0.67	0.73	0.74	2
HUI3 OUS	0.21	0.33	0.22	0.30	0.63	0.96	0.68	3
Criterion C (n=62)								
EQ-VAS	10.04	21.11	19.05	16.20	0.48	0.53	0.62	6
EQ-Index	0.39	0.41	0.47	0.36	0.94	0.84	1.07	1
PCS-36	7.38	11.95	5.20	9.00	0.62	1.42	0.82	3
MCS-36	7.13	11.28	10.44	11.30	0.63	0.68	0.63	4
HUI2 OUS	0.12	0.24	0.19	0.19	0.52	0.66	0.67	5
HUI3 OUS	0.26	0.35	0.22	0.30	0.74	1.19	0.84	2
Criterion D (n=51)								
EQ-VAS	13.84	22.52	19.05	16.20	0.61	0.73	0.85	5
EQ-Index	0.49	0.38	0.47	0.36	1.28	1.05	1.34	1
PCS-36	8.51	11.83	5.20	9.00	0.72	1.64	0.95	4
MCS-36	5.92	14.08	10.44	11.30	0.42	0.57	0.52	6
HUI2 OUS	0.18	0.23	0.19	0.19	0.78	0.96	0.97	3
HUI3 OUS	0.32	0.30	0.22	0.30	1.05	1.48	1.05	2

As hypothesized, the MCS-36 was the least responsive of the summary scores according to criterion D. The PCS-36 did not receive the hypothesized highest ranking using criterion D, but the effect size for PCS-36 change scores using criterion D was highest among the 4 criteria within summary score. For all 4 criteria, the responsiveness indices ranked the EQ-Index most highly for self-assessed summary scores.

The number of patients in the 'no change' group was much smaller than the 'improved' group and only criterion A and D identified 10 or more patients as unchanged (Table 21). As a facile means of examining the suitability of external anchors for global health change in patients, the mean difference summary scores of patients 'changed' group were expected to demonstrate less responsiveness than those in the 'no change' subgroup. In this regard, criterion A performed poorly. The mean difference scores by self-assessment were higher for the 'no change' group than for the 'change' group for every summary score except the HUI2 OUS, and the effect sizes were non-trivial. Thus, criterion A (patient rates self) appears to be a sensitive but not specific method of categorizing patients into groups of change. In contrast, smaller effect sizes were observed for all summary scores when comparing the 'no change' group to the 'improved' group based upon criterion D. Using criterion D, the effect sizes were trivial (< 0.20) in the 'no change' group for all summary scores except for the EQ-Index (ES=0.39) and for the MCS-36 (ES=0.42). The MCS-36 change scores were not expected to correspond well with criterion D.

Table 21: 'No Change' Group Responsiveness of Self-Assessed Scores

Summary Score	Mean change score (A)	SDch (B)	SDst (C)	SDbl (D)	SRM (A/B) (1)	GRS (A/C) (2)	Effect Size (A/D) (3)	Median Rank (1,2,3)
Crit A (n=11)								
EQ-VAS	3.36	17.15	19.05	16.80	0.20	0.18	0.21	6
EQ-Index	0.49	0.45	0.47	0.45	1.09	1.04	1.36	1
PCS-36	4.14	11.43	5.20	11.50	0.36	0.80	0.46	5
MCS-36	8.93	11.31	10.44	11.30	0.79	0.86	0.79	3
HUI2 OUS	0.10	0.14	0.19	0.14	0.71	0.53	0.53	4
HUI3 OUS	0.25	0.32	0.22	0.32	0.78	1.14	0.83	2
Crit D (n=25)								
EQ-VAS	-0.48	21.86	19.05	24.00	-0.02	-0.03	-0.03	6
EQ-Index	0.14	0.35	0.47	0.35	0.40	0.30	0.39	2
PCS-36	1.03	12.28	5.20	12.88	0.08	0.20	0.11	4.5
MCS-36	4.69	12.21	10.44	12.34	0.38	0.45	0.42	1
HUI2 OUS	-0.02	0.17	0.19	0.17	-0.12	-0.11	-0.11	4.5
HUI3 OUS	0.05	0.32	0.22	0.33	0.16	0.23	0.17	3

Only criteria A (patient rates self) identified 10 or more patients as 'declined' in overall health (Table 22). Contrary to the hypotheses, the direction of change was positive (improved) for the EQ-Index, HUI2 OUS and HUI3 OUS for these 12 patients. The EQ-VAS, PCS-36, and MCS-36 all captured a decrease in mean difference scores, but the effect sizes were an order of magnitude lower than for the patients who were classified as 'improved' using the same criterion.

Table 22: 'Declined' Group Responsiveness of Self-Assessed Scores

Summary Score	Mean change score (A)	SDch (B)	SDst (C)	SDbl (D)	SRM (A/B) (1)	GRS (A/C) (2)	Effect Size (A/D) (3)	Median Rank (1,2,3)
Crit A (n=12)								
EQ-VAS	-7.42	16.64	19.05	16.2	-0.45	-0.39	-0.46	1
EQ-Index	0.03	0.44	0.47	0.36	0.07	0.06	0.08	--
PCS-36	-1.67	5.11	5.20	9.00	-0.33	-0.32	-0.19	2
MCS-36	-2.44	11.07	10.44	11.3	-0.22	-0.23	-0.22	3
HUI2 OUS	0.04	0.26	0.19	0.19	0.15	0.21	0.21	--
HUI3 OUS	0.08	0.32	0.22	0.30	0.25	0.36	0.27	--

4.3.1.4 Self-Assessed Scores - 'Change' versus 'No Change'

The responsiveness statistics (in section 4.3.1.2) examined patients who were categorized as 'improved or declined' to the exclusion of the patients categorized as 'no change' based on the 4 criteria. Independent sample t-tests were performed to compare mean change scores between patients categorized as 'changed' to those categorized as 'not changed' according to each criterion. Contrary to expectations, the mean change scores were greater in the no change group than in the change group for several self-assessed scores (see criteria A, B, and C in Table 23). This was most problematic for criterion A, where the patient was responsible for rating self as changed or not (a negative sign in front of the t-statistic was used to denote a larger mean difference score in 'no change' group than in 'change' group). These problems were foreshadowed by the findings in section 4.3.1.3, where the 'no change' group, according to criterion A, had larger effect sizes than the 'improved' group. In the case of the EQ-Index, the self-assessed scores of the 'no change' group experienced significantly more change than those who said they had 'changed' using criterion A.

This may be indicative of a problem with misclassification error using criterion A to categorize patients, and/or limitations associated with EQ-Index-based scoring algorithm. The small subgroup of 'no change' patients according to criterion C (clinician rated patient) limited the ability to make statistical inferences. The mean difference scores in the 'no change group' were larger than in the 'change' group for both the EQ-VAS and MCS-36. This supports the notion that clinician based evaluations may focus more on the observable aspects of health.

Table 23: Comparison of 'Change' vs. 'No Change' Groups (Patient Scores)

	Change Group			No Change Group			t-stat	Sig
Summary Score	N	Mean Change	SD	N	Mean Change	SD		
Criterion A								
EQ-VAS	75	19.60	13.90	11	11.70	12.50	1.78	0.08
EQ-Index	75	0.37	0.31	11	0.57	0.34	-2.00	*0.05
PCS-36	75	7.50	6.70	11	9.65	6.87	-1.01	0.32
MCS-36	75	11.00	8.80	11	11.70	8.06	-0.24	0.81
HUI2 OUS	75	0.20	0.13	11	0.14	0.10	1.40	0.06
HUI3 OUS	75	0.30	0.21	11	0.34	0.22	-0.51	0.61
Criterion B								
EQ-VAS	57	20.90	14.70	26	14.40	14.70	1.99	*0.05
EQ-Index	57	0.38	0.32	26	0.43	0.33	-0.68	0.50
PCS-36	57	7.90	6.80	26	7.60	6.80	0.18	0.86
MCS-36	57	11.50	8.80	26	10.40	8.50	0.54	1.12
HUI2 OUS	57	0.21	0.13	26	0.16	0.12	1.50	0.14
HUI3 OUS	57	0.33	0.22	26	0.28	0.19	1.16	0.25
Criterion C								
EQ-VAS	70	18.70	14.80	7	21.40	25.40	-0.28	0.67
EQ-Index	70	0.43	0.34	7	0.37	0.46	0.38	0.71
PCS-36	70	10.60	9.30	7	7.22	4.59	0.95	0.35
MCS-36	70	11.10	8.40	7	13.20	13.30	-0.40	0.70
HUI2 OUS	70	0.22	0.14	7	0.16	0.14	1.13	0.26
HUI3 OUS	70	0.33	0.24	7	0.22	0.22	1.12	0.27
Criterion D								
EQ-VAS	51	20.20	17.00	25	16.80	13.60	0.86	0.39
EQ-Index	51	0.51	0.35	25	0.25	0.28	3.52	*0.001
PCS-36	51	11.20	9.30	25	8.60	8.60	1.16	0.25
MCS-36	51	11.80	9.60	25	10.80	7.16	0.48	0.63
HUI2 OUS	51	0.25	0.15	25	0.13	0.10	4.29	*0.0001
HUI3 OUS	51	0.37	0.23	25	0.22	0.23	2.60	*0.01

*p-value <0.05

Criterion D was the only method of patient categorization that consistently demonstrated larger mean difference scores in the 'change' group than in the 'no change' group for each of the self-assessed summary scores. Further, the EQ-Index, HUI2 OUS and HUI3 OUS had mean difference scores that were significantly greater in the 'change' group than in the 'no change' group. The MCS-36 did not have mean difference scores that were significantly greater in the 'change group' than in the 'no change' group, but this was not surprising because the basis for criterion D is the Barthel Index, a measure of functional ability.

4.3.2 Longitudinal Construct Validity – Proxy-Assessed HRQL Scores

4.3.2.1 Sensitivity – Proxy Assessed Scores

Tests of statistical significance (paired t-tests) were performed on the pre- and post-summary scores of patients who were categorized as changed (improved or declined) according to criteria A to D. All of the proxy-assessed HRQL summary scores were sensitive (p-value <0.001) using criterion A through D to define patients whose health changed (Table 24). Ratios were comparable and rankings may be misleading.

Table 24: Sensitivity of Proxy-Assessed Summary Scores

Summary Score	Absolute mean difference score	Std Dev	t-stat*	Squared t-test ratio	Rank
Cnterion A (n=73)					
EQ-VAS	17.10	14.50	10.00	1.00	3
EQ-Index	0.34	0.27	10.50	1.10	1
PCS-36	6.66	5.96	9.54	0.91	6
MCS-36	10.30	8.70	10.10	1.02	2
HUI2 OUS	0.15	0.13	9.78	0.96	4
HUI3 OUS	0.26	0.23	9.58	0.92	5
Cnterion B (n=57)					
EQ-VAS	19.30	14.70	9.92	1.00	1
EQ-Index	0.37	0.29	9.61	0.94	3
PCS-36	6.92	6.26	8.34	0.71	6
MCS-36	11.50	8.88	9.77	0.97	2
HUI2 OUS	0.16	0.14	8.42	0.72	4.5
HUI3 OUS	0.27	0.25	8.39	0.72	4.5
Cnterion C (n=68)					
EQ-VAS	22.00	18.70	9.70	1.00	5
EQ-Index	0.40	0.32	10.10	1.08	3
PCS-36	10.20	8.78	9.60	0.98	6
MCS-36	9.67	8.09	9.90	1.04	4
HUI2 OUS	0.22	0.16	11.70	1.45	1
HUI3 OUS	0.33	0.26	10.80	1.24	2
Criterion D (n=49)					
EQ-VAS	27.00	19.10	9.89	1.00	4
EQ-Index	0.49	0.35	9.73	0.97	5

PCS-36	11.50	9.59	8.41	0.72	6
MCS-36	11.00	7.60	10.10	1.04	3
HUI2 OUS	0.25	0.16	10.70	1.17	1
HUI3 OUS	0.38	0.26	10.30	1.08	2

*all t-statistics p-value < 0.001

For criteria C and D, the HUI2 OUS and HUI3 OUS were observed to have the largest squared t-test ratios for proxy assessed summary scores. For criteria A, the EQ-Index had the largest squared t-test ratio among proxy-assessed scores. For criteria B, the EQ-VAS had the largest squared t-test ratio among proxy-assessed scores. The PCS-36 had the lowest ranking among the HRQL summary scores for all 4 criteria.

The magnitude of t-statistics of the proxy-assessed summary scores was similar to those of patient self-assessment: they were all relatively comparable. The HUI2 OUS and HUI 3 OUS were the most sensitive summary scores for both patient self-assessment and proxy assessment using criteria C (clinician rated patient) and D (Barthel Index-based change).

4.3.2.2 Responsiveness of Proxy-Assessed Scores (Any Change)

Similar to the patient self-assessed scores, important change between time points was captured by each of the proxy-assessed HRQL summary scores, using criteria A to D as external anchors of important change (Table 25). The EQ-VAS and MCS-36 had the highest median ranks for criterion A. There was very little variation in the range for both effect size (0.75 to 0.87) and for the SRM (1.12 to 1.23) using criterion A to compare summary scores. The EQ-VAS ranked highest using criteria B to group patients as 'changed'. The HUI2 OUS and HUI3 OUS were ranked highest according to the responsiveness indices for criteria C and D.

All proxy-assessed summary scores had large effect sizes for criteria A through D, ranging from 0.83 (criterion A: PCS-36) to 1.39 (criterion D: EQ-Index). The SRM and GRS were generally larger than the effect size statistic. The SRM ranged from 1.12 (criterion A: PCS-36) to 1.73 (criterion D: HUI2 OUS).

Table 25: Responsiveness of Proxy Assessed Scores (Any Change)

Summary Score	Mn change score (A)	SDch (B)	SDst (C)	SDbl (D)	SRM (A/B) (1)	GRS (A/C) (2)	Effect Size (A/D) (3)	Md Rank (1,2, 3)
Criterion A (n=73)								
EQ-VAS	17.10	14.50	7.65	19.60	1.18	2.23	0.87	2
EQ-Index	0.34	0.27	0.42	0.40	1.23	0.79	0.83	3
PCS-36	6.70	6.00	4.36	8.40	1.12	1.53	0.79	4
MCS-36	10.30	8.70	7.49	12.10	1.18	1.38	0.85	2.5
HUI2 OUS	0.15	0.13	0.07	0.20	1.14	2.18	0.75	4
HUI3 OUS	0.26	0.23	0.07	0.33	1.12	3.84	0.78	5
Criterion B (n=57)								
EQ-VAS	19.30	14.70	7.65	19.60	1.31	2.52	0.98	1
EQ-Index	0.37	0.29	0.42	0.40	1.27	0.86	0.91	3
PCS-36	6.90	6.30	4.36	8.40	1.11	1.59	0.82	5
MCS-36	11.50	8.90	7.49	12.10	1.29	1.54	0.95	2
HUI2 OUS	0.16	0.14	0.07	0.20	1.11	2.26	0.78	5
HUI3 OUS	0.27	0.25	0.07	0.33	1.11	4.08	0.83	4
Criterion C (n=68)								
EQ-VAS	22.00	18.70	7.65	19.60	1.18	2.88	1.12	3
EQ-Index	0.40	0.32	0.42	0.40	1.23	0.93	0.98	5
PCS-36	10.20	8.80	4.36	8.40	1.16	2.34	1.21	4
MCS-36	9.70	8.10	7.49	12.10	1.20	1.29	0.80	5
HUI2 OUS	0.22	0.16	0.07	0.20	1.42	3.22	1.11	2
HUI3 OUS	0.33	0.26	0.07	0.33	1.31	4.98	1.01	2
Criterion D (n=49)								
EQ-VAS	27.00	19.10	7.65	19.60	1.41	3.53	1.38	3
EQ-Index	0.49	0.35	0.42	0.40	1.39	1.15	1.21	5
PCS-36	11.50	9.60	4.36	8.40	1.20	2.64	1.37	4
MCS-36	11.00	7.60	7.49	12.10	1.45	1.47	0.91	5
HUI2 OUS	0.25	0.16	0.07	0.20	1.52	3.63	1.25	2
HUI3 OUS	0.38	0.26	0.07	0.33	1.47	5.71	1.16	2

Mn=mean; Md=Median

GRS was very large for several proxy-assessed summary scores owing to the small amount of variance among the patient classified as stable. The range of GRS was from 0.79 (criterion A: EQ-Index) to 5.71 (criterion D: HUI3 OUS). GRS was greater than 1 for all proxy-assessed summary scores with the exception of the EQ-Index, due to the large standard deviation in the change scores (measurement noise) associated with patients classified as stable.

4.3.2.3 Responsiveness of Proxy-Assessed Scores (Subgroups)

A positive mean difference score for each of the summary scores assessed by proxy was observed, as hypothesized, for the 'improved' change group based on each of the criteria (Table 26). Medium-small to medium-large effect sizes were observed for all of the proxy assessed summary scores based on criteria A and B. The hypotheses that the effect sizes for summary scores would be lower for proxy assessment than for patient self assessment using criterion A and higher using criterion B were not consistently confirmed. Criterion B was more stringent than criterion A for categorizing patients as having improved in health, and slightly larger effect sizes were observed across all summary scores when comparing criterion B to A.

Similar to the findings for self-assessed scores, the effect sizes associated with proxy-assessment were generally larger for criteria C and D than for criteria A and B. As hypothesized, the MCS-36 was the least responsive summary score when ranked using criterion D. While the PCS-36 did not receive the highest ranking, the effect size for PCS-36 using criterion D was higher than for the other criteria. The EQ-Index had the highest median ranked responsiveness indices based on criteria A and B, while the EQ-VAS was the most responsive summary scores using criteria C and D for proxy-assessed summary scores.

Table 26: Responsiveness of Proxy-Assessed Scores (Improved Group)

Summary Score	Mean change score (A)	SDch (B)	SDst (C)	SDbl (D)	SRM (A/B) (1)	GRS (A/C) (2)	Effect Size (A/D) (3)	Median Rank (1,2,3)
Cnterion A (n=61)								
EQ-VAS	12.56	19.95	7.65	20.95	0.63	1.64	0.64	2
EQ-Index	0.26	0.37	0.42	0.42	0.69	0.61	0.64	1.5
PCS-36	3.41	8.37	4.36	9.11	0.41	0.78	0.41	5
MCS-36	3.69	13.90	7.49	12.09	0.27	0.49	0.30	6
HUI2 OUS	0.10	0.18	0.07	0.19	0.52	1.37	0.48	4
HUI3 OUS	0.18	0.32	0.07	0.34	0.57	2.68	0.54	3
Cnterion B (n=55)								
EQ-VAS	13.30	20.50	7.65	19.60	0.65	1.74	0.68	2
EQ-Index	0.29	0.37	0.42	0.40	0.78	0.68	0.71	1
PCS-36	3.71	8.50	4.36	8.40	0.43	0.85	0.44	5
MCS-36	4.38	13.90	7.49	12.10	0.32	0.59	0.36	6
HUI2 OUS	0.11	0.18	0.07	0.20	0.58	1.55	0.54	4
HUI3 OUS	0.19	0.33	0.07	0.33	0.58	2.79	0.57	3
Cnterion C (n=60)								
EQ-VAS	21.95	21.04	7.65	19.60	1.04	2.87	1.12	1
EQ-Index	0.38	0.39	0.42	0.40	0.97	0.89	0.93	3
PCS-36	8.80	11.05	4.36	8.40	0.80	2.02	1.05	4
MCS-36	6.83	11.38	7.49	12.10	0.60	0.91	0.56	6
HUI2 OUS	0.18	0.22	0.07	0.20	0.84	2.65	0.92	3
HUI3 OUS	0.28	0.34	0.07	0.33	0.83	4.17	0.85	4
Cnterion D (n=49)								
EQ-VAS	25.29	21.44	7.65	19.60	1.18	3.31	1.29	2
EQ-Index	0.47	0.38	0.42	0.40	1.22	1.10	1.15	3
PCS-36	9.79	11.38	4.36	8.40	0.86	2.25	1.17	4
MCS-36	6.97	11.46	7.49	12.10	0.61	0.93	0.58	6
HUI2 OUS	0.20	0.22	0.07	0.20	0.94	2.95	1.02	4
HUI3 OUS	0.32	0.33	0.07	0.33	0.97	4.79	0.97	3

The 'no change' group was much smaller than the 'improved' group and only cnteria A and D identified 10 or more patients as unchanged (Table 27). Unexpectedly, the proxy-assessed EQ-VAS and EQ-Index had effect sizes that were larger for the 'no change' group than for the 'improved group' according to cnterion A. The effect sizes were non-trivial for all summary scores using both criteria A and D to define patients who did not change. Disconcertingly, mean difference scores exceeded MCIDs for every summary score except the PCS-36 (using criteria A and D) and the MCS-36 (criteria D).

Generally, smaller effect sizes were observed using criterion D compared to criterion A. However, relative to the patient assessed scores where most of the effect sizes were trivial (< 0.20), the proxy assessments generated larger effect sizes in the same patients categorized as not having changed (using the same criteria).

Table 27: Responsiveness of Proxy-Assessed Scores (No Change Group)

Summary Score	Mean change score (A)	SDch (B)	SDst (C)	SDbl (D)	SRM (A/B) (1)	GRS (A/C) (2)	Effect Size (A/D) (3)	Median Rank (1,2,3)
Cnterion A (n=10)								
EQ-VAS	15.40	15.09	7.65	19.60	1.02	2.01	0.79	2
EQ-Index	0.30	0.28	0.42	0.40	1.09	0.71	0.75	3
PCS-36	1.96	6.90	4.36	8.40	0.28	0.45	0.23	6
MCS-36	5.14	11.96	7.49	12.10	0.43	0.69	0.42	5
HUI2 OUS	0.14	0.12	0.07	0.20	1.17	1.99	0.69	3
HUI3 OUS	0.20	0.17	0.07	0.33	1.17	2.93	0.60	1.5
Criterion D (n=25)								
EQ-VAS	12.04	17.94	7.65	19.60	0.67	1.57	0.61	1
EQ-Index	0.14	0.31	0.42	0.40	0.45	0.33	0.35	5
PCS-36	3.33	8.30	4.36	8.40	0.40	0.76	0.40	3.5
MCS-36	3.94	11.19	7.49	12.10	0.35	0.53	0.33	5
HUI2 OUS	0.08	0.18	0.07	0.20	0.44	1.14	0.40	3
HUI3 OUS	0.15	0.26	0.07	0.33	0.58	2.21	0.45	2

No criteria other than criterion A (patient rates self) identified 10 or more patients as having 'declined' in overall health. Contrary to the hypotheses, the direction of change was positive (improved) for the EQ-VAS, PCS-36, HUI2 OUS and HUI3 OUS. The EQ-Index and MCS-36 had negative mean difference scores, but the effect sizes were much smaller than for the patients who were classified as 'improved' using the same criterion.

Table 28: Responsiveness of Proxy-Assessed Scores (Declined Group)

Summary Score	Mean change score (A)	SDch (B)	SDst (C)	SDbl (D)	SRM (A/B) (1)	GRS (A/C) (2)	Effect Size (A/D) (3)	Med-ian Rank (1,2,3)
Criterion A (n=12)								
EQ-VAS	5.17	15.09	7.65	19.60	0.34	0.68	0.26	--
EQ-Index	-0.04	0.33	0.42	0.40	-0.12	-0.09	-0.10	2
PCS-36	5.20	7.11	4.36	8.40	0.73	1.19	0.62	--
MCS-36	-2.74	7.95	7.49	12.10	-0.35	-0.37	-0.23	1
HUI2 OUS	0.06	0.16	0.07	0.20	0.34	0.80	0.28	--
HUI3 OUS	0.02	0.24	0.07	0.33	0.07	0.24	0.05	--

4.3.2.4 Proxy-Assessed Scores: 'Change' Versus 'No Change'

The mean difference scores in the change group, as defined by each of the criteria, were compared to the 'no change' group using independent t-tests (Table 29). The EQ-VAS had a larger mean difference score in the 'no change' group than the 'change' group according to criteria A, but the null hypothesis that there was no difference between the mean change scores was not rejected (p-value > 0.05). The clinician-based criterion C was associated with the two other instances (EQ-VAS, MCS-36) where a larger mean difference scores was larger in the 'no change' group than the 'change' group.

Similar to the self-assessed scores, criterion D appeared to be a superior method of patient categorization that demonstrated statistically significantly larger mean difference scores in the 'change' group than in the 'no change' group for each of the self-assessed summary scores, with the exception of the MCS-36. Because the basis for criterion D is the Barthel Index, a measure of functional ability, mean difference scores for the MCS-36 were expected to be independent of this criterion.

Table 29: Comparison of 'Change' vs. 'No Change' Groups (Proxy Scores)

	Change Group			No Change Group			t-stat	Sig
Summary Score	N	Mean change	SD	N	Mean change	SD		
Criterion A								
EQ-VAS	73	17.10	14.50	10	18.40	10.70	-0.27	0.79
EQ-Index	73	0.34	0.27	10	0.30	0.28	0.39	0.71
PCS-36	73	6.70	6.00	10	4.85	5.07	0.91	0.37
MCS-36	73	10.30	8.70	10	9.75	8.17	0.19	0.85
HUI2 OUS	73	0.15	0.13	10	0.14	0.12	0.30	0.77
HUI3 OUS	73	0.26	0.23	10	0.20	0.16	0.75	0.45
Criterion B								
EQ-VAS	57	19.30	14.70	26	12.80	11.70	1.98	*0.05
EQ-Index	57	0.37	0.29	26	0.26	0.23	1.71	0.09
PCS-36	57	6.90	6.30	26	5.39	4.82	1.11	1.53
MCS-36	57	11.50	8.90	26	7.47	7.34	2.02	*0.05
HUI2 OUS	57	0.16	0.14	26	0.13	0.10	0.71	0.48
HUI3 OUS	57	0.27	0.25	26	0.20	0.15	1.40	0.17
Criterion C								
EQ-VAS	68	22.00	18.70	7	29.70	18.70	-1.03	0.30
EQ-Index	68	0.40	0.32	7	0.52	0.41	-0.95	0.34
PCS-36	68	10.20	8.80	7	7.18	6.56	0.88	0.38
MCS-36	68	9.70	8.10	7	14.60	8.15	-1.55	0.13
HUI2 OUS	68	0.22	0.16	7	0.19	0.09	0.97	0.35
HUI3 OUS	68	0.33	0.26	7	0.22	0.23	1.10	0.27
Criterion D								
EQ-VAS	49	27.00	19.10	25	15.20	15.30	2.70	*0.01
EQ-Index	49	0.49	0.35	25	0.27	0.21	3.39	*0.001
PCS-36	49	11.50	9.60	25	7.06	5.34	2.57	*0.01
MCS-36	49	11.00	7.60	25	7.92	8.71	1.49	0.14
HUI2 OUS	49	0.25	0.16	25	0.16	0.11	2.80	*0.002
HUI3 OUS	49	0.38	0.26	25	0.21	0.21	3.11	*0.003

* t-statistic $p < 0.05$

4.4.0 Comparison of Patient and Proxy Assessment Scores

Patient self-assessed and proxy-assessed change scores were contrasted for each of the 4 criteria (Tables 30 to 33). The sensitivity of self- and proxy-assessed HRQL summary scores were expected to differ depending on the criteria used to categorize patients as changed or not. As expected, change scores by patient self-assessment were larger than by proxy-assessment for patient self-rated global scale of change (criterion A), but this difference was statistically significant between perspectives only for the HUI2 OUS (p<0.01) (Table 30).

Table 30: Comparison of Patient and Proxy Scores (Criterion A)

Score	Patient-Assessed Scores		Proxy-Assessed Score		Difference Between Patient and Proxy Change Scores		
	Mean change between t_0 and t_1	Std Dev	Mean change between t_0 and t_1	Std Dev	Mean	Std Dev	t-stat
EQ-VAS	19.60	13.90	17.10	14.50	2.75	18.52	1.27
EQ-Index	0.37	0.31	0.34	0.27	0.04	0.40	0.80
PCS-36	7.47	6.67	6.66	5.96	0.86	8.07	0.91
MCS-36	11.00	8.82	10.30	8.70	0.63	11.29	0.47
HUI2 OUS	0.20	0.13	0.15	0.13	0.05	0.15	•2.97
HUI3 OUS	0.30	0.21	0.26	0.23	0.05	0.23	1.81

•p-value <0.01

For criterion B (patient and proxy agree patient changed between baseline and 1 month), the HUI2 OUS change scores were significantly different (p<0.01). Differences were not expected because patient and proxy agreed change took place. However, they did not have to agree on the magnitude of change.

Table 31: Comparison of Patient and Proxy Scores (Criterion B)

Score	Patient-Assessed Scores (n=57)		Proxy-Assessed Score (n=57)		Difference Between Patient and Proxy Change Scores		
	Mean change between t_0 and t_1	Std Dev	Mean change between t_0 and t_1	Std Dev	Mean	Std Dev	t-stat
EQ-VAS	20.90	14.70	19.30	14.70	1.67	19.40	0.65
EQ-Index	0.38	0.32	0.37	0.29	0.01	0.40	0.22
PCS-36	7.89	6.79	6.92	6.26	0.98	7.74	0.95
MCS-36	11.50	8.82	11.50	8.88	0.01	11.37	0.01
HUI2 OUS	0.21	0.13	0.16	0.14	0.05	0.14	•2.69
HUI3 OUS	0.33	0.22	0.27	0.25	0.05	0.25	1.59

•p-value <0.01

The clinical assessor global rating of patient change between baseline and 6 months (criterion C) demonstrated no differences between patient self- and proxy-assessed change scores (Table 32).

Table 32: Comparison of Patient and Proxy Scores (Criterion C)

Instrument	Patient-Assessed Scores (n=68)		Proxy-Assessed Scores (n=68)		Difference Between Patient and Proxy Change Scores		
	Mean change between t_0 and t_6	Std Dev	Mean change between t_0 and t_6	Std Dev	Mean	Std Dev	t-stat
EQ-VAS	18.70	14.80	22.00	18.70	-3.07	20.30	-1.25
EQ-Index	0.43	0.34	0.40	0.32	0.04	0.35	0.86
PCS-36	10.60	9.35	10.20	8.78	0.50	9.23	0.45
MCS-36	11.10	8.39	9.67	8.09	1.52	11.88	1.06
HUI2 OUS	0.22	0.14	0.22	0.16	-0.01	0.18	-0.38
HUI3 OUS	0.33	0.24	0.33	0.26	-0.01	0.28	-0.30

If differences arose between self- and proxy-assessed change scores, proxy-assessed change scores were expected to be larger than patient-assessed change scores for criterion D. Proxy-assessed EQ-VAS scores were significantly lower than patient-self assessed scores among the patients categorized as changed using criterion D (p<0.05) (Table 33).

Table 33: Comparison of Patient and Proxy Scores (Criterion D)

Instrument	Patient-Assessed Scores (n=49)		Proxy-Assessed Scores (n=49)		Difference Between Patient and Proxy Change Scores		
	Mean change between t_0 and t_6	Std Dev	Mean change between t_0 and t_6	Std Dev	Mean	Std Dev	t-stat
EQ-VAS	20.20	17.00	27.00	19.10	-6.44	21.42	*2.11
EQ-Index	0.51	0.35	0.49	0.35	0.03	0.41	0.53
PCS-36	11.20	9.30	11.50	9.59	-0.23	8.37	0.19
MCS-36	11.80	9.58	11.00	7.60	0.93	12.56	0.52
HUI2 OUS	0.25	0.15	0.25	0.16	-0.002	0.20	0.09
HUI3 OUS	0.37	0.23	0.38	0.26	-0.02	0.29	0.41

*p-value <0.05

Self- and proxy assessed scores were also compared by testing for significant differences over time. For the EQ-VAS, an interaction effect between time and respondent type was observed ($_{(0.05)}F_{(3,204)}$=7.894; p<0.001). Therefore, the null hypothesis that there was no difference between self- and proxy assessed EQ-VAS scores over time was rejected. The observed power was 0.935. The difference in scores was illustrated by the graph where mean scores began apart at baseline and converged from month 1 to month 6 (Appendix 7).

For the EQ-Index, an interaction effect between time and respondent type was not observed ($_{(0.05)}F_{(2.54,173.01)}$=0.703; p=0.529). The observed power was 0.184, which limited the ability to detect a difference between EQ-Index scores by self-assessment and proxy assessment.

An interaction effect between time and respondent type was not observed for the PCS-36 (${}_{(0.05)}F_{(2.58,175.22)}=2.13$; p=0.11) or for the MCS-36 (${}_{(0.05)}F_{(2.63,178.74)}=1.45$; p=0.23). The observed power was 0.50 and 0.35, respectively.

Similarly, no interaction effect between time and respondent type was observed for the HUI 2 OUS (${}_{(0.05)}F_{(3,204)}=1.65$; p=0.18), nor for the HUI 3 OUS (${}_{(0.05)}F_{(3,204)}=0.353$; p=0.79). The observed power to detect an effect was 0.43 and 0.12, respectively.

4.5.0 Agreement between Patient and Proxy Assessments

For the cross-sectional and change scores for each HRQL measure summary score, the average absolute value of the difference between self- and proxy assessments and the standard deviations of those values were calculated. The means of patient and proxy scores were compared using paired t-tests. No adjustment was made for multiple tests, but tests with p-values less than 0.05 and 0.005 were identified. Pearson's product-moment correlation coefficients, a one-way random-effects ICC and the non-parametric ICC (Robinson, 1957) were used to compare summary scores generated by self- and proxy-assessment at the different time periods examined in the study.

4.5.1 Comparison of Cross-Sectional Self- and Proxy-Assessed Scores

Mean self-assessed summary scores were higher than those assessed by proxy in almost every instance except for EQ-VAS at 1 month (-2.17), 6 months (-1.39), and MCS-36 (-0.12). The latter differences did not reach the CID threshold, as previously defined. Statistically significant differences between patient and proxy assessed scores (no items imputed) were detected on the EQ-VAS at baseline, the EQ-Index scores at 1 month, PCS-36 scores at baseline and 1 month, and the MCS-36 at month 3 (Table 34).

Table 34: Comparison of Patient and Proxy-Assessed Cross-Sectional Scores

	Patient			Proxy			Difference Scores					
Time	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**t-stat**	**sig**	**SRM**
EQ-VAS												
Baseline	97	60.56	16.21	97	51.27	19.58	97	9.29	21.56	4.24*	0.00	0.43
Month 1	86	61.99	18.95	83	63.98	18.47	83	-2.17	19.79	-1.00	0.32	-0.11
Month 3	79	67.67	17.02	79	65.67	20.42	77	1.73	19.07	0.79	0.43	0.09
Month 6	77	69.86	18.65	76	70.24	20.01	75	-1.39	18.49	-0.65	0.52	-0.08
EQ-Index												
Baseline	97	0.23	0.36	97	0.18	0.40	97	0.05	0.40	1.14	0.26	0.12
Month 1	86	0.51	0.38	81	0.43	0.36	81	0.08	0.25	2.74*	0.01	0.30
Month 3	80	0.57	0.32	79	0.50	0.37	78	0.07	0.30	1.91	0.06	0.22
Month 6	76	0.60	0.33	74	0.53	0.39	72	0.04	0.24	1.51	0.14	0.18
PCS-36												
Baseline	97	28.89	9.03	96	25.57	8.43	96	3.38	9.13	3.63*	0.00	0.37
Month 1	85	32.26	10.54	83	29.22	10.50	82	2.67	7.26	3.32*	0.00	0.37
Month 3	76	33.65	11.39	79	32.81	11.95	74	1.03	7.96	1.11	0.27	0.13
Month 6	77	35.06	12.44	76	33.01	12.48	75	1.90	9.70	1.7	0.09	0.20
MCS-36												
Baseline	97	47.10	11.33	96	46.59	12.10	96	0.63	14.04	0.44	0.66	0.04
Month 1	85	50.76	13.00	83	49.84	12.31	82	0.65	13.33	0.44	0.66	0.05
Month 3	76	51.94	12.80	79	48.92	12.45	74	3.80	11.93	2.74*	0.01	0.32
Month 6	77	52.66	10.36	76	52.16	11.94	75	-0.16	11.48	-0.12	0.90	-0.01
HUI2 OUS												
Baseline	93	0.52	0.18	93	0.50	0.20	89	0.02	0.17	0.97	0.34	0.10
Month 1	82	0.63	0.21	78	0.60	0.20	74	0.03	0.16	1.59	0.12	0.18
Month 3	79	0.64	0.21	76	0.63	0.23	74	0.01	0.18	0.66	0.51	0.08
Month 6	74	0.64	0.23	73	0.65	0.24	69	0.02	0.17	-0.86	0.40	0.10
HUI3 OUS												
Baseline	91	0.22	0.29	92	0.18	0.34	86	0.03	0.30	1.01	0.32	0.11
Month 1	83	0.40	0.35	78	0.34	0.32	75	0.05	0.28	1.72	0.09	0.20
Month 3	79	0.42	0.33	76	0.39	0.35	74	0.04	0.29	1.16	0.25	0.14
Month 6	73	0.45	0.35	73	0.43	0.36	68	0.02	0.28	0.44	0.66	0.05

*p<0.05; SRM

No paired t-tests detected statistically significant differences between rater types for the HUI2 OUS and HUI3 OUS. All statistically significant differences between assessment scores by rater type had small to moderate magnitudes of effect size; all other comparisons at each time period between rater assessment scores were not statistically significant and had effect sizes less than 0.25. A CID between rater scores was observed at baseline for EQ-VAS (difference in scores >5); at all time points for the EQ-Index (difference in scores >0.036); and at baseline, 1 month and 3 months for the HUI3 OUS (difference in scores >0.03) (Table 34).

The statistics of association (Pearson's r) and agreement (1-way ANOVA-based ICC and Robinson's ICC) were observed to have similar patterns of findings (Tables 35).

Table 35: Cross-sectional Agreement between Self- and Proxy Assessment

	Agreement between Patient and Proxy		
Summary Scores	**Pearson's r***	**Case 1 ICC (one-way)**	**Robinson's ICC**
EQ-VAS			
Baseline	0.29	0.20 (0.01, 0.39)	0.20
Month 1	0.45	0.45 (0.26, 0.62)	0.44
Month 3	0.47	0.47 (0.27, 0.62)	0.46
Month 6	0.51	0.51 (0.32, 0.66)	0.51
EQ-Index			
Baseline	0.46	0.45 (0.28, 0.60)	0.45
Month 1	0.78	0.76 (0.66, 0.84)	0.76
Month 3	0.60	0.58 (0.42, 0.71)	0.58
Month 6	0.77	0.76 (0.64, 0.84)	0.75
PCS-36			
Baseline	0.46	0.41 (0.23, 0.56)	0.40
Month 1	0.76	0.73 (0.61, 0.82)	0.73
Month 3	0.77	0.76 (0.65, 0.85)	0.76
Month 6	0.70	0.69 (0.56, 0.80)	0.69
MCS-36			
Baseline	0.28	0.29 (0.09, 0.46)	0.28
Month 1	0.44	0.44 (0.25, 0.60)	0.44
Month 3	0.54	0.51 (0.32, 0.66)	0.51
Month 6	0.44	0.45 (0.25, 0.61)	0.44
HUI2 OUS			
Baseline	0.59	0.59 (0.43, 0.71)	0.58
Month 1	0.73	0.72 (0.59, 0.81)	0.72
Month 3	0.65	0.65 (0.49, 0.76)	0.64
Month 6	0.72	0.72 (0.58, 0.82)	0.72
HUI3 OUS			
Baseline	0.55	0.55 (0.38, 0.68)	0.55
Month 1	0.67	0.66 (0.52, 0.77)	0.66
Month 3	0.64	0.64 (0.48, 0.76)	0.63
Month 6	0.69	0.69 (0.54, 0.80)	0.68

*p-value <0.005, test of null hypothesis that r = 0

Pearson's r and the ICCs were nearly the same except for scores at time points where a systematic difference occurred (typically for scores where the t-statistic was significant). Thus, ICCs were slightly lower than Pearson's r. The 1-way ANOVA-based ICC and Robinson's ICC gave almost identical results. Specific hypothesis on agreement between cross-sectional scores are addressed below:

Hypotheses (H1): greater agreement between cross-sectional scores is observed as more time elapses and the patient stabilizes.

A trend towards greater agreement on successive data collection waves was found for the EQ-VAS. Point estimates of agreement using ICCs indicated that agreement at baseline was poorer relative to successive data collection times for all of the summary scores. However, the confidence intervals for the ICCs were wide and the trend was not statistically significant.

Hypotheses (H2): greater agreement between self- and proxy-assessed cross-sectional summary scores is expected for the more observable domains (i.e. PCS scores will agree more than MCS scores).

The ICC point estimates appeared to support this hypothesis, with PCS-36 scores having agreement considered fair (baseline) to excellent (month 1, 3), while MCS-36 scores demonstrated poor agreement at baseline, and fair agreement thereafter (month 1, 3, and 6). However, the confidence intervals of the ICCs for the PCS-36 and MCS-36 overlapped at each time point except at 1 month, when the PCS-36 clearly demonstrated greater agreement.

Hypotheses (H3): poorer agreement is expected for the EQ-VAS than the other summary scores, because the EQ-VAS reflects both the patient's assessment and *valuation of health, whereas the other summary scores reflect the patient's assessment of health status, and use standard scoring algorithms.*

Point estimates of the Pearson's r and ICCs for EQ-5D VAS scores were lower than all other summary scores with the exception of the MCS-36 scores. There was some overlap of the confidence intervals between the EQ-5D VAS scores and other scores at each time period except for the PCS-36 at 3 months.

4.5.2 Comparison of Patient and Proxy Assessed Change Scores

The incremental change in scores between data collection points ('change scores') were compared on the basis of statistically significant change, magnitude of change, CID, and association/agreement, similar to the cross-sectional scores. Statistically significant differences between assessor types for changes scores between time periods were detected on the EQ-VAS at t_0/t_1 and t_0/t_6 (moderate effect sizes). Significant differences were also observed for PCS-36 scores between 1 and 3 months (ES=0.31), and MCS-36 scores between 3 and 6 months (ES=0.30) (Table 36). All other comparisons between rater assessment change scores were not statistically significant and had effect sizes less than 0.25. For the EQ-Index, HUI2 OUS and HUI 3 OUS, no statistically significant differences were between rater types on the and all effect sizes < 0.20. A clinically important difference between raters was observed for the time periods t_0/t_1 and t_0/t_6 for EQ-VAS (>5); t_0/t_1 and t_3/t_6 for the EQ-Index (>0.036); and between t_0/t_6 and t_3/t_6 for both the HUI2 OUS and HUI3 OUS (>0.03).

Table 36: Comparison of Patient and Proxy Assessed Change Scores

	Patient			Proxy			Difference in Values					
Time	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**n**	**Mean**	**SD**	**t**	**sig**	**ES**
EQ-VAS												
T_0/T_1	86	0.49	23.32	83	11.83	18.92	83	-11.22	24.59	4.15*	0.00	-0.46
T_1/T_3	77	5.81	18.92	75	1.75	16.05	73	2.97	22.92	1.11	0.27	0.13
T_3/T_6	73	2.10	17.63	74	4.86	16.74	72	-3.56	19.91	1.52	0.13	-0.18
T_0/T_6	77	8.88	23.12	76	19.30	23.68	75	-11.69	24.18	4.19*	0.00	-0.48
EQ-Index												
T_0/T_1	86	0.27	0.43	81	0.23	0.37	81	0.05	0.46	0.87	0.39	0.10
T_1/T_3	78	0.07	0.34	74	0.08	0.28	73	0.00	0.33	-0.03	0.98	0.00
T_3/T_6	72	0.02	0.29	72	0.03	0.35	69	-0.04	0.36	-0.94	0.35	-0.11
T_0/T_6	76	0.36	0.41	74	0.33	0.43	72	0.02	0.42	0.38	0.70	0.05
PCS-36												
T_0/T_1	85	2.88	9.84	82	3.64	7.94	81	-1.02	10.28	-0.89	0.38	-0.10
T_1/T_3	74	0.84	8.11	75	3.61	8.35	70	-2.51	8.00	2.63*	0.01	-0.31
T_3/T_6	70	1.82	7.33	74	0.34	8.15	69	0.91	10.19	0.74	0.46	0.09
T_0/T_6	77	5.84	12.46	75	7.21	11.01	74	-1.51	12.05	-1.08	0.28	-0.13
MCS-36												
T_0/T_1	85	3.61	13.58	82	2.91	13.16	81	0.38	16.72	0.21	0.84	0.02
T_1/T_3	74	2.33	12.73	75	-0.73	10.71	70	2.62	13.30	1.65	0.10	0.20
T_3/T_6	70	-0.53	10.73	74	2.86	10.99	69	-4.11	13.75	2.48*	0.02	-0.30
T_0/T_6	77	5.45	13.35	75	5.64	12.36	74	-0.73	15.80	-0.4	0.69	-0.05
HUI2 OUS												
T_0/T_1	79	0.10	0.21	74	0.10	0.18	68	0.01	0.19	0.24	0.81	0.03
T_1/T_3	73	0.02	0.16	69	0.04	0.13	64	-0.01	0.19	-0.37	0.71	-0.05
T_3/T_6	70	0.00	0.18	70	0.03	0.20	65	-0.04	0.21	-1.61	0.11	-0.20
T_0/T_6	70	0.12	0.22	69	0.14	0.24	63	-0.04	0.24	-1.36	0.18	-0.17
HUI3 OUS												
T_0/T_1	77	0.18	0.33	74	0.17	0.29	66	0.01	0.35	0.18	0.85	0.02
T_1/T_3	74	0.04	0.27	69	0.07	0.20	65	-0.02	0.29	-0.63	0.53	-0.08
T_3/T_6	69	0.01	0.25	70	0.05	0.26	64	-0.05	0.32	-1.17	0.25	-0.15
T_0/T_6	67	0.23	0.31	69	0.25	0.35	60	-0.05	0.36	-1.00	0.32	-0.13

* $p < 0.05$

Similar to the cross-sectional scores, Pearson's r and ICCs were nearly identical except where systematic difference arose (mainly in cases where the t-test was significant), in which case the ICCs were slightly lower than Pearson's r (Tables 36 and 37). Regarding the hypotheses for agreement on difference scores:

Hypotheses (H4): greater agreement is expected for the change scores between baseline and 1 month, and 1 month and 3 months, than between 3 and 6 months (as the health status of the patient stabilizes).

Change score agreement was difficult to generalize across measures (Table 37). Considerable overlap occurred for the confidence intervals of the one-way ICCs. The trend of the point estimates for Pearson's r and the ICCs for the EQ-Index, PCS-36, MCS-36 and HUI3 OUS were consistent with the hypothesis (T_0/T_1 and T_1/T_3 > T_3/T_6). For the HUI2 OUS and EQ-VAS, point estimates indicated T_3/T_6 > T_1/T_3, but there was much overlap between confidence intervals as determined by the 1-way ANOVA ICC.

Hypotheses (H5): greater agreement will be observed for the cross-sectional scores than for the change scores.

This hypothesis was generally confirmed, as Pearson's r and the ICCs between proxy and patient assessments were fair to excellent for cross-sectional scores and poor to fair for change scores across all of the summary scores (Table 37).

Table 37: Change Score Agreement between Patient and Proxy

	Extent of Agreement or difference between Patient and Proxy		
Scores	**Pearson's r**	**ICC: one-way (95% CI)**	**Robinson's ICC**
EQ-VAS			
T_0/T_1	.35†	0.26 (0.05, 0.45)	0.25
T_1/T_3	.08	0.08 (-0.15, 0.30)	0.07
T_3/T_6	.20	0.19 (-0.05, 0.40)	0.19
T_0/T_6	.41†	0.33 (0.11, 0.51)	0.33
EQ-Index			
T_0/T_1	.35†	0.34 (0.14, 0.52)	0.34
T_1/T_3	.40†	0.40 (0.18, 0.57)	0.40
T_3/T_6	.15	0.15 (-0.08, 0.37)	0.15
T_0/T_6	.45†	0.45 (0.25, 0.62)	0.45
PCS-36			
T_0/T_1	.36†	0.35 (0.15, 0.53)	0.35
T_1/T_3	.51†	0.48 (0.28, 0.64)	0.48
T_3/T_6	.04	0.04 (-0.19, 0.28)	0.04
T_0/T_6	.48†	0.48 (0.28, 0.63)	0.47
MCS-36			
T_0/T_1	.23*	0.23 (0.01, 0.43)	0.22
T_1/T_3	.38†	0.36 (0.14, 0.55)	0.35
T_3/T_6	.20	0.16 (-0.08, 0.38)	0.15
T_0/T_6	.21	0.22 (-0.01, 0.42)	0.21
HUI2 OUS			
T_0/T_1	.55†	0.54 (0.35, 0.69)	0.54
T_1/T_3	.18	0.18 (-0.07, 0.41)	0.17
T_3/T_6	.28*	0.27 (0.03, 0.48)	0.26
T_0/T_6	.41†	0.40 (0.17, 0.59)	0.39
HUI3 OUS			
T_0/T_1	.38†	0.38 (0.16, 0.57)	0.38
T_1/T_3	.22	0.22 (-0.03, 0.43)	0.21
T_3/T_6	.15	0.15 (-0.10, 0.38)	0.14
T_0/T_6	.40†	0.40, (0.16, 0.59)	0.39

*$p < 0.05$; †$p < 0.005$

4.6.0 Cross-Sectional Construct Validity

4.6.1 Cross-Sectional Construct Validity – Patient Self-Assessment

The cross-sectional construct validity of patient self-assessed scores at baseline were evaluated against *a priori* hypothesized relationships stated in section 3.4.4.

The hypothesized correlations between clinical measures and the patient self-assessed summary scores of HRQL measures were generally confirmed (Appendix 12, Table A). For instance, the CES-D was strongly correlated with MCS-36 ($r = 0.60$; p-value < 0.01), while correlation between the CES-D and PCS-36 was absent, as expected. The HUI2 OUS and HUI3 OUS were strongly correlated with each of the clinical measures, with the exception of moderate correlation with the CES-D. Most of the HRQL summary scores had moderate to strongly correlation with each other. The PCS-36 was moderately correlated with the MCS-36.

Hypothesized correlations between the EQ-5D and single attribute scores from the HUI2 and HUI3 were generally confirmed (Appendix 12, Table C). Most interrelationships were absent, as expected based on the Statistics Canada report (2000). Strong correlations between ambulation (saa3) and mobility (eqmo), saa3 and self-care (eqsc), and saa3 and usual activities (equa) on the EQ-5D were confirmed. Strong correlations were also observed between dexterity and the same EQ-5D domains, but these associations were not hypothesized, as Statistics Canada (2000) did not detect relationships between those domains. Pain and discomfort on the EQ-5D (eqpd) was moderately correlated with HUI3 pain (sap3); a strong correlation was hypothesized. Similarly, anxiety and depression on the EQ-5D (eqad) had a moderate rather than strong correlation, as expected, in relation to HUI3 emotion (sae3).

The mean self-assessed domain and summary scores were compared to population-based norms. The self-assessed scores for the SF-36 were lower than the Canadian normative data relative to every age group (Hopman et al, 2000) for each score (Appendix 11, Table A). A ceiling effect was observed for the SF-36 that has been previously cited (Hopman et al, 2000), and was particularly striking for the role physical domain scores, in which more than half the sample scored 0 for data collection at baseline, 1 month and 3 months. Among the SF-36 domain scores, the mental health scores were the closest to the Canadian norms.

Relative to HUI 3 population-based norms (Source: National Population Health Survey, 1996-97), self-assessed scores (Appendix 11, Table E) were lower than the Canadian normative data relative to every age group for each score. The exception was hearing, where the mean single attribute utility score (sah3) in the stroke sample was comparable to hearing of the age group ages 85 and older (mean score = 0.87). The HUI3 OUS mean score of 0.22 by patient self-assessment at baseline was much lower than the HUI3 OUS mean score norms for males of 91.4 (95%CI: 91.1, 91.7) and females 90.4 (90.1, 90.7).

The distribution of responses by self-assessment on the EQ-5D were compared to reported findings from the Statistics Canada (2000) survey and an Alberta-based general population survey. The extent of problems reported in the stroke population was much greater than in either of the studies aimed at a more general population. The majority of stroke patients reported some or extreme problems on each of the EQ-5D domains at baseline (Appendix 11, Table G). This pattern persisted across the time periods for the domains of mobility, self-care, and pain/discomfort, while the prevalence of problems on the domains of usual activities and anxiety/depression diminished after baseline.

4.6.2 Cross-Sectional Construct Validity – Proxy-Assessment

The cross-sectional construct validity of proxy-assessed scores at baseline were evaluated against *a priori* hypothesized relationships stated in section 3.4.4. The findings were similar to those reported in section 4.6.1 (Appendix 11, Tables B, D, F, G; Appendix 12, Tables B and D).

CHAPTER 5: SUMMARY AND DISCUSSION

5.1.0. Summary of Results

Longitudinal studies of the evaluative properties of the SF-36, EQ-5D, and HUI are necessary to provide evidence on the validity of their application to clinical trials in different patient subgroups. Comparisons between generic health status and HRQL measures in the same set of study participants are useful to developers and users of these instruments as means of understanding their strengths and weaknesses, relative to each other. This study combined both elements by comparing the longitudinal construct validity of 3 families of generic HRQL measures in a cohort of stroke patients. Responsiveness of the summary scores of the HRQL measures was examined for self-assessment and proxy-assessment, and by comparing and contrasting the results for each perspective on patient health. Assessment of the patient from the proxy perspective also provided an external anchor of change for evaluating responsiveness. Collection of self- and proxy-assessments of patient health status facilitated an evaluation of agreement and the potential substitutability of proxy-assessments for self-assessments. This section further discusses the interpretation and limitations of the study design and findings on the evaluative properties of the EQ-5D, HUI and SF-36 and agreement between patient and proxy assessment, followed by implications for clinical trials and future research.

5.1.1 Longitudinal Construct Validity

One of the primary objectives of this study was to evaluate the longitudinal construct validity (LCV) of the summary scores of several commonly used generic health status measures in post-stroke recovery. For the purposes of this study, LCV was operationalized in terms of sensitivity and responsiveness. Sensitivity is the ability to capture statistically significant change regardless of whether those changes are considered meaningful or clinically important. Responsiveness is the ability to detect clinically meaningful change. Central to these assessments of LCV is the determination of changes in health, to which change in health status scores could be compared. This was accomplished using external criteria of change as benchmarks of change in health status changes subsequent to the stroke.

5.1.1.1 Sensitivity: Self-Assessment

Tests of statistical significance demonstrated that all of the summary scores were able to detect change. These tests of 'sensitivity' did not indicate whether the change detected was meaningful or not. However, they did show that the study was adequately powered to detect statistically significant change and/or that patients recovering from stroke experience large changes.

Sensitivity has implications for sample size requirements in that summary scores with relatively higher squared t-test ratios would not require as large a sample size to achieve statistical significance and to avoid a type 2 error, *ceteris parabis*. This study found the HUI2 OUS consistently (according to each of the external anchor-based approaches) provided the greatest power of the summary scores for this study design, with the HUI3 OUS consistently ranking second. These two outcome measures would have required fewer patients to detect a statistically significant difference in pre/post scores by patient self-assessment between baseline and 1 month, and baseline and 6 months. Conversely, the PCS-36 was the lowest ranked of the summary scores in terms of magnitude of change reflected by the change in summary score among patients who were deemed 'changed' using criteria A to C. For these criteria, the PCS-36 would have required a larger sample size than the other outcome measures to detect a statistically significant difference in change scores.

5.1.1.2 Responsiveness: Self-Assessment

All of summary scores demonstrated responsiveness when evaluated on the basis of the absolute values of change scores. The patient self-assessed EQ-VAS, EQ-Index, PCS-36, MCS-36, HUI2 OUS, and HUI3 OUS scores consistently displayed large magnitudes of change (e.g. ES > 0.80), with SRMs greater 1.00 and GRSs greater than 0.80. Every summary score had change scores representing clinically important and meaningful change between baseline and 1 month (evaluated by criteria A and B), and between baseline and 6 months (evaluated by criteria C and D). However, the derivation of responsiveness indices by aggregating patients who improved or declined via absolute values obscured some of the limitations of the methods used to evaluate responsiveness.

Separate calculation of the responsiveness statistics for each of the patient subgroups 'improved', 'no change' and 'declined' clarified some of the potentially misleading statistics generated by using absolute values. Some patients, for instance, self-rated their global health as 'declined' yet had positive change scores. Calculation of responsiveness statistics by subgroups of change more clearly presented the direction and magnitude of change among patients who fell into the categories of change as defined by each external anchor.

Most patients in the cohort were categorized as 'improved' by each external anchor of change. According to all 4 of the external criteria, the self-assessed EQ-Index was the most responsive of the summary scores for patients who 'improved'. Mean change scores in the 'improved' group were in the appropriate (positive) direction and were generally very large, i.e. a multiple of 5 to 10 times the MCID, for the EQ-Index, the HUI2 OUS and HUI3 OUS. This was evident between baseline and 1 month (according to criteria A and B) as well as between baseline and 6 months (according to criteria C and D). Meaningful change was also captured by the PCS-36, MCS-36, and EQ-VAS between baseline and 6 months (criteria C and D). However, the positive increases in scores for patients defined as changed between baseline and 1 month (by criteria A and B) were statistically significant but insufficient to represent a clinically important difference on the PCS-36, MCS-36 and EQ-VAS.

Only criteria A and D categorized sufficient patients (i.e., n >10) in 'no change' category. Change scores associated with patients categorize as 'no change' were expected to be smaller than in the 'improved' or 'declined' group, with trivial effect sizes. This was not the case. Patients classified as 'no change' using criterion A had change scores with medium to large effect sizes for all summary scores except the EQ-VAS. However, the magnitude of the change scores on the EQ-VAS were larger for the 'no change' group (ES=0.21) than for the 'improved' group (ES=0.09), using criterion A. In comparing the extent of noise generated by the summary scores of patients classified as 'no change' by criterion A, the largest responsiveness statistics (greatest noise) were observed in the EQ-Index. For criterion D (movement between categories based on Barthel Index score), magnitudes of change were smaller, with trivial ES < 0.20 for all summary scores, except for the EQ-Index (ES=0.39) and MCS-36 (ES=0.42).

Criterion A was the only approach that categorized more than 10 patients as having 'declined' (n=12). The EQ-VAS, PCS-36 and MCS-36 had self-assessed change scores that were consistent with the expected direction (negative). The magnitude of change scores in this group were much smaller than for those in the 'improved' group.

In ranking the order of summary scores, ES and SRM were typically very similar, but rankings based on the GRS sometimes diverged from those using the SRM and ES. This was the case for the HUI3, which had a very small standard deviation association with OUS change scores from patients categorized as 'stable' (a small subgroup of 7) and as a consequence, usually received the highest ranking based on GRS. This was not viewed as problematic because median ranking of responsiveness indices was used (Birbeck et al, 2000). A single aberration was inconsequential if another measure was superior according to both the ES and SRM.

5.1.1.4 Sensitivity - Proxy Assessment

The findings on the sensitivity of proxy-assessed scores were comparable to self-assessment in terms of magnitude of t-test statistics. The HUI2 OUS, followed by the HUI3 OUS, were the most highly ranked summary scores according to criteria C and D, similar to the findings for patient self-assessed scores. The HUI2 OUS and HUI3 OUS were ranked in the bottom half of the summary scores based on criteria A and B, where the EQ-Index and EQ-VAS were rated first and second. The EQ-VAS and MCS-36 performed noticeably better than their self-assessed counterparts using criteria A and B. This was surprising in the sense that criterion A is based patient self-rating of global health change and the EQ-VAS and MCS-36 may be considered more subjective than some of the other summary scores. Aside from the relative ranking, however, the summary scores were comparable in their relative sensitivity. As a consequence, no particular summary score was advantageous to maximize the power of a study.

5.1.1.5 Responsiveness – Proxy Assessment

Clinically important improvements were represented by the change scores of every summary score for each of the external anchor-based criteria when absolute values were used to derive estimates of responsiveness by combining 'improved' and 'declined' patients into an aggregated ' any change' category. The responsiveness indices for proxy-assessed summary scores consistently ranked the EQ-VAS among the most responsive of the measures. This contrasted with the results for patient self-assessment, where EQ-VAS scores were ranked the lowest based on criteria A or B (criteria involving self-rated global health change). However, effect sizes were very similar for all of the summary scores.

Among patients categorized as 'improved', the EQ-VAS and EQ-index change scores were the top ranked summary scores when assessed by proxy. The MCS-36 consistently received the lowest ranking. For criteria A and B, the PCS-36 and MCS-36 change scores between baseline and 1 month did not reach MCID. All other change scores exceeded the pre-defined MCID levels.

Proxy-assessed EQ-VAS scores did not perform favorably in the 'no change' group, with a medium large ES that reflected the largest amount of instrument noise amongst the outcome measures according to both criteria A and D. Otherwise, the 'no change' group demonstrated medium to large effect sizes for all summary scores except for PCS-36 and MCS-36 according to criteria A. All differences were clinically important with exception of the MCS-36 (criterion D) and PCS-36 (criterion A and D). Criterion D was again the preferred external anchor for identifying patients who changed less than for the other criteria, having smaller mean change scores for each of the summary scores relative to the results for criterion A.

Few conclusions can be drawn from the change scores generated by the patients categorized as 'declined'. Only criterion A (patient rates self) identified more than 10 patients as 'declined', and the only proxy-assessed scores that demonstrated decline in the 'declined' subgroup were the EQ-Index and MCS-36, with trivial (ES=0.10) and small (ES=0.23) magnitudes of change, respectively, based on criterion A. These results raise the possibility that the validity criterion A (patient rates self) as an external anchor of change may be questionable in some stroke patients.

The comparison of proxy-assessed difference scores for the 'change' versus 'no change' groups indicated that criterion D (Barthel Index category change) was the most preferred approach for statistically differentiating between the two groups. This reinforced the same findings as for self-assessed scores.

5.1.2 Comparison of Patient and Proxy Assessment Scores

In general, both self- and proxy-assessed mean scores charted similar recovery patterns. Cross-sectional mean scores, the magnitude of mean change scores and the variance of those scores were comparable between self-assessment and proxy-assessment. The most dramatic improvement in patient health status as measured by the mean summary scores of the cohort were observed between baseline (within 2 weeks of the stroke event) and 1 month later, with a less dramatic slope between 1 and 3 months (Appendix 7). Patient self-assessments appeared to plateau between 3 and 6 months, while the angle of the slope of summary scores by proxy assessment between 3 and 6 months was similar to the slope between 1 and 3 month.

In contrasting the summary scores by assessment perspective, only the EQ-VAS demonstrated that assessment perspective was significantly different over time. This may be attributed to differences in baseline EQ-VAS scores, with self-assessed scores almost 10 points higher than the proxy-assessed scores. Mean summary scores were not significantly different between patient and proxy assessment for the other scores, but the power to detect such differences was low.

Few significant differences in change scores between self- and proxy-assessment were detected. The patient self-assessed HUI2 OUS was statistically significantly larger when patient self-rated global change formed the basis for change groups (criteria A and B). The proxy-assessed change scores were larger for the EQ-VAS scores based on criterion D.

5.1.3 Agreement between Patient and Proxy Assessments

Evaluation of the extent of agreement between self- and proxy assessments of patient HRQL has practical implications for determining whether proxy assessments can reliably substitute for patient self-assessment in studies such as clinical trials. Although no battery of statistical tests and study designs can answer this unequivocally, several tests were performed for insight into the issue of substitutability. Cross-sectional scores by self- and proxy-assessment were examined for systematic differences and extent of association and agreement between the perspectives. Changes in summary scores were evaluated in the same way.

5.1.3.1 Cross-Sectional Agreement

Systematic differences between the mean scores of the self- and proxy assessments were not consistently detected for any specific summary score or at a particular time point. No statistically significant differences in mean scores between self- and proxy assessment were observed for most of the summary scores, even for summary scores that may be considered more subjective (i.e. MCS-36 and EQ-VAS). While not always reaching statistical significance, the difference between self- and proxy-assessed scores exceeded MCID or CID levels at certain times: the EQ-VAS at baseline, the EQ-Index at all time points, and the HUI3 OUS at baseline, 1 month, 3 months. When mean scores differed, self-assessment tended to be higher than proxy-assessment, consistent with the results summarized by Sprangers and Aaronson (1992). The difference between self- and proxy-assessed scores for MCS-36 scores at baseline, 1 and 6 months and the EQ-VAS scores after baseline were not significant and did not represent a clinically important difference.

Point estimates of agreement between patient and proxy assessments using ICCs on cross-sectional HRQL summary scores were fair to excellent when the assessment were performed least one month after baseline. The EQ-VAS and MCS-36 demonstrated no better than fair agreement at any point in time. The EQ-Index generated point estimates at months 1 and 6 that were acceptable (>0.70), as did the PCS-36 after baseline. The HUI 2 and 3 OUS had the highest Pearson's r and ICC at baseline (fair), and generated good agreement thereafter. If the criterion for acceptable agreement were strictly adhered to (ICC> 0.70), no scores were acceptable because all confidence intervals had lower bounds below 0.70. Acceptable point estimates were determined for: the EQ-Index at 1 and 6 months; the PCS-36 at 1, 3, and 6 months; and the HUI2 OUS at 1 and 6 months. The mean systematic differences between self- and proxy assessments appeared to have only a minor impact on agreement, with no more than 3 points difference between the Pearson's r and the ICC in most instances.

5.1.3.2 Change Score Agreement

Agreement between patient self- and proxy-assessed change scores was generally poor to fair, with unacceptable levels of agreement for every summary score. No ICC point estimates exceeded 0.70. The absence of correlation between many of the self- and proxy-assessed changes scores for the interval between 3 and 6 months, particularly on the PCS-36, would support the assertion that most functional recovery in these patients occurs within the first 3 months after stroke (Duncan, 1994; Kelly-Haynes et al, 1989). The pattern of association can be observed on scatterplots of summary scores comparing patient-proxy scores (Appendices 7 and 8). A scatterplot of the self- versus proxy-assessed PCS-36 difference scores shows the dispersion of difference scores around 0 in every direction, conveying absence of agreement between the change scores (Appendix 9: Figure 9k).

5.2.0 Discussion

5.2.1 Longitudinal Construct Validity

The ability to detect clinically meaningful important is a requisite characteristic for a health status or HRQL measure to be validly applied for evaluative purposes. Longitudinal studies of the evaluative properties of the SF-36, HUI2, HUI3, and EQ-5D are necessary to determine the validity of their application in clinical trials. There is a dearth of stroke-based studies evaluating the responsiveness of the SF-36, HUI2/3 or EQ-5D, either alone or in combination with other measures. Previous studies of stroke patients that have investigated the psychometric properties of these investigations have been primarily through cross-sectional study designs (Anderson et al, 1996, Mathias et al, 1997; Dorman et al, 1997a; Dorman et al, 1998; Duncan et al, 1997; Grootendorst et al, 2000).

Much of the work on responsiveness and clinically important differences has been conducted in arthritis/degenerative bone and connective tissue-related conditions (Liang et al, 1985; Deyo and Centor, 1986; Clinch et al, 2001; Stucki et al, 1995; Stucki et al, 1996b; Liang et al, 1990; Ruta et al, 1998; Kazis et al, 1989; Bessette et al, 1998; Brazier et al, 1999) and in conditions that involve the airways (Garratt et al, 2000; Juniper et al, 1996a; Juniper et al, 1997; Ware et al, 1998; Jenkinson et al, 1997). The results of this study not only demonstrated that the summary scores of the EQ-VAS, EQ-Index, HUI2 OUS, HUI3 OUS, MCS-36 and PCS-36 are sensitive to statistically significant improvements, but also that post-stroke recovery involves considerable clinically important change that is captured by these measures.

Relative performance of summary scores of the various generic measures depended on the perspective of the assessor and on the external criteria for global health change. The findings indicated that for self-assessment, the HUI2 OUS and HUI3 OUS provide more power to detect statistically significant differences, and would be preferable if sample size limitations are an issue among a group of patients defined as 'improved' using any of the external criteria of change. For proxy-assessment, the differences in sensitivity between summary scores were minimal. Based on rankings, the HUI2 OUS followed by the HUI3 OUS would be more sensitive according to external anchors of change based on clinical criteria. Using self-assessed global rating of change, the proxy-assessed EQ-VAS and EQ-Index were most sensitivie among patients classfied as 'improved'.

General conclusions have been related in terms of findings on the 'improved' group because the majority of patients were categorized by each of the criterion for grouping global health change as 'improved'. So few patients were available for analysis when evaluating and comparing the summary scores for those classified as 'no change' and 'declined' that discussion will primarily focus on the limitations of those findings. Separate analysis and discussion of responsiveness by subgroups of change is further warranted because responsiveness statisics in the 'improved' group were an order of magnitude larger than changes scores for patients categorized as 'declined'. Such an argument against the combining 'improved' and 'declined' patients using absolute values to calculate responsiveness has been made by Norman et al (1997), but was not illustrated with data. Brazier et al (1999) and Harper et al (1997) presented data that would lend support to the argument that patients who are classified as worse and better have change scores of differing magnitudes.

The present study analysis was not able to assess the evaluative properties of each measure with respect to 'no change'. Unfortunately, examination of the usefulness of the summary scores in the 'no change' and 'declined' groups were limited by the small numbers of patients in those groups. Incorporating the full sample (n=124) in future analysis and including the patients who died among those categorized as 'declined' since baseline will increase the sample size in the 'no change' and 'declined' groups, providing more evidence upon which to make conclusions. For the present analysis, only patient rating self (criterion A) and the Barthel Index-based approach (criterion D) identified more than 10 patients who did not change.

Interestingly, the EQ-Index demonstrated the greatest responsiveness among the patient categorized as 'improved', but also appeared to generate the most noise, with larger magnitudes of change in patients categorized as experiencing 'no change'. This experience with the EQ-Index may be comparable to the findings of Brazier et al (1999), who reported that in terms of magnitude of responsiveness, the EQ-5D Index was comparable to the best-performing dimensions of the SF-36 in a group of patients with knee replacement. This study appears to substantiate claims by Brazier et al (1999) that while the EQ-5D's crude description of status in any given dimension makes it efficient for large changes, its brevity may compromise the ability of the measure to reflect more subtle and diverse changes in health status with scores consistent with subtle change.

Other studies comparing the responsiveness of generic health status and HRQL measures to each other have reported conflicting results, perhaps because the results are highly sample dependent and specific to the condition being studied. Inter-study comparisons are not particularly relevant in the study of responsiveness because the results are sample dependent. More meaningful comparisons between measures can be made when various measures are studied in the same sample cohort, where in principle at least, within patient variance is a constant for comparative purposes.

Findings of other studies have arrived at different conclusions regarding the responsiveness of various health status and HRQL measures, depending on the sample and condition being evaluated. For instance, a pre/post study of magnetic imaging of the knee reported no change in the EQ-VAS (pre= 75.7 (SD=18), post= 75.4 (SD=19);

$p > 0.05$) while the EQ-Index score improved (pre= 61.3 (SD=16), post= 70.0 (SD=19); $p < 0.001$) and significant changes occurred ($p < 0.05$) on 5 domains of the SF-36 (change scores not stated) (Hollingworth et al, 1995). Rheumatology clinic patients who perceived that they felt better had change scores of 9.7 (SD=7.5) on the EQ-Index, 10.3 (SD=16.3) on the EQ-VAS, and large effect sizes (>0.80) on the physical functioning, role limitations (physical), pain, and general health perceptions domains of the SF-36 (Brazier et al, 1999). The domain-based scores of the SF-36 and EQ-Index were favored over the EQ-VAS in both studies, with Brazier et al (1999) commenting on the inconsistent performance of the EQ-VAS. The findings of the present study were similar in the respect that the performance of the EQ-VAS was inconsistent across perspectives, times, and criteria of change.

However, a study of COPD patients by Harper et al (1997) concluded that the EQ-VAS, but not the EQ-Index, was responsive to the minor changes in health which are typical of patient with chronic disease. The EQ-VAS was able to detect statistically significant differences while the EQ-Index was not. Also, a study of treatment for sleep apnoea reported moderately large ES for the MCS-36 and PCS-36, contrasting with an ES=0.24 for the EQ-Index (Jenkinson et al, 1997). The authors cited the EQ-Index's failure to measure the aspects of HRQL related to severe sleep fragmentation, although the mean difference score of 5.0 did exceed the MCID criterion used in the present study. The SF-36 was applied in all of the above mentioned studies with generally favorable reviews, while the EQ-Index and EQ-VAS each met with mixed reviews. The results of the present study contrasted with Jenkinson et al (1997) in the sense that the MCS-36 and

PCS-36 change scores were generally not as responsive as the EQ-Index. However, the MCS-36 and PCS-36 generated less measurement noise than the EQ-Index.

The HUI2 OUS and HUI3 OUS have not been compared in a head-to-head comparison of generic health status measures. A study that focused on the ability of the HUI to respond to health changes in a general population cohort reported that provisional HUI3 scores based on the HUI2 system respond to changes in health status associated with serious chronic illnesses, but changes in the HUI did not always coincide with changes in self-reported health (Kopec et al, 2001). The findings of the present study echo the concern that changes in the HUI2 OUS and HUI3 OUS did not always coincide with self-reported changes (or self-report of no change) in health. Several points to consider are that similar criticism may as also be levelled against the item being used to represent self-reported change in health, and that summary scores such as those generated by the HUI and EQ-Index are not the same concept as self-rated global health. The present study found that benchmarking clinically important differences using clinically-based measure (the Barthel Index) to categorize change in patient global health provided larger change scores in the appropriate direction and less measurement noise when no change occurred relative to a patient self-rating of global health change.

As an external anchor-based criterion, movement between categories on the Barthel Index favorably compared to the other criteria. This criterion was associated with summary scores that had the largest magnitudes of change in patients categorized as 'improved', and the smallest change scores in patients categorized as 'no change'. Preference for an external anchor of change that is not based on retrospective ratings of change (e.g. the respondent was asked if they felt the patient had changed since the last assessment) supports the recommendation by Norman et al (1997) that retrospective methods of computing responsiveness should not be used as a basis for choice of an instrument for applications to clinical trials. Norman and colleagues (1997) assert that patients' judgment of change may be heavily influenced by their present health state, and challenge the presumption that the global scale is independent of the HRQL measure, i.e., the errors of measurement in the two scales are assumed to be uncorrelated. Among the external anchors of change, the Barthel Index-based was associated with the most desirable performance characteristics and was arguably more independent of the patient and proxy assessments of health status than the other criteria.

In the examination of the change score differences between self- and proxy-assessment using the 4 external criteria, few significant results were found. However, patient self-assessed change scores demonstrated a tendency to be larger when patient self-rated global change forms the basis for change groups (criteria A and B). The patient self-assessed HUI2 OUS was statistically significantly higher in this context. Conversely, proxy-assessed change scores may be larger when the criterion is based on a clinical measure, as demonstrated by the EQ-VAS scores using criterion D.

Evaluation of the responsiveness of self- assessed health status was compromised by the small sample size of some of the subgroups categorized by the external criteria. However, proxy-assessment provided convergent validity of the appropriateness of the criteria, and indicated that the external anchors of change were generally sensitive, but not specific, means of identifying patients whose global health changed.

The measurement and subsequent interpretation of change on each of the HRQL measure is the subject of ongoing debate (Hays and Woolley, 2000; Norman et al, 1997;

Neymark et al, 1998; Wyrwich and Wolinsky, 2000; Lydick and Epstein, 1993). Few HRQL measure developers are willing to commit to an MCID. Even if a consensus for some critical value or range of values was agreed upon and promoted by an academic society such as the International Society for Quality of Life Research (ISOQOL), such attempts to define MCIDs are likely to encounter criticism. More positively, research such as this helps to contribute to a general understanding of the interpretability of the difference scores through the combined use of multiple HRQL measures with clinical measures and external anchors of change.

5.2.2 Agreement

Due to the inability of some stroke survivors to self-complete health status questionnaires, proxy assessments have been used in a number of stroke studies. Agreement between self- and proxy-assessments has been studied for several generic health status instruments, including the Health Utilities Index (HUI) (Mathias et al, 1997), the EQ-5D (Dorman et al, 1997b), the Sickness Impact Profile (SIP) (Rothman et al, 1991), and the Health Status Questionnaire (a variation on the SF-36 of the Medical Outcomes Study) (Segal and Schall, 1994). These cross-sectional investigations of agreement between patient-proxy assessment in stroke have had mixed results. All studies generally agreed that proxy responses for the less observable, psychosocial attributes are less predictive of patient responses than proxy responses to the more physically-based, observable attributes. Unlike the present study, no previous study of stroke patients assessed agreement for multiple time points or evaluated the agreement for multiple generic health status measures on the same cohort.

Mathias et al (1997) reported moderate to high agreement in interrater reliability between stroke patients and proxies on the HUI2, suggesting that family caregivers can complete the HUI reliably when patients are unable to do so. Their indicators of agreement for the HUI2 OUS were ICC=0.72 and Pearson R=0.70. Their study was conducted within 3 months of the stroke event, and reassuringly, ICCs were very similar to the results of the present study, which were 0.72 at 1 month, 0.65 at 3 months, and 0.72 at 6 months. The present study found that agreement for the HUI3 OUS was slightly lower than for the HUI2 OUS, between 0.65 and 0.70 after baseline, an outcome Mathias et al (1997) did not evaluate because the scoring algorithm for the HUI3 OUS had yet to be published.

Dorman et al (1997b) concluded that the HRQL information obtained by proxy on the more observable domains of the EQ-5D may be sufficiently valid and unbiased to be useable in most types of trials and surveys, but found poor agreement for the domain that assessed emotional function. The overall EQ-Index based score was not examined in the paper, but agreement between patient and proxy assessment using the EQ-VAS had an overall ICC of 0.49, similar to the results of this study. Such a level of agreement is not acceptable for substitution of proxy-assessment for self-assessment. For the present study, the EQ-Index was found to have an acceptable level of agreement at time points 1 or more months after the stroke event.

Segal and Schall (1994) indicated that proxy agreement for the HSQ (SF-36) scales was poor, with a median ICC of 0.32 for the eight dimensions. Agreement was highest on the physical functioning dimension (ICC = 0.67), but was otherwise poor for the other dimensions that largely consisted of more subjective items. Segal and Schall

(1994) postulated that poorly educated respondents had more difficulty with comprehension of the HSQ items, further detracting from interrater reliability. The PCS-36 and MCS-36 were not reported. For the present study, the MCS-36, being derived primarily from the more subjective items, was found to have inadequate agreement. In contrast, the PCS-36 produced an acceptable level of agreement at 1 month and thereafter.

In comparing the agreement statistics among the summary scores, the HUI2 OUS and HUI3 OUS had the highest Pearson's r and ICC at baseline. The HUI2/3 questionnaire poses concrete response options that may be considered less open to interpretation. Interpretation of patient health status by the proxy may have been more elusive at baseline because of less opportunity to communicate with the patient at time of the initial survey that was completed at hospital prior to discharge. The HUI 2 and 3 OUS performed similarly to the PCS-36 and the EQ-index after baseline.

Agreement between change scores has not been examined in the published literature on stroke. The level of agreement was unacceptable for all of the summary scores at each of the data collection points for the purposes of substituting for patient scores. Interestingly, the mean difference scores at the group level were not statistically significantly different for the EQ-Index, HUI2 OUS or HUI3 OUS at any of the time points, the magnitudes of the differences were trivial (ES < 0.20), and the differences were less than or only slightly more than MCID. It is conceivable that agreement would be higher for changes scores for milder subtypes of stroke, if the sample size had been sufficient to allow for stratification and subgroup analysis.

Although the focus of this evaluation was the agreement and/or association between self- and proxy-assessments of stroke patients, disagreement does not have to be viewed as undesirable in every respect. Multiple viewpoints are valid, especially in stroke. Some stroke patients deny impairment, experiencing a state known as "neglect". Caregivers may recognize such physical limitations. Who is the appropriate source of information on HRQL? It could be argued that both the patient and proxy perspectives are valid, and further study of different perspectives as they correlate with other health outcomes is warranted (Feeny, 1999).

To summarize, the agreement between patient self-assessment and proxy-assessment of stroke patients produces an acceptable level of reliability using the EQ-Index, PCS-36, HUI2 OUS and HUI3 OUS when cross-sectional score is sought a month or more after the initial stroke event. By 6 months, no systematic differences in mean scores were detectable for any of the summary scores at the group level, and all mean difference scores were less than the MCID. Use of two or more consecutive summary scores generated by proxy-assessment for the purpose of substituting for self-assessment is not recommended for any of the generic HRQL measures, as reflected by the lack of acceptable levels of agreement for change scores. Decisions to use proxy-assessment to substitute for patient-assessment should be weighted against the alternatives, such as other perspectives, statistically-driven data imputation, or leaving the response as missing data.

5.3.0 Limitations

5.3.1 Longitudinal Construct Validity Issues

HRQL and health status measures used for evaluative purposes are understood to be sensitive and responsive, yet despite their widespread use, relatively few published studies examine longitudinal construct validity. The present study only focused on the longitudinal construct validity of summary scores. Examination of the domain scores of the health status measures is needed to more comprehensively evaluate longitudinal validity.

A consensus is lacking on what constitutes a responsive measure, and how responsiveness should be quantified (Husted et al, 2000). Ironically, recent attempts to define terminology in the literature have presented different terms for similar concepts (Liang, 2000; Husted et al, 2000).

This study employed several responsiveness statistics (ES, SRM, and GRS), (Husted et al, 2000) each with its own limitations (Hays et al in Staquet, 1998; Norman et al, 1997; Wyrwich and Wolinsky, 2000). The most frequently cited statistic used to compare responsiveness in the literature is the SRM, which has been used alone to evaluate responsiveness in some studies (Garrett et al, 2000; Harper et al, 1997; Ruta et al, 1998; Bouchet et al, 2000), or combined with other metrics in other studies. The ability to demonstrate similar ranking of the measures regardless of the choice of responsiveness statistic, which was generally the case in this study, strengthened the robustness of conclusions.

Because no 'gold standard' exists for establishing change in HRQL, several external anchor-based criteria were used to evaluate responsiveness. Three of the 4 criteria were based on a single global health change question. The cutoff values used to trifurcate the patients as 'improved' 'no change' or 'declined' were based on previously published ranges in the literature (e.g. improved included patients who scored +2 to +7 on a 15-point item of global health change) (Juniper et al, 1996; Jaeschke et al, 1989). The cutoff values applied in this study were from previous studies, but they may not be the optimal cutoff points for change in this patient subgroup. As previously mentioned, the assumption that the external anchor is independent of the health measure is not met for self-assessed scores using criteria A and B, nor for proxy-assessed scores using criterion B. It is also debatable that the clinician-based judgment is independent, as the clinical assessor must rely in part on feelings and perceptions reported to them by the patient (Norman et al, 1997).

Comparisons of difference scores for patients defined as 'declined', 'improved' or 'no change' assisted in distinguishing between problems that where associated with the anchors (e.g. misclassification error using criterion A) and problems potentially related to the completion and scoring of HRQL measures. The direction and magnitude of mean difference scores of patients using the different external anchors of change provided insights into the appropriateness of the external anchors. None of the external anchors of important change in global health successfully demonstrated all 3 of the following characteristics that were intuitively desirable: (1) positive mean change scores that equal or surpass the MCID for a summary score based on patients classified as 'improved'; (2) negative mean change scores that equal or surpass the MCID for a summary score in patients classified as 'declined'; (3) change scores, either positive or negative, that were less than the MCID. Clinically important differences were observed between baseline and 1 month for every summary score in those patients classified as 'no change'. The EQ-VAS, PCS-36 and MCS-36 were the only scores that displayed the appropriate direction of change in the 'declined'subgroup for criterion A. Conclusions about the suitability of the criteria are weakened by the small samples available for subgroup analysis, for instance, the 12 patients assigned to the 'no change' group via criterion A. In comparative terms, criterion D (based on the Barthel Index) was the only external anchor that consistently resulted in larger difference score being obtained for the 'change' group than for the 'no change' group, and those differences were statistically significant for the EQ-Index, HUI2 OUS, and HUI3 OUS.

The use of absolute values to evaluate responsiveness by combining all patients whose HRQL changed (either improved or declined) into the indices of responsiveness was somewhat misleading as difference scores were not always in the direction of change indicated by the external anchor (the global rating of change). Thus, a patient who was classified as 'improved' according to criteria A, for instance, could have a negative difference score between baseline and 1 month and the calculation of responsiveness was not penalized by this logical inconsistency. There were some instances where the group classified as 'no change' had a greater mean difference score than the group classified as 'changed'. This occurred for both the self-assessed and proxy-assessed difference scores (Table 23; Table 29). In one case, the mean difference score for the EQ-Index based scores in the group classified as 'no change' was significantly higher (p-value <0.05) than for the group classified as 'changed' (Table 23).

Further clarification was supplied by the subsequent calculation of responsiveness within subgroups of change, in which the indices of responsiveness were attenuated, in part because absolute values were not used. The magnitude of effect appeared to be different for the improved group versus the declined group. A larger sample size in the 'declined' health group would have provided greater confidence in this observation.

The choice of time periods over which to evaluate responsiveness limited the inferences that could be made on the relative responsiveness of the different summary scores. Responsiveness was evaluated between baseline and 1 month, and baseline and 6 months. The examination of time periods in which considerable important change occurred did not assist in differentiating the ability of the instruments and their summary scores to capture change. Examination of responsiveness between 1 and 3 months, or even 3 and 6 months using the criteria for change is an avenue for future research that may help to gauge the relative responsiveness of each summary score and lend greater insight into ability to capture smaller degrees of meaningful change. Unfortunately, the clinical assessment was not performed at 1 month or 3 months, which precludes the use of criteria C or D for these assessments.

Considering that post-stroke depression is estimated to occur in between 20-50% of stroke patients in the first year after stroke (Hosking et al, 1996; Kotila et al, 1998), the psychosocial burden of morbidity due to stroke reflected by the MCS-36 in this study cohort should be further studied. Mean MCS-36 scores were less than one standard deviation below general population norms. This anomalous finding may be a result of the imposed orthogonal scoring system for the summary scores of the SF-36 using the New England Medical Center scoring system. Despite "convincing empirical evidence favoring the orthogonal principal components in summarizing SF-36 information" (Ware et al, 1994), evidence of problems with the scoring methodology is beginning to accumulate. The pattern of results observed as a result of negative scoring coefficients may not fully reflect cross-sectional differences in functioning or overall health across conditions such as depression (Simon et al, 1998). The MCS appears to overestimate mental health in multiple sclerosis (Nortvedt et al, 2000), and the SF-36 scoring system also appears to hinder responsiveness (Birback et al, 2000). Limitations of the scoring algorithm may also be evaluated by looking at SF-36 domain scores individually, rather than just summary scores (McHorney 1998). Data reanalysis is planned for responses to the SF-36 items using the RAND Health Status Inventory scoring algorithm (Hays, 1998), which allows for correlation between physical and mental component summary scores.

5.3.2 Internal Validity

This natural history study lacked an intervention or control group. In a sense, time is the intervention in a longitudinal natural history study. If all extraneous variables that might affect the outcome measured (health status) can be held constant or eliminated, the research can attribute the observed outcomes (changes in health status as reflected by the measures) to the treatment variable (time). Among the threats to the internal validity of experimental designs (Gall et al, 1996), history, maturation, testing, instrumentation, statistical regression, and experimental mortality can be discussed in the context of this study.

History, which refers to events that happened during the study, was unlikely to contribute serious concern about the changes in health status captured by the measures on a group level. However, on an individual basis, other conditions and events had the potential to affect the assessment of health status, such as a heart attack or divorce.

Maturation refers to changes that occur in study subjects that are unrelated to an intervention. Maturation is undesirable in an experimental design, but in an observational study such as this one, maturation represents the essence of what the present study attempts to measure: natural changes over time (unrelated to an intervention).

Testing-related issues presented numerous potential concerns to this study. To what extent did the approach to survey administration contribute to the observed findings? The ordering of the generic health status measures was not randomized because of the logistic complications and relatively small sample size. The clinical examination was performed prior to the administration of survey at baseline and 6 months. The majority of surveys were completed in the presence of a RA but approximately 15-20% were completed by mailout at 1 and 3 months. Discussion or collusion between patient and proxy assessors was possible. A different RA oversaw the patient and proxy completion of the survey at 1 and 3 months than at baseline and 6 months.

Steps were taken to minimize the potential for bias. All RAs were similarly trained in questionnaire administration and how to deal with questions from participants so as to preserve the integrity of the assessor's responses. The RAs actively discouraged any deliberations between mailout respondents. No statistically significant differences were detected between those aspects of the study that were testable for bias, such as observer bias. Comparisons of agreement on summary scores between mailouts and RA visits had overlapping confidence intervals. However, sample size limited the ability of these tests to detect systematic differences.

A retest within days of the initial assessment performed at baseline was not conducted. This would have instilled greater confidence in the reliability (test-retest) of the survey instruments in this particular patient sample. Knowledge of the test-retest reliability of baseline self-assessments of the EQ-VAS would have been helpful to clarify observed inconsistencies. Due to resource constraints and patient burden, a retest was not planned.

Threats to the internal validity of findings related to the instrumentation may have arisen due to repeated use and greater familiarity with the survey measures. Respondents initially confronted with valuing their health for the first time may have reflected upon it and refined or re-calibrated their valuation. Adaptation to long-term conditions can bias self-assessments of well-being by individuals (Groot, 2000). Recent health problems can affect participants' reporting of limitations, consistent with a recalibration-type response shift (Daltroy et al, 1999), and may be a plausible explanation for the inconsistent behavior of self-assessed EQ-VAS scores, particularly between baseline and 1 month post-stroke. Intuitively the EQ-VAS would appear to be more susceptible to response shift, as it involves both the self-assessment of health and the valuation of health by the patient.

Other instrumentation issues that affect the ability of a HRQL measure to capture important change can arise from ceiling effects, floor effects, and a limited number of responses available for each item. Ceiling effects were observed for all of the summary scores in this study, especially at month 6 of follow-up, where the distributions of scores became skewed to the left (Appendix 6; Appendix 11). The limited number of response options on the EQ-5D may restrict the ability of that measure to detect change when meaningful change occurs because no intermediate response options are available between no problems, some problems and extreme problems (Table 11G and 11H).

Statistical regression (to the mean) presupposes the tendency for individuals whose scores fall at either extreme on a variable (e.g. the EQ-VAS) to score near the mean when the variable is measured a second time. Regression to the mean is primarily a concern when making inferences about the reliability of scores for individual patients. When a similar number of study participants with high and low scores regress to the mean, this tendency is presumably negated at the group-level.

Experimental mortality or attrition is of particular concern in studies of HRQL because there is a possibility that patients who drop out, or are lost to follow-up, are sicker or differ from the patients retained in the study. To alleviate some concern, baseline summary scores and demographic and clinical characteristics of those who were retained for the duration of the study were compared to dropouts. Patient characteristics and summary scores between the groups were generally very similar. No statistically significant differences were found, but the dropout sample was relatively small. The retention rate for both patient self-assessment (77 of 97 completed the 6 month assessment) and proxy (76/97) was almost 80% in both instances. This is a reasonable retention rate for any longitudinal study and is evidence of the considerable efforts expended by the project team to track and follow-up the patients. A similarly designed longitudinal study of stroke outcomes had a retention rate of 42% (Kim et al, 1999).

The ability to detect differences between scores or other characteristics was limited by the sample size. Lack of power to detect systematic differences primarily compromised the conclusive validity of statements regarding systematic differences between patient and proxy assessments and formats of survey administration (RA and mailout). The study had adequate power to detect statistically significant changes in summary scores between the time periods of interest for ascertaining sensitivity, which was the objective addressed by sample size calculations in Table 1.

5.3.3 External Validity

Approximately 62% of all patients who met the selection criteria consented to participate in the study. The patients and/or caregivers who declined to participate sometimes cited the additional stress of a study being too much to handle at the time. Patients without caregivers could not participate. One would anticipate these stroke patients to differ in the extent of their social support. Lack of social support conceivably impacts recovery and HRQL. The absence of this patient subgroup compromises generalizability of the sample. Graphical data illustrated possible differences in the health of patients who dropped out prior to the study endpoint at 6 months. Some degree of selection bias occurred in this study, but it is difficult to predict how it might affect the findings. In relation to those patients who completed the study, the 11 patients who dropped out after baseline had poorer health status, while the 9 patients who dropped out after 1 month had better health status. Further study of unit non-respondents should be pursued, with the possible inclusion of utility scores for those who were dropouts because they died."

Findings on the substitutability of proxy-assessments for self-assessment may inform exploratory research into other conditions requiring proxy-assessment, but cannot be generalized to those conditions.

These results cannot be generalized to all stroke patients, such as patients who are cognitively impaired. This cohort included a greater proportion of moderate and severe stroke in relation to the incidence of stroke as described for larger, comprehensive studies of stroke (Bamford et al, 1990). Patients were recruited within 2 weeks of the stroke event, but not all patients had exactly the same baseline starting point. Some patients suffered more severe strokes, others were temporarily aphasic, and some patients were ready to enter the study within a week but could not be recruited until a suitable proxy was located. Rapid discharge (within hours of admission to hospital) of patients with less debilitating strokes prevented recruitment of such patients into the study. Thus, a greater proportion of moderate and severe stroke patients are believed to be present in this patient sample.

The composition of the stroke cohort probably contributed to an avoidance of a ceiling effect on the Barthel Index at baseline, a limitation that has been associated with its use in patients with milder stroke (Duncan et al, 1994; Williams et al, 1999b). The Barthel Index performed favorably as an external criterion for change but may not perform as well in a patient cohort primarily composed of mild stroke.

The different summary scores reflected slightly different patterns of recovery across the cohort. For both self- and proxy-assessment, the EQ-VAS, PCS-36 and MCS-36 showed more continuous improvement between baseline and 6 months than did the HUI2 OUS, HUI3 OUS and EQ-Index, which appeared to plateau to a greater extent at 3 months. Because the clinical measures were not applied at month 1 or 3, it is difficult to assess whether the clinical condition of patients began to plateau over these same time periods. Thus, this study cannot generalize to what extent the health status measures reflected clinical measures of stroke recovery at months 1 and 3.

The findings on agreement between self- and proxy assessments and interpretations of those findings may not be applicable other conditions or patient groups such as Alzheimer's disease. A more detailed investigation of the agreement between patient and proxy assessment for the specific domains of HRQL and functional ability is needed to better examine describe the findings of this study in relation to other studies.

5.4.0 Study Implications

5.4.1 Implications for Clinical Trials

This study has demonstrated that in the post-stroke period where large magnitudes of change are expected, the SF-36, HUI2, HUI3, EQ-VAS, and EQ-Index are likely to be responsive. The inconsistent performance of the EQ-VAS relative to the other outcome measures, particularly for patient self-assessment, raises concern about the validity of using the EQ-VAS, particularly at baseline when neurological recovery is still taking place. The summary scores of each measure were responsive for both patient and proxy assessment. The HUI2 OUS and HUI3 OUS provided more power for study designs eliciting self-assessment. When using proxy-assessment, the relative sensitivity of the summary scores was similar. However, numerous caveats apply to these generalizations.

Responsiveness was assessed using 4 anchor-based criteria, 3 of which were retrospective rating of global change (criteria A, B, and C). This study exemplified some of the problems with retrospective ratings of global change. Evaluation of responsiveness confounded by retention of those patients whose health improves while those who deteriorate are more likely to be lost to attrition/dead. Giving patient who died utility scores of 0 and including those patients in the "declined" group would have attenuated this bias and increased the sample.

The majority of patients rated themselves as changed, and those who rated themselves as "no change" had summary scores on almost every measure that would indicate otherwise, in some instances having significantly greater change scores than the "improved" or "declined" group. The "no change" and "declined" groups were very small, particularly for criteria A, B and C. Criterion B indicated that there was only 1 patient for whom both patient and proxy agreed no change had taken place. This provides insight into how it was possible the patient-rated (criterion A) "no change" group had large change scores: according to the proxy-rating of global change, they had experienced meaningful change. Criterion C, clinician-assessed global change, was informed by the patient and proxy assessments, yet responsiveness indices were larger when anchored to a clinical measure (criterion D). In order to more comprehensively understand and make conclusions about the suitability of each criteria, the association between changes in summary scores and the gradients of the 15-point global rating scale should be examined.

Between baseline and 1 month, MCIDs or CIDs were achieved for the EQ-Index, HUI2 OUS and HUI3 OUS, while MCIDs or CIDs were not observed for the PCS-36, MCS-36 and EQ-VAS scores in this study. Studies with HRQL outcomes that exceed the MCID or CID are more likely to attract favorable reviews when attempting to demonstrate the benefit of an intervention. This is because the magnitude of the difference is more likely to be interpreted as meaningful.

Instrumentation limitations such as ceiling and floor effects should be considered in relation to the patient group to be studied. The initial burden of morbidity may be a factor in the ability of an instrument to detect change. Patients with greater levels of initial morbidity have more opportunity to improve. A relatively healthy cohort gives less room for responsiveness to be demonstrated.

Trialists may be reluctant to use the EQ-VAS, particularly for self-assessment in stroke patients within the first month of stroke. This study produced self-assessed EQ-VAS scores that were inconsistent with the pattern of recovery captured by the other self-assessed summary scores. However, the proxy-assessed EQ-VAS scores were similar to those exhibited by the other HRQL scores and may represent a preferable alternative to self-assessment in certain situations.

For summary scores on the HRQL measures studied, proxy assessment produced sufficient agreement with patient self-assessment to substitute for a single cross-sectional data collection point using the preference-based index scores (i.e. HUI2 OUS, EQ-5D Index) and the PCS-36 if the data collection occurs at least 1 month post-stroke. HUI3 OUS was on the margin, nearly meeting the criterion for acceptability of an ICC greater than or equal to 0.70. The point estimates of the correlation and agreement statistics would substantiate a conclusion that proxy-assessed HUI2 OUS, EQ-5D, EQ-Index and PCS-36 scores can substitute for patient self-assessment in cross-sectional studies, but the confidence intervals around those estimates somewhat weaken this conclusion. Proxy-assessed EQ-VAS and MCS-36 are of questionable reliability when substituting for patient self-assessment anytime during the 6-month post-stroke recovery, especially at baseline (within 2 weeks of the stroke event).

Agreement between self-assessed and proxy-assessed change scores was at best moderate (<0.60) for the HRQL measures studied. Based on the findings of this study, one should be reticent to recommend of the use of proxy-assessments as a substitute for self-assessment in 2 or more successive periods in order to generate a change score. If such an approach is being seriously considered for a large clinical trial, a pilot study of agreement is recommended.

5.4.2 Future Methodological Research

The scope of this investigation was restricted to examining only summary scores, and much remains to be done in analyzing longitudinal construct validity of the health status measures. Correlations between change scores of the domains and summary scores of HRQL and clinical measures need to be evaluated. Study of the responsiveness of specific attributes and domains of health measured by each instrument will impart greater insight into which aspects of health status are driving the summary scores. Use of the mean is limited as a threshold for scrutinizing the appropriateness of the external anchors as was done in this study, particularly for scores that may not be normally distributed. Thus, comparisons of the suitability of the different external anchors used to assess changes in HRQL scores is planned via Receiver Operand Curves (Deyo et al, 1992) and regression-based approaches (Husted et al, 2000). Number of hospitals days was collected in the study and may serve as an informative external criterion that is not based on a retrospective rating of change.

The relationship between changes in stroke-specific and clinical measures such as the Barthel Index and changes in the different generic health status measures should be explored. Interpretation of changes scores on each measure may be enhanced by studying the relationships between changes on each measure, and an expert panel of clinicians familiar with the interpretation of changes on, for instance, the NIHSS, may be assembled to explore how those changes correspond with the generic measures used in the study. Comparisons of such findings to other benchmarks for representing clinically important differences cited in the literature, such as SEM (Wyrwich et al, 1999) and ES (Samsa et al, 1999; Kazis et al, 1989; Norman et al, 2001) would be useful.

Data reanalysis is planned for responses to the SF-36 items using the RAND Health Status Inventory scoring algorithm (Hays, 1998), which allows for correlation between physical and mental component summary scores. The limitations of the orthogonal scoring algorithm have been discussed in several papers (Simon et al, 1998; Nortvedt et al, 2000). The New England Medical Center's scoring system (Ware et al, 1994) demonstrated less responsiveness than the item-response theory (IRT) based scoring system for the RAND in a recent study (Birback et al, 2000). Further comparisons based on conceptualization (orthogonal versus oblique principal factors) and derivation of scoring systems (simple summation versus IRT-based) are needed to clarify the differences between competing scoring systems for the SF-36 items.

The agreement between assessors on specific HRQL domains remains to be examined. Domain specific findings should be compared to the existing body of evidence for the various measures in stroke (Mathias et al, 1997; Dorman et al, 1997; Segal and Schall, 1994). Incorporating predictor variables such as presence of depression, living with patient, gender, and proxy relationship to patient into multivariate regression models may better explain the differences between self- and proxy-assessments. Comparisons of agreement between different proxy perspectives (e.g.

proxy assessment of patient health status from the perspective of the patient versus proxy assessment of patient health status from the perspective of the proxy) would help to shed light on the extent to which proxy assessment are dependent upon the perspective elicited from the proxy. However, such a research question would have required a different study design.

Construct validity should be examined between the domains and attributes of the generic health status measures and the CES-D, both cross-sectionally and longitudinally. Data were gathered during this study on the health of the proxy (caregiver) by self-assessment, but no analysis has been performed thus far. This represents another avenue of research that may include determining impact of stroke on the health status of caregivers who live versus do not live with the stroke patient. Previous research that examined the relationship between SF-12 and SF-36 in stroke patients found that proxy assessment appeared to be influenced by patient age above and beyond that which may be explained by the health status of the patient (Pickard et al, 1999), but only the patient or proxy assessment of health status was available to the investigators of that study. Further insight into that relationship can be obtained through the data collected in this study.

5.5.0 Conclusions

The summary scores of the HUI2, HUI3, SF-36 and EQ-5D demonstrated that over the course of 6 months post-stroke recovery, health status and HRQL improves dramatically in the majority of patients admitted to hospital for stroke. Improvement primarily occurs in the first 4 to 6 weeks of stroke. External anchors of global health change were employed to compare and facilitate the interpretation of health status and HRQL, and these included: (1) patient self-rating of change; (2) proxy and patient rate and agree on change; (3) clinician-based rating of change; (4) and movement between Barthel Index score-based categories. All external criteria indicated that the majority of patients 'improved'. Among the 'improved' patients, the proxy- and self-assessed scores of the HUI2 OUS, HUI3 OUS, EQ-Index, and PCS-36 (plus the EQ-VAS for proxy-assessed scores) were responsive to meaningful change in stroke patients. Effect sizes for the change scores on the measures were generally medium to large (>0.60) using each of the external anchors. Self- and proxy-assessed HUI2 OUS, HUI3 OUS and EQ-Index scores (and proxy-assessed EQ-VAS) were relatively more sensitive and were associated with larger magnitudes of change. These scores may provide more power to detect statistically significant differences than the other summary scores. The self-assessed EQ-VAS scores demonstrated inconsistencies that raised questions as to its suitability, possibly being subject to response shift, although this cannot be proven with the data available.

The EQ-Index generated more noise than the other outcome measures in the 'no change' group. This may be an indication that the EQ-Index captures large change scores associated with major changes in health, but is unable to reflect smaller change scores when less dramatic changes in health occur. The HUI3 OUS also appeared to reflect this tendency, but to a lesser extent. The criterion based upon movement between Barthel Index score-based categories performed favorably relative to the other criteria in that: 'improved' patients experienced larger magnitudes of change; and patients classified as 'no change' were associated with smaller change scores (i.e. noise), with the exception of the MCS-36.

Cross-sectional mean self- and proxy assessed summary scores for the collective cohort were very similar for most instruments. Minimum clinically important differences between self- and proxy assessment were exceeded, albeit marginally, on the EQ-VAS (baseline), EQ-Index (baseline, 1 month, 3 months, 6 months) and HUI3 OUS (baseline, 1 month, 3 months). Acceptable levels of agreement (ICC > 0.70) between self- and proxy-assessed scores were generally observed on the EQ-Index, PCS-36, and HUI2 OUS after baseline. For these measures, proxy-assessed summary scores may reliably substitute for self-assessment cross-sectionally. Consecutive imputation for the same patient is not recommended due to the generally poor to fair (0.08<ICC<0.55) level of reliability for all change scores.

The small number of patients whose health did not change or declined limited a more comprehensive evaluation of responsiveness of the summary scores of the general health status measures in this study. Internal validity of the study may be compromised by the lack of a 'gold standard' for establishing levels of, and changes in, health status and HRQL, a problem inherent to all studies of responsiveness of HRQL measures. The use of multiple criteria to qualify change was an attempt to mitigate this concern. Lack of a retest to evaluate reliability of scores, follow-up of patients and proxy with different observers, and loss to follow-up are all potential threats to the validity of the conclusions of this study. The generalizability of the findings apply, in the strictest sense, to stroke patients who would meet the selection criteria, such as patients who are not cognitively impaired.

BIBLIOGRAPHY

Adams HP Jr, Davis PH, Leira EC, Chang K-C, Bendixen BH, Clarke WR, Woolson RF, Hansen MD. Baseline NIH Stroke Scale score strongly predicts outcome after stroke. Neurology 1999;53:126-131.

AHCPR Publication no. 95-0062. Post-Stroke Rehabilitation Clinical Guideline No. 16. May 1995. [On-line <http://www.text.nlm.nih.gov>].

Algina J. Comment on Bartko's "on various intraclass correlation reliability coefficients". Psychological Bulletin 1978;85(1):135-138.

Anderson C, Laubscher S, Burns R. Validation of the Short Form 36 (SF-36) health survey questionnaire among stroke patients. Stroke 1996;27:1812-1816.

Bamford J, Sandercock P, Dennis M, Burn J, Warlow C. Classification and natural history of clinically identifiable subtypes of cerebral infarction. Lancet 1991;337:1521-1526.

Barkto JJ. On various intraclass correlation reliability coefficients. Psychological Bulletin 1976;83(5):762-765.

Beck AT, Ward CH, Mendelson M. An inventory for measuring depression. Archives of General Psychiatry 1961;4:561-571.

Bendz M. Rules of relevance after a stroke. Social Science and Medicine 2000;51:713-723.

Berzon RA. Chapter 1: understanding and using health-related quality of life instruments within clinical research studies. In: Quality of Life Assessments in Clinical Trials. Edited by MJ Staquet, RD Hays, PM Fayers, Oxford University Press, Oxford, 1998: 3-18.

Bessette L, Sangha O, Kuntz KM, Keller RB, Lew RA, Fossel AH, Katz JN. Comparative responsiveness of generic versus disease-specific and weighted versus unweighted health status measures in carpal tunnel syndrome. Medical Care 1998;36(4):491-502.

Birbeck GL, Kim S, Hays ED, Vickrey BG. Quality of life measures in epilepsy. Neurology 2000;54:1822-1827.

Bouchet C, Guillemin F, Paul-Dauphin A, Briancon S. Selection of quality-of-life measures for a prevention trial: a psychometric analysis. Controlled Clinical Trials 2000;21:30-43.

Boyle MH, Furlong W, Feeny D, Tormace GW, Hatcher J. Reliability of the Health Utilities Index-Mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. Qual Life Res 1995;4:249-257.

Brazier JE, Harper R, Munro J, Walters SJ, Snaith ML. Generic and condition-specific outcome measures for people with osteoarthritis of the knee. Rheumatology 1999;38:870-877.

Brooks R. EuroQol: the current state of play. Health Policy 1996;37(1):53-72.

Brott T, Adams HP, Olinger CP, Marler JR, Barsan WG, Biller J, Spilker J, Holleran R, Eberle R, Hertzberg V, et al. Measurement of acute cerebral infarction: a clinical examination scale. Stroke 1989;20:864-870.

Buck D, Jacoby A, Massey A, Ford G. Evaluation of measures used to assess quality of life after stroke. Stroke 1999;31(8):2004-2010.

Canadian Coordinating Office for Health Technology Assessment. Guidelines for economic evaluation of pharmaceuticals: Canada. Second edition. Ottawa: Canadian Coordinating Office for Health Technology Assessment (CCOHTA); 1997.

Chan B, Hayes B. Cost of stroke in Ontario, 1994/95. JAMC 1998;159(Supp 6):S2-S8.

Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am J Ment Defic. 1981;86:127-137.

Clinch J, Tugwell P, Wells G, Shea B. Individualized functional priority approach to the assessment of HRQL in rheumatology. Journal of Rheumatology 2001;28:445-451.

Coast J, Peters TJ, Richards SH, Gunnell DJ. Use of the EuroQol among elderly acute care patients. Quality of Life Research. 1998;7:1-10.

Coen RF. Chapter 14: Individual Quality of Life and Assessment by Carers or 'Proxy' Respondents. In: Individual quality of life: approaches to conceptualisation and assessment, edited by Joyce CRB, McGee HM, O'Boyle CA, Harwood Academic Publishers, Netherlands, 1999: 185-196.

Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality of life instruments. Pharmacoeconomics 2000;17(1):13-35.

Curran D, Fayers PM, Molenberghs G, Machin D. Chapter 14: Analysis of incomplete quality of life data in clinical trials In: Quality of Life Assessments in Clinical Trials, Second Edition, edited by Staquet MJ, Hays RD, Fayers PM, Oxford University Press, New York, 1998: 249-280.

D'Olhaberriague L, Litvan I, Mitsias P, Mansbach HH. A reappraisal of reliability and validity studies in stroke. Stroke 1996;27:2331-2336.

Daltroy LH, Larson MG, Eaton HM, Phillips CB, Liang MH. Discrepancies between self-reported and observed physical function in the elderly: the influence of response shift and other factors. Social Science and Medicine 1999;48:1549-1561.

De Haan R, Aaronson N, Limburg M, Langton Hewer R, van Crevel H. Measuring quality of life in stroke. Stroke. 1993;24:320-327.

Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. Journal of Chronic Diseases 1986;39(11):897-906.

Deyo RA, Diehr P, Patrick DL. Reproducilibility and responsiveness of health status measures. Controlled Clinical Trials 1991;12:142S-158S.

Dolan P. Modeling valuations for EuroQol health states. Medical Care 1997;35(11):1095-1108.

Dombovy ML, Basford JR, Whisnant JP, Bergstrahl EJ. Disability and use of rehabilitation services following stroke in Rochester, Minnesota, 1975-1979. Stroke. 1987;18:830-836.

Dorman PJ, Slattery JM, Farrell B, Dennis MS, Sandercock PA, and the United Kingdom Collaborators in the International Stroke Trial. A randomized comparison of the EuroQol and SF-36 after stroke. BMJ 1997a; 315:461.

Dorman PJ, Slattery JM, Farrell B, Dennis MS, Sandercock PA, and the United Kingdom Collaborators in the International Stroke Trial. Qualitative comparison of the reliability of health status assessments with the EuroQol and SF-36 questionnaires after stroke. Stroke 1998;29:63-68.

Dorman PJ, Waddell F, Slattery JM, Dennis M, Sandercock PA, and the United Kingdom Collaborators in the International Stroke Trial. Are proxy assessments of health status after stroke with the EuroQol questionnaire feasible, accurate and unbiased? Stroke 1997b;28(10):1883-1887.

Drummond MF, O'Brien BJ, Stoddart GL, Torrance GW. Methods for the Economic Evaluation of Health Care Programmes. Second Edition. Toronto: University Oxford Press, 1997:164.

Duncan PW, Jorgensen HS, Wade DT. Outcome measures in acute stroke trials: a systematic review and some recommendations to improve practice. Stroke 2000; 31(6):1429-1438.

Duncan PW, Samsa GP, Weinberger M, Goldstein LB, Bonito A, Witter DM, Enarson C, Matchar D. Health status of individuals with mild stroke. Stroke 1997;28:740-745.

Duncan PW. Stroke disability. Physical Therapy 1994;74:399-407.

Duncan PW, Wallace D, Sue ML, Johnson D, Embretson S, Laster LJ. The stroke impact scale version 2.0: Evaluation of reliability, validity, and sensitivity to change. Stroke 1999;30(10):2131-2140.

Epstein AM, Hall JA, Tognetti J, Son LH, Conant L, Jr. Using proxies to evaluate quality of life. Can they provide valid information about patients' health status and satisfaction with medical care? Medical Care 1989;27:S91-S98.

Evans RL, Noonan WC, Bishop DS, Hendricks RD. Caregiver assessment of personal adjustment after stroke in a Veterans Administration Medical Center outpatient cohort. Stroke 1989;20(4):483-487.

Fairclough DL. Chapter 13: Methods of analysis for longitudinal studies of health-related quality of life. In: Quality of Life Assessments in Clinical Trials. Edited by MJ Staquet, RD Hays, PM Fayers, Oxford University Press, Oxford, 1998: 227-247.

Fairclough DL. Imputation for non-randomly missing QOL data in longitudinal studies. ISOQOL 2000 workshop.

Feeny DH. Personal communication, November 27, 1998.

Feeny DH, Torrance, Furlong WJ. Chapter 26: Health Utilities Index. In: Quality of Life and Pharmacoeconomics in Clinical Trials, Second Edition, edited by B.Spilker, Lippincott-Raven Publishers, Philadelphia, 1996: 239-251.

Feeny, David. The assessment of health-related quality of life. What roles should it play in pediatric oncology? Med Pediatr Oncol 1999;33(3):184 (Abstract).

Furlong W. Personal communication, September 22, 2000.

Furlong W, Feeny D, Torrance GW, Goldsmith CH, DePauw S, Denton M, Zhu Z, Boyle M. Multiplicative multi-attribute utility function for the Health Utilities Index Mark 3 (HUI3) System: a technical report. McMaster University Centre for Health Economics and Policy Analysis Working Paper No. 98-11, December 1998.

Gall MD, Borg WR, Gall JP. Experimental designs, part 1. In: Educational research: an introduction. 6th edition. White plains NY: Longman, 1996: 463-504.

Garrett AM, Hutchinson A, Russell I. Patient-assessed measures of health outcome in asthma: a comparison of four approaches. Respiratory Medicine 2000;94:597-606.

Glossary [Medical Care]: Health outcomes methodology. Medical Care 2000;38(9):7-14.

Golomb BA, Vickrey BG, Hays, RD. A review of health-related quality of life measures in stroke. Pharmacoeconomics 2001;19(2):155-185.

Granger CV, Seltzer GB, Fishbein CF. Chapter 9: The completed stroke. In: Primary Care of the Functionally Disabled, Lippincott-Raven Publishers, Philadelphia, 1987: 191-209.

Groot W. Adaptation and scale of reference bias in self-assessments of quality of life. Journal of Health Economics 2000;19:403-420.

Grootendorst P, Feeny D, Furlong W. Health Utilities Index Mark 3: Evidence of construct validity for stroke and arthritis in a population health survey. Medical Care 2000;38(3):290-299.

Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. Journal of Chronic Disease. 1987;40(2):171-178.

Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Annals of Internal Medicine 1993;118(8):622-629.

Guyatt GH, Jaeschke R, Feeny DH, Patrick DL. Chapter 5: Measurements in clinical trials: choosing the right approach, Second Edition, edited by B.Spilker, Lippincott-Raven Publishers, Philadelphia, 1996: 41-48.

Guyatt GH, Juniper EF, Griffith LE, Feeny DH, Ferrie PJ. Children and adult perceptions of childhood asthma. Pediatrics 1997;99(2):165-168.

Hackett ML, Duncan JR, Anderson CS, Broad JB, Bonita R. Health-related quality of life among long-term survivors of stroke : results from the Auckland Stroke Study, 1991-1992. Stroke, 30(12):2585-91, 1999 Dec (47 ref) 2000; 31(2):440-447.

Hakim AM, Silver F, Hodgson C. Organized stroke care: A new era in stroke prevention and treatment. JAMC 1998;159(Supp 6):S1.

Han B, Haley WE. Family caregiving for patients with stroke: review and analysis. Stroke 1999; 30(7):1478-1485.

Harper R, Brazier JE, Waterhouse JC, Walters SJ, Jones NMB, Howard P. Comparison of outcome measures for patients with chronic obstructive pulmonary disease (COPD) in an outpatient setting. Thorax 1997;52:879-887.

Hays RD, Anderson R, Revicki D. Chapter 10: Assessing reliability and validity of measurement in clinical trials. In: Quality of Life Assessments in Clinical Trials. Edited by MJ Staquet, RD Hays, PM Fayers, Oxford University Press, Oxford, 1998: 169-182.

Hays RD, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. Quality of Life Research. 1993;2:441-449.

Hays RD, Vickrey BG, Hermann BP, Perrine K, Cramer J, Meador K, Spritzer K, Devinsky O. Agreement between self reports and proxy reports of quality of life in epilepsy patients. Qual Life Res 1995;4:159-168.

Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality of life research: how meaningful is it? Pharmacoeconomics 2000;18(5):419-423.

Hays RD. R36 HSI: Rand-36 Health Status Inventory. Psychological Corporation, San Antonio, CA: Harcourt Brace and Company, 1998.

Hillers TK, Guyatt GH, Oldridge N, Crowe J, Willan A, Griffith L, Feeny DH. Quality of life after myocardial infarction. Journal of Clinical Epidemiology. 1994;47(11):1287-1296.

Hodgson C. Prevalence and disabilities of community-living seniors who report the effects of stroke. JAMC 1998;159(Supp 6):S9-S14.

Hollingworth W, Mackenzie R, Todd CJ, Dixon AK. Measuring changes in quality of life following magnetic resonance imaging of the knee: SF-36, EuroQol or Rosser index? Quality of Life Research 1995;4:325-334.

Hop JW, Rinkel JE, Algra A, van Gijn J. Quality of life in patients and partners after aneurysmal subarachnoid hemorrhage. Stroke 1998;29:798-804.

Hopman WM, Towheed T, Anastassiades T, Tenenhouse A, Poliquin S, Berger C, et al. Canadian normative data for the SF-36 health survey. CMAJ 2000;163(3):265-271.

Hosking SG, Marsh NV, Friedman PJ. Poststroke depression: prevalence, course and associated factors. Neuropsychology Review 1996;6(3):107-133.

Hulley SB, Cummings SR. Appendix 15 B in Designing clinical research: an epidemiologic approach. Baltimore MD, Williams and Wilkins, 1988.

Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. Journal of Clinical Epidemiology 2000;53:459-468.

Indredavik B, Bakke F, Slodahl SA, Rokseth R, Haheim LL. Stroke unit treatment improves long-term quality of life. Stroke. 1998;29:895-899.

Jaeschke R, Singer J, Guyatt GH. Measurement of health status: Ascertaining the minimal clinically important difference. Controlled Clinical Trials 1989;10:407-415.

Jenkinson C, Stradling J, Petersen S. Comparison of three measures of quality of life outcome in the evaluation of continuous positive airways pressure therapy for sleep apnoea. Journal of Sleep Research 1997;6:199-204.

Johnson JA, Coons SJ, Ergo A, Szava-Kovats G. Valuation of EuroQOL (EQ-5D) health states in an adult US sample. Pharmacoeconomics 1998;13(4):421-33.

Johnson JA, Pickard AS. Comparison of the EQ-5D and SF-12 health surveys in a general population survey in Alberta, Canada. Medical Care 2000;38(1):115-21.

Jorgensen HS, Kammersgaard LP, Nakayama H, Raaschou HO, Larsen K, Hubbe R, Skyhoj Olsen. Treatment and rehabilitation on a stroke unit improves 5-year survival. Stroke 1999;30:930-933.

Jonkman EJ, de Weerd AW, Vrijens NLH. Quality of life after a first ischemic stroke. Acta Neurol Scand 1998;98:169-175.

Jorgensen HS, Nakayama H, Raaschou HO, Vive-Larsen J, Stoier M, Olsem TS. Outcome and time course of recovery in stroke. Part II: time course of recovery. The Copenhagen stroke study. Arch Phys Med Rehabil 1995;76:406-412.

Juniper EF, Guyatt GH, Feeny DH, Ferrie PJ, Griffith LE, Townsend M. Measuring quality of life in children. Qual of Life Res 1996;5:35-46.

Juniper EF, Guyatt GH, Feeny DH, Griffith LE, Ferrie PJ. Minimum skills required by children to complete health-related quality of life instrument for asthma: comparison of measurement properties. Eur Respir J 1997;10:2285-2294.

Juniper EF, Guyatt GH, Jaeshke R.. Chapter 6: How to develop and validate a new health-related quality of life instrument. In: Quality of Life and Pharmacoeconomics in Clinical Trials, Second Edition, edited by B.Spilker, Lippincott-Raven Publishers, Philadelphia, 1996: p.54.

Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Medical Care 1989;27:S178-S189.

Kelly-Hayes M, Wolf PA, Kase CS, Gresham GE, Kannel WB, D'Agostino RB. Time course of functional recovery after stroke: the Framingham Study. Journal of Neurological Rehabilitation. 1989;3(2):65-70.

Kim T, Smurawska L, Norris JW, Bayer N, Nadareishvili Z, Oh PI. Stroke outcomes study – prospective evaluation of health-related quality of life and cost in acute stroke (abstract unpublished) Neurology Outcomes Research Sessions, October 10th, 1999.

Knapp P, Hewison J. Disagreement in patient and carer assessment of functional abilities after stroke. Stroke 1999;30:934-938.

Kopec JA, Schultz SE, Goel V, Williams JI. Can the Health Utilities Index measure change? Medical Care 2001;39(6):562-574.

Kotila M, Waltimo O, Niemi ML, Laaksonen R, Lempinen M. The profile of recovery from stroke and factors influencing outcome. Stroke 1984;15:1039-1044.

Kotila M, Numminen H, Waltimo O, Kaste M. Depression after stroke. Stroke 1998;29:368-372.

Kressin NR, Spiro III A, Skinner KM. Negative affectivity and health-related quality of life. Medical Care 2000;28(8):858-867.

Kwa VIH, Limburg M, de Haan RJ. The role of cognitive impairment in the quality of life after ischaemic stroke. Journal of Neurology 1996; 243(8):599-604.

Lai SM, Duncan PW. Evaluation of the American Heart Association Stroke Outcome Classification. Stroke 1999;30(9):1840-3.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;22:159-177.

Lenert L, Kaplan RM. Validity and interpretation of preference-based measures of health-related quality of life. Medical Care 2000;38(suppl2):138-150.

Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 1985;28:542-547.

Liang MH, Fossel AH, Larson MG.Comparisons of five health status instruments for orthopedic evaluation. Medical Care 1990;28(7):632-42.

Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. Medical Care 2000;38(suppl2):84-90.

Locke BA, Putman P. Center for Epidemiological Studies Depression Scale (CES-D). Epidemiology and Psychpathology Research Branch, Division of Epidemiology and Services Branch, Public Health Service, National Institutes of Health, National Institutes of Mental Health. 1971.

Lydick E, Epstein RS. Interpretation of quality of life changes. Quality of Life Research 1993;2:221-226.

MacKeigan LD, Pathak DS. Overview of health-related quality of life measures. American Journal of Hospital Pharmacy. 1993;49:2236-2245.

Magaziner J, Simonsick EM, Kashner TM, Hebel JR. Patient-proxy response comparability on measures of patient health and functional status. Journal of Clinical Epidemiology 1988;41:1065-1074.

Mahoney F, Barthel D. Functional evaluation: the Barthel Index. Maryland Medical Journal 1965;14:61-65.

Manocchia M, Bayliss MS, Conner J, Keller SD, Shiely JC. SF-36 health survey annotated bibliography second edition (1988-1996). Boston, MA: The Health Assessment Lab, New England Medical Center, 1998.

Marshall GN, Hays RD, Nicholas R. Evaluating agreement between clinical assessment methods. International Journal of Methods in Psychiatric Research 1994;4:249-257.

Mathias SD, Bates MM, Pasta DJ, Cisternas MG, Feeny D, Patrick DL. Use of the Health Utilities Index with stroke patients and their caregivers. Stroke 1997;28(10):1888-1894.

Mayo NE, Wood-Dauphinee S, Cote R, Gayton D, Carlton J, Buttery J et al. There's no place like home: An evaluation of early supported discharge for stroke. Stroke 2000a; 31(5):1016-1023.

Mayo N, Poissant L, Wood-Dauphinee S, Clarke A. A stroke-specific module to the EQ-5D. Discussion Paper, EuroQol Group Scientific Meeting, Pamplona, Spain, September 28-29, 2000b.

McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. Second Edition. New York: Oxford University Press, 1996.

McHorney CA, Ware JE, Raczek AE. The MOS 36-item short-form health survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Medical Care 1993;31:247-263.

McHorney CA. Methodological inquiries in health status assessment. Medical Care 1998;36(4):445-448.

McVilly KR, Burton-Smith RM, Davidson JA. Concurrence between subject and proxy ratings of quality of life for people with and without intellectual disabilities. Journal of Intellectual and Developmental Disability 2000;25(1):19-39.

Moinpour CM, Lyons B, Schmidt SP, Chansky K, Patchell RA. Substituting proxy ratings for patient ratings in cancer clinical trials: an analysis based on a Southwest Oncology Group trial in patients with brain metastases. Quality of Life Research 2000;9:219-231.

Neumann PJ, Kuntz KM, Leon J, Araki SS, Hermann RC, Hsu M-A, Weinstein MC. Health Utilities in Alzheimer's Disease. Medical Care 1999;37(1):27-32.

Neymark N, Kiebert W, Torfs K, Davies L, Fayers P, Hillner B, Gelber R, Guyatt G, Kind P, Machin D, Nord E, Osoba D, Revicki D, Schilman K, Simpson K. Methodological and statistical issues of quality of life (Qol) and economic evaluation in cancer clinical trials: report of a workshop. Eur J Cancer 1998;34(9):1317-1333.

Norman GR, Sridhar FG, Guyatt GH, Walter SD. Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. Med Care 2001;39(10):1039-47.

Norman GR, Stratford P, Reghr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. Journal of Clinical Epidemiology 1997;50(8):869-879.

Nortvedt MW, Riise T, Myhr K-M, Nyland HI. Performance of the SF-36, SF-12 and RAND-36 summary scales in a multiple sclerosis population. Medical Care 2000;38:1022-1028.

O'Mahony PG, Rodgers H, Thomson RG, Bobson R, James OFW. Is the SF-36 suitable for assessing health status of older stroke patients? Age and Ageing. 1998;27:19-22.

Ontario Ministry of Health Report of the Joint Stroke Strategy Working Group, June 2000.

Patrick DL, Chiang Y-P. Measurement health outcomes in treatment effectiveness evaluations: conceptual and methodological challenges. Medical Care 2000a;38(suppl2):14-25.

Patrick DL, Chiang Y-P. Preface: convening health outcomes methodologists. Medical Care 2000b;38(suppl2):3-6.

Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. Medical Care 1989;27:S217-232.

Patrick DL, Erickson P. Health status and health policy: quality of life in health care evaluation and resource allocation. New York: Oxford University Press, 1993.

Pickard AS, Johnson JA, Penn A, Lau F, Noseworthy T. Replicability of SF-36 summary scores by the SF-12 in stroke patients. Stroke 1999;30(6):1213-7.

Pickard AS, Topfer LA, Feeny DH. A rapid structured review of studies on health-related quality of life and economic evaluation in pediatric acute lymphoblastic leukemia. Institute of Health Economics, Edmonton, AB, Working Paper 00-1, 2000.

Pickard AS, Weijnen ThJG, Niewenhuizen MGM, Johnson JA, de Charro FTh. A comparison of Canadian and European VAS-based valuations of EQ-5D health states (abstract). The Canadian Journal of Clinical Pharmacology 2001 Vol 8(1):23.

Pierre U, Wood-Dauphinee S, Korner-Bitensky N, Gayton D, Hanley J. Proxy use of the Canadian SF-36 in rating health status of the disabled elderly. J Clin Epidemiol 1998;51(11):983-990.

Portney LG, Watkins MP. Foundations of Clinical Research: applications to practice. New Jersey: Prentice Hall Health, 2000.

Postulart D, Adang EMM. Response shift and adaptation in chronically ill patients. Medical Decision Making 2000;20:186-193.

Radloff LS, Locke BZ. Chapter 9: The community mental health assessment survey and the CES-D scale. In: Weissman MM, Myers JK, Ross CE, eds. Community surveys of psychiatric disorders. New Jersey: Rutgers University Press, 1986:177-189.

Rankin J. Cerebral vascular accidents in people over the age of 60. II. Prognosis. Scot Med J. 1957;2:200-215.

Revicki DA, Gold K, Buckman D, Chan K, Kallich JD, Woolley JM. Imputing physical health status scores missing owing to mortality. Medical Care 2001;39:61-71.

Robinson WS. The statistical measurement of agreement. American Sociological Review 1957;22(1):17-25.

Ronning OM, Guldvog B. Outcome of subacute stroke rehabilitation: a randomized control trial. Stroke 1998;29:779-784.

Rothman ML, Hedrick SC, Bulcroft KA, Hickam DH, Rubenstein LZ. The validity of proxy-generated scores as measures of patient health status. Medical Care 1991;29:115-124.

Ruta DA, Hurst NP, Kind P, Hunter M, Stubbings A. Measuring health status in British patients with rheumatoid arthritis: reliability, validity and responsiveness of the short form 36-item health survey (SF-36). British Journal of Rheumatology 1998;37:425-436.

form 36-item health survey (SF-36). British Journal of Rheumatology 1998;37:425-436.

Samsa GP, Edelman D, Rothman ML, Williams GR, Lipscomb J, Matchar D. Determining clinically important differences in health status measures. Pharmacoeconomics 1999;15(2):141-155.

Samsa GP, Matchar DB, Goldstein L, Bonito A, Duncan PW, Lipscomb J et al. Utilities for major stroke: results from a survey of preferences among persons at increased risk for stroke. Am Heart J 1998; 136(4 Pt 1):703-713.

Scandinavian Stroke Study Group. Multicenter trial of hemodilution in ischemic stroke: Background and study protocol. Stroke 1985;16(5):885-890.

Segal ME, Schall RR. Determining functional/health status and its relation to disability in stroke survivors. Stroke 1994;25(12):2391-2397.

Segal ME, Schall RR. Life satisfaction and caregiving stress for individuals with stroke and their primary caregivers. Rehabilitation Psychology 1996; 41(4):303-320.

Segatore M. Understanding central post-stroke pain. Journal of Neuroscience Nursing 1996;28(1):28-35.

Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 1979;86(2):420-428.

Shuaib, A. Personal communication, June 18th, 2001.

Simon GE, Revicki DA, Grothaus L, Vonkorff M. SF-36 summary scores: are physical and mental health truly distinct? Medical Care 1998;36:567-572.

Sneeuw KCA, Aaronson NK, Osoba D, Muller MJ, Hsu MA, Yung WK, Brada M, Newlands ES. The use of significant other as proxy raters of the quality of life of patients with brain cancer. Medical Care 1997a;35:490-506.

Sneeuw KCA, Aaronson NK, de Haan RJ, Limburg M. Assessing the quality of life after stroke: the value and limitations of proxy ratings. Stroke 1997b;28:1541-1549.

Sneeuw KCA, Aaronson NK, Sprangers MAG, Detmar SB, Wever LDV, Schornagel JH. Comparison of patient and proxy EORTC QLQ-C30 ratings in assessing the quality of life of cancer patients. Journal of Clinical Epidemiology 1998;51(7):617-631.

Solomon NA, Glick HA, Russo CJ, Lee J, Schulman KA. Patient preferences for stroke outcomes. Stroke 1994;25:1721-1725.

Spilker B, Revicki DA. Chapter 3: Taxonomy of quality of life. In: Quality of Life and Pharmacoeconomics in Clinical Trials, Second Edition, edited by B.Spilker, Philadelphia: Lippincott-Raven Publishers, 1996: 25-31.

Sprangers MA, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. J Clin Epidemiol 1992; 45(7):743-760.

Sprangers MAG, Schwartz CE. Integrating response shift into health-related quality of life research: a theoretical model. Social Science and Medicine 1999;48:1507-1515.

Statistics Canada (Belanger A, Berthelot J-M, Guimond E, Houle C). A head-to-head comparison of two generic health status measures in the household population: McMaster Health Utilities Index (Mark 3) and the EQ-5D. April 2000.

Stratford PW, Binkley JM, Riddle DL. Health status measures: strategies and analytic methods for assessing change scores. Phys Ther 1996;76(10):1109-1123.

Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. Second edition. Oxford: Oxford University Press 1995.

Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. Journal Clinical Epidemiology 1996a;49(7):711-7.

Stucki G, Daltroy L, Liang MH, Lipson SJ, Fossel AH, Katz JN. Measurement properties of a self-administered outcome measure in lumbar spinal stenosis. Spine 1996b;21(7):796-803.

Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. Journal of Clinical Epidemiology 1995;48(11):1369-78.

Stucki G, Liang MH, Stucki S, Katz JN, Lew RA. Application of statistical graphics to facilitate selection of health status measures for clinical practice and evaluative research. Clinical Rheumatology 1999;18(2):101-5.

Sulter G, Steen C, De Keyser J. Use of the Barthel Index and Modified Rankin Scale in acute stroke trials. Stroke 1999;30:1538-1541.

Thompson SC, Bundek, NI, Sobolew-Shubin A. The caregivers of stroke patients: an investigation of factors associated with depression. Journal of Applied Social Psychology 1990;20(2):115-129.

Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multi-Attribute Preference Functions for A Comprehensive Health Status Classification System: Health Utilities Index Mark 2. Medical Care 1996;34(7):702-722.

Tu JV, Porter J. Stroke care in Ontario: Hospital survey results. 1999 Toronto.

Van Straten A, de Haan RJ, Limburg M, Schuling J, Bossuyt PM, van den Bos GA. A stroke-adapted 30-item version of the Sickness Impact Profile to assess quality of life (SA-SIP30). Stroke 1997;28(11):2155-2161.

Van Swieten JC, Koudstall PJ, Visser MC, Schouten HJA, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. Stroke 1988;19:604-607.

Varni JW, Katz ER, Seid M, Quiggins DJL, Friedman-Bender A, Castro CM. The Pediatric Cancer Quality of Life Inventory (PCQL). I. Instrument development, descriptive statistics and cross-informant variance. J Behav Med 1998;21(2)179-204.

Ware JE, Kemp JP, Buchner DA, Singer AE, Nolop KB, Goss TF. The responsiveness of disease-specific and generic health measures to changes in the severity of asthma among adults. Quality of Life Research 1998;7:235-244.

Ware JE, Kosinski M, Keller SD. SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales. Boston, MA: The Health Institute, New England Medical Center, 1995.

Ware JE, Kosinski M, Keller SD. SF-36 physical and mental health summary scales: a user's manual. Boston, MA: The Health Institute, New England Medical Center, 1994.

Warlow CP. Epidemiology of stroke. Lancet 1998;352(suppl 3):1-4.

Weinberger M, Oddone EZ, Samsa GP, Landsman PB. Are health-related quality of life measures affected by the mode of administration? Journal of Clinical Epidemiology 1996;49(2):135-140.

Weissman MM, Sholomskas D, Pottenger M, Prusoff BA, Locke B. Assessing depressive symptoms in five psychiatric populations: a validation study. American Journal of Epidemiology 1977;108:203-214.

Wells GA, Tugwell P, Kraag GR, Baker PRA, Groh J, Redelmeier DA. Journal of Rheumatology 1993;20:557-560.

World Health Organization. Proposal for the multinational monitoring of trends and determinants in Cardiovascular disease (MONICA Project). Rev 1. Geneva Switzerland: World Health Organization;1983: WHO/MHC/82.1

Wilkinson PR, Wolfe CDA, Warburton FG, Rudd AG, Howard RS, Ross-Russell RW, Beech RR. A long-term follow-up of stroke patients. Stroke 1997;28:507-512.

Williams LS. Health-related quality of life outcomes in stroke. Neuroepidemiology 1998;17:116-120.

Williams LS, Weinberger M, Harris LE, Clark DO, Biller J. Development of a stroke-specific quality of life scale. Stroke 1999;30(7):1362-1369.

Wu AW, Jacobson DL, Berzon RA, Revicki DA, van der HC, Fichtenbaum CJ et al. The effect of mode of administration on medical outcomes study health ratings and EuroQol scores in AIDS. Quality of Life Research 1997;6(1):3-10.

Wyrwich KW, Nienaber NA, Tierney WM, Wolinsky FD. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Medical Care 1999;37(5):469-478.

Wyrwich KW, Wolinsky FD. Identifying meaningful intra-individual change standards for health-related quality of life measures. Journal of Evaluation in Clinical Practice 2000;6(1):39-49.

Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, Leirer VO. Development and validation of a geriatric depression screening scale: a preliminary report. Journal of Psychiatric Research 1982-83;17(1):37-49.

Zung WWK. A self-rating depression scale. Archives of General Psychiatry 1965;12:63-70.

Appendix 1: Study Information Sheet and Consent Form

Patient and Caregiver Information Letter and Consent Form

Study: Health-Related Quality of Life Assessment in Stroke
Dr. Ashfaq Shuaib, MD, FRCPC, Director of Neurology, University of Alberta Hospital
Dr. Jeffrey A. Johnson, PhD; Dr. David Feeny, PhD; Simon Pickard, BscPharm,
University of Alberta

You have been asked to be in a study about quality of life and the use of health care services after stroke. This is an information letter to explain the study. Doctors and researchers at the University of Alberta are interested in how you feel about your health during the stroke recovery process

You and a family member will be asked to answer some questions in a survey about your health. It is important that you express your feelings about your health. The point of view that is important to us is your own. Your family caregiver will also complete surveys on your health and their own health. You will also be asked to perform some basic tasks when you enroll, and in 6 months from now.

We will schedule follow-up appointments with you and the same family member to fill out the health surveys at 1, 3 and 6 months of recovery after stroke. We will phone to confirm the appointment with you and your caregiver to complete the survey at home. The visit will take about 30 minutes while you complete the surveys about health and health care. The people conducting the survey are trained to answer any questions related to the study.

Please be aware that by signing this form you are providing consent to participate in this study. Your participation is voluntary and you may withdraw your consent at any time. Whether you choose to participate or not, your decision will not affect the care you are given. There are no known risks or benefits to completing surveys. However, your participation in this study will help to understand how patients and their caregivers perceive health changes after a stroke. In order to track the use of health services required during stroke recovery, such as length of hospital stay, we are asking your permission to access to such records. The information you provide will be treated confidentially. All information will be anonymous. Following University policy, we will keep the data for 7 years. Any reports will contain only summary information on the participants who enroll in this study.

We hope to learn how stroke impacts the health of patients and their caregivers. Your participation in this project would be greatly appreciated. If you have any questions, please call Dr. Shuaib at 407-6395, or Simon Pickard at 448-4881. To speak with someone who is not directly involved in this study, please call the CHA Patient Concerns Office at 407-1040.

This study was explained to me by: ______________________________

I agree to take part in this study.

Signature of Participant (Patient) Printed Name Date

Signature of Participant (Caregiver) Printed Name Date

Witness Printed Name Date

I believe that the person signing this form understands what is involved in the study and voluntarily agrees to participate.

Signature of Investigator or Designee Date

(COPY TO THE STUDY PARTICIPANT)

Appendix 2: Pre-Questionnaire Forms

A2.1 Reasons for Non-Participation in Study

(For completion by Clinical Assessor)
Please complete this form for every patient/caregiver pair approached during the recruitment process. Due to limited information on medical chart when patient data is reviewed for adherence to inclusion and exclusion criteria, patient and/or caregiver may not participate for the following reason(s):

Please check the item(s) that best describe reason(s) for non-participation

1. ______ Exclusion criteria: life expectancy of < 6 months for any medical reason.
2. ______ Exclusion criteria: previous degenerative or space occupying brain disorder.
3. ______ Exclusion criteria: hemorrhagic stroke
4. ______ Exclusion criteria: subarachnoid hemorrhage or transient ischemic attack.
5. ______ Exclusion criteria: coma, or with global or Wernicke's aphasia.
6. ______ Exclusion criteria: history of dementia prior to stroke.
7. ______ Exclusion criteria: patient cognitively impaired
8. ______ Exclusion criteria: patient lives too far away (>150kms) from Edmonton
9. ______ patient not interested or afraid of participating in research
10. ______ patient unwilling to participate due to perceived strain of respondent burden
11. ______ patient does not understand English (unable to respond to survey)
12. ______ patient cannot read and won't complete survey with assistance
13. ______ too much time has elapsed since stroke (> 2 weeks)
14. ______ no caregiver
15. ______ caregiver does not want patient to participate
16. ______ caregiver cognitively impaired to the extent that they cannot participate
17. ______ caregiver unwilling to participate due to strain of respondent burden
18. ______ caregiver not interested or afraid of participating in research
19. ______ caregiver from out of province or lives too far away from Edmonton
20. ______ caregiver does not understand English (unable to respond to survey)
21. ______ caregiver cannot read and won't complete survey with assistance
22. _____ Other reason. Specify: ______________________________

A2.2 Patient Demographic Data Form

To be completed by recruiter upon initial recruitment of patient and caregiver into study
*Extract from patient medical chart

Patient ID: **Hospital ID:** ______________/______

Name of Patient: ______________________ Initials:
Street Address: ____
City, Province ________________________
Postal Code: ____________________
Telephone Number (____)____________
Patient Date of Birth (DD/MM/YY): (__/__/__) ___________ years old
Sex of patient: 1 Male 2 Female
Date of stroke (DD/MM/YY): (__/__/__)
History of stroke: 1 Yes, previous stroke 2 No
Type of stroke (ICD code):___________________________
Bamford classification of stroke : 1 TACI (total anterior circulation infarct)
2 PACI (partial anterior circulation infarct)
3 POCI (posterior circulation infarct)
4 LACI (lacunar infarct)

Name of Caregiver: _________________________
Street Address: ____
City, Province ________________________
Postal Code: ____________________
Telephone Number (____)____________
Emergency Contact Phone No. (___)____________
Relationship of Caregiver to Patient: 1 Spouse
2 Daughter
3 Son
4 Sister
5 Brother
6 Other: Specify:_________________

Caregiver Date of Birth (DD/MM/YY): (__/__/__) ___________ years old
Sex of caregiver: 1 Male 2 Female

A2.3 Questionnaire Cover Sheet

Patient ID: Respondent* Wave (0, 1, 3, 6)

*1= patient self-report, 2 = proxy by caregiver, 3 = caregiver self assessment

Initials of Subject:

Was the questionnaire completed? 1 Yes 2 No

Date of Visit A: _____ _____ _____
dd mm yy

If questionnaire was not completed on Visit A, Specify the reason(s) why the questionnaire was not completed (tick those that apply). If questionnaire was completed, tick item 7 (not applicable)

_____ 1. Patient/caregiver kept appointment for examination, but could not complete questionnaire due to illness

_____ 2. Patient/caregiver kept appointment for examination, but refused to complete questionnaire for reason other than illness. Specify reason: ________________

_____ 3. Patient/caregiver did not keep appointment. Specify reason: ________________

_____ 4. Patient/caregiver could not be contacted.

_____ 5. Questionnaire not administered due to institutional error. Explain: ____________

_____ 6. Other reason, specify: ________________

_____ 7. Completed; Not applicable

If questionnaire completed on second visit, list date: _____ _____ _____
dd mm yy

Start Time: ____________ a.m./p.m.
End Time: ____________ a.m./p.m.
Total Length of Time to Complete: ________________ minutes
Were *all* questions answered? 1 ▶ Yes 2 ▶ No If no, give reason__________
Was assistance required? 1 ▶ Yes 2 ▶ No If yes, give reason _________
Where was the questionnaire completed? 1 ▶ home 2 ▶ hospital 3 ▶ another centre

Interviewer Number: 1 ▶ Nasser 2 ▶ RA I (AH) 3 ▶ RA II (AW) 4 ▶ RAIII (MG) 5 ▶ RA IV(SP)

Level of assistance from interviewer: Code: ___________ (from next page; coded from 001 to 004)

Attrition (if patient drops out of study): 1 ▶ deceased 2 ▶ voluntary 3 ▶ invol untary

A2.4 Checklist of Codified Responses to Respondent Questions about Survey

Codes used when assistance is provided to respondent during completion of survey

Code:

0 Could not read due to visual impairment

1 Did not understand a particular word or phase of item: # _____

2 Required explanation of what was meant by item

3 Required physical assistance to fill out survey

If not listed please describe type of assistance required: ____

____

____

____

Appendix 3: Instruments Completed by Clinical Assessor

A3.1 National Institutes of Health Stroke Scale (NIHSS)

1.a. Level of Consciousness:	0 Alert
	1 Not alert, but arousable with minimal stimulation
	2 Not alert, requires repeated stimulation to attend
	3 Coma
1.b. Ask patient the month and their age:	0 Answers both correctly
	1 Answers one correctly
	2 Both incorrect
1.c. Ask patient to open and close eyes and	0 Obeys both correctly
	1 Obeys one correctly
	2 Both incorrect
2. Best gaze (only horizontal eye movement):	0 Normal
	1 Partial gaze palsy
	2 Forced deviation
3. Visual Field testing:	0 No visual field loss
	1 Partial hemianopia
	2 Complete hemianopia
	3 Bilateral hemianopia (blind including cortical blindness)
4. Facial Paresis (Ask patient to show teeth or raise eyebrows and close eyes tightly):	0 Normal symmetrical movement
	1 Minor paralysis (flattened nasolabial fold, asymmetry on smiling)
	2 Partial paralysis (total or near total paralysis of lower face)
	3 Complete paralysis of one or both sides (absence of facial movement in the upper and lower face)
5. Motor Function - Arm (right and left): Right arm ____ Left arm ____	0 Normal (extends arms 90 (or 45) degrees for 10 seconds without drift)
	1 Drift
	2 Some effort against gravity
	3 No effort against gravity
	4 No movement
	9 Untestable (Joint fused or limb amputated)
6. Motor Function - Leg (right and left): Right leg ____	0 Normal (hold leg 30 degrees position for 5 seconds)
	1 Drift
	2 Some effort against gravity

Left leg ____	3 No effort against gravity
	4 No movement
	9 Untestable (Joint fused or limb amputated)
7. Limb Ataxia:	0 No ataxia
	1 Present in one limb
	2 Present in two limbs
8. Sensory (Use pinprick to test arms, legs, trunk and face -- compare side to side):	0 Normal
	1 Mild to moderate decrease in sensation
	2 Severe to total sensory loss
9. Best Language (describe picture, name items, read sentences)	0 No aphasia
	1 Mild to moderate aphasia
	2 Severe aphasia
	3 Mute
10. Dysarthria (read several words):	0 Normal articulation
	1 Mild to moderate slurring of words
	2 Near unintelligible or unable to speak
	9 Intubated or other physical barrier
11. Extinction and inattention:	0 Normal
	1 Inattention or extinction to bilateral simultaneous stimulation in one of the sensory modalities
	2 Severe hemi-inattention or hemi-inattention to more than one modality

A3.2 The Scandinavian Stroke Scale (SSS-48)

Item	Score	Prognostic Score	Long-term Score
Consciousness			
Fully conscious	6		
Somnolent, can be awakened to full consciousness	4	▶	
Reacts to verbal command, but is not fully conscious	2		
Eye Movements			
No gaze palsy	4		
Gaze palsy present	2	▶	
Conjugate eye deviation	0		
Arm, motor power*			
Raises arm with normal strength	6		
Raises arm with reduced strength	5		
Raises arm with flexion in elbow	4	▶	▶
Can move, but not against gravity	2		
Paralysis	0		
HAND, MOTOR POWER*			
Normal strength	6		
Reduced strength in full range	4		▶
Some movement, fingertips do not reach palm	2		
Paralysis	0		
Leg, motor power			
Normal strength	6		
Raises straight leg with reduced strength	5		
Raises leg with flexion of knee	4	▶	▶
Can move, but not against gravity	2		
Paralysis	0		
ORIENTATION			
Correct for time, place and person	6		
2 of these	4		▶
1 of these	2		
Completely disorientated	0		
SPEECH			
No aphasia	10		
Limited vocabulary or incoherent speech	6		▶
More than yes/no, but no longer sentences	3		
Only yes/no or less	0		
Facial palsy			
None/dubious	2		▶
Present	0		
GAIT			
Walks 5 m without aids	12		
Walks without aids	9		
Walks with help of another person	6		▶
Sits without support	3		
Bedridden/wheelchair	0		
Maximum Score		22	48

*motor power is assessed only on the affected side

A3.3 Barthel Index (15-item version)

Item	**Can do by myself**	**Can do with help of someone else**	**Cannot do at all**
Drinking and eating	10	3	3
Dressing upper body	5	3	0
Dressing lower body	5	2	0
Donning brace or prosthesis	0	-2	N/A
Grooming	5	0	0
Washing or bathing	4	0	0
Perineal care	4	0	0
Managing urination	10	5	0
Managing bowel movements	10	5	0
Getting in and out of a chair	15	7	0
Getting on /off a toilet	6	3	0
Getting in and out of a tub or Shower	1	0	0
Walking 50 m on the level	15	10	0
Going up /down one flight of Stairs	10	5	0
IF NOT WALKING propelling Or pushing a wheelchair 50 m	5	0	N/A
Barthel Total: Best score is 100, worst score is 3.		/100	

A3.4 The Modified Rankin Handicap Scale Grades

0 = No symptoms at all
1 = No significant disability despite symptoms; able to carry out all usual duties and activities
2 = Slight disability: unable to carry out all previous activities but able to look after own affairs without assistance
3 = Moderate disability: requiring some help, but able to walk without assistance
4 = Moderately severe disability: unable to walk without assistance, and unable to attend to own bodily needs without assistance
5 = Severe disability: bedridden, incontinent, and requires constant nursing care and attention

A3.5 Clinical Assessment of Patient Global Health Change (at 6 months)

Has there been any change in the subject's health since baseline?

(check one box)

Worse	A very great deal worse	-7
	A great deal worse	-6
	A good deal worse	-5
	Moderately worse	-4
	Somewhat worse	-3
	A little worse	-2
	Almost the same, hardly any worse at all	-1
	No change	0
Better	Almost the same, hardly any better at all	1
	A little better	2
	Somewhat better	3
	Moderately better	4
	A good deal better	5
	A great deal better	6
	A very great deal better	7

Appendix 4: Patient Questionnaire (6 month version)

A4.1 Health Utilities Index Mark 2 and Mark 3 (HUI23S1.15Q)

Health status classification system: HUI2

Attribute	Level	Description
SENSORY	1	Able to see, hear and speak normally for age
	2	Requires equipment to see or hear or speak
	3	Sees, hears, or speaks with limitations even with equipment
	4	Blind, deaf or mute
MOBILITY	1	Able to walk, bend, lift, jump and run normally for age
	2	Walks, bends, lifts, jumps or runs with some limitations but does not require help
	3	Requires mechanical equipment (such as canes, crutches, braces or wheelchair) to walk or get around independently
	4	Requires the help of another person to walk or get around and requires mechanical equipment as well
	5	Unable to control or use arms and legs
EMOTION	1	Generally happy and free from worry
	2	Occasionally fretful, angry, irritable, anxious, depressed, or suffering "nigh terrors"
	3	Often fretful, angry, irritable, anxious, depress or suffering "night terrors"
	4	Almost always fretful, angry, irritable, anxious, depressed
	5	Extremely fretful, angry, irritable or depressed usually requiring hospitalization or psychiatric institutional care
COGNITIVE	1	Learns and remembers school work normally for age
	2	Learns and remembers school work more slowly than classmates as judged by parents and/or teachers
	3	Learns and remembers very slowly and usually requires special educational assistance
	4	Unable to learn and remember
SELF-CARE	1	Eats, bathes, dresses and uses the toilet normally for age
	2	Eats, bathes, dresses or uses the toilet independently with difficulty

	3	Requires mechanical equipment to eat, bathe, dress or use the toilet independently
	4	Requires the help of another person to eat, bathe, dress or use the toilet
PAIN	1	Free of pain and discomfort
	2	Occasional pain. Discomfort relieved by non-prescription drugs or self-control activity without disruption of normal activities
	3	Frequent pain. Discomfort relieved by oral medicines with occasional disruption of normal activities
	4	Frequent pain; frequent disruption of normal activities. Discomfort requires prescription narcotics for relief
	5	Severe pain. Pain not relieved by drugs and constantly disrupts normal activities
FERTILITY	1	Able to have children with a fertile spouse
	2	Difficulty in having children with a fertile spouse
	3	Unable to have children with a fertile spouse

Health status classification system: HUI3

VISION

1 Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, without glasses or contact lenses.

2 Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, but with glasses.

3 Able to read ordinary newsprint with or without glasses but unable to recognize a friend on the other side of the street, even with glasses.

4 Able to recognize a friend on the other side of the street with or without glasses but unable to read ordinary newsprint, even with glasses.

5 Unable to read ordinary newsprint and unable to recognize a friend on the other side of the street, even with glasses.

6 Unable to see at all.

HEARING

1 Able to hear what is said in a group with at least three other people, without a hearing aid.

2 Able to hear what is said in a conversation with one other person in a quiet room without a hearing aid, but requires a hearing aid to hear what is said in a group conversation with at least three other people.

3 Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, and able to hear what is said in a group conversation with at least three other people, with a hearing aid.

4 Able to hear what is said in a conversation with one other person in a quiet room, without a hearing aid, but unable to hearing what is said in a group conversation with at least three other people even with a hearing aid.

5 Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid.

6 Unable to hear at all.

SPEECH

1 Able to be understood completely when speaking with strangers or friends.

2 Able to be understood partially when speaking with strangers but able to be understood completely when speaking with people who know me well.

3 Able to be understood partially when speaking with strangers or people who know me well.

4 Unable to be understood when speaking with strangers but able to be understood partially by people who know me well.

5 Unable to be understood when speaking to other people (or unable to speak at all).

AMBULATION 1 Able to walk around the neighbourhood without difficulty, and without walking equipment.

2 Able to walk around the neighbourhood with difficulty; but does not require walking equipment or the help of another person.

3 Able to walk around the neighbourhood with walking equipment, but without the help of another person.

4 Able to walk only short distances with walking equipment, and requires a wheelchair to get around the neighbourhood.

5 Unable to walk alone, even with walking equipment; able to walk short distances with the help of another person, and requires a wheelchair to get around the neighbourhood.

6 Cannot walk at all.

DEXTERITY 1 Full use of two hands and ten fingers.

2 Limitations in the use of hands or fingers, but does not require special tools or help of another person.

3 Limitations in the use of hands or fingers, is independent with use of special tools (does not require the help of another person).

4 Limitations in the use of hands or fingers, requires the help of another person for some tasks (not independent even with use of special tools).

5 Limitations in use of hands or fingers, requires the help of another person for most tasks (not independent even with use of special tools).

6 Limitations in use of hands or fingers, requires the help of another person for all tasks (not independent even with use of special tools).

EMOTION

1 Happy and interested in life.

2 Somewhat happy.

3 Somewhat unhappy.

4 Very unhappy.

5 So unhappy that life is not worthwhile.

COGNITION

1 Able to remember most things, think clearly and solve day to day problems.

2 Able to remember most things, but have a little difficulty when trying to think and solve day to day problems.

3 Somewhat forgetful, but able to think clearly and solve day to day problems.

4 Somewhat forgetful, and have a little difficulty when trying to think or solve day to day problems.

5 Very forgetful, and have great difficulty when trying to think or solve day to day problems.

6 Unable to remember anything at all, and unable to think or solve day to day problems.

PAIN

1 Free of pain and discomfort.

2 Mild to moderate pain that prevents no activities

3 Moderate pain that prevents a few activities.

4 Moderate to severe pain that prevents some activities.

5 Severe pain that prevents most activities.

For the standard version of the HUI Mark 2 and 3 self-administered, self-assessed "one-week" health status assessment contact:

Health Utilities Inc.,
Dundas ON, Canada
L9H 2V3,
phone (905)525-9140
url: <www.healthutilities.com>

A4.2 EQ-5D

By placing a check-mark in one box in each group below, please indicate which statements best describe your own state of health today.

Mobility

I have no problems in walking about □

I have some problems in walking about □

I am confined to bed □

Self-Care

I have no problems with self-care □

I have some problems washing or dressing myself □

I am unable to wash or dress myself □

Usual Activities (*e.g. work, study, housework, family or leisure activities)*

I have no problems with performing my usual activities □

I have some problems with performing my usual activities □

I am unable to perform my usual activities □

Pain/Discomfort

I have no pain or discomfort □

I have moderate pain or discomfort □

I have extreme pain or discomfort □

Anxiety/Depression

I am not anxious or depressed □

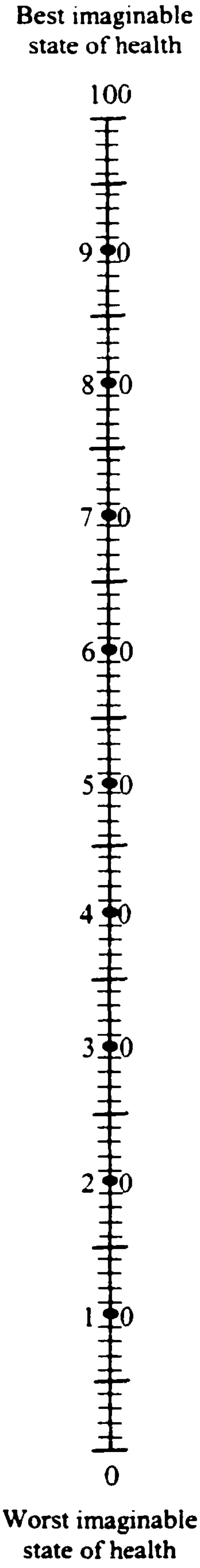
I am moderately anxious or depressed □

I am extremely anxious or depressed □

To help people say how good or bad their state of health is, we have drawn a scale (rather like a thermometer) on which the best state you can imagine is marked 100 and the worst state you can imagine is marked 0.

We would like you to indicate on this scale how good or bad your own health is today, in your opinion. Please do this by drawing a line from the box below to whichever point on the scale indicates how good or bad your state of health is today.

Your own state of health today


Best imaginable state of health
100
90
80
70
60
50
40
30
20
10
0
Worst imaginable state of health

A4.3 CES-D Scale

Circle the number for each statement which best describes how often you felt or behaved this way -- **DURING THE PAST WEEK:**	Rarely or None of the Time (Less than 1 Day)	Some or a Little of the Time (1-2 Days)	Occasionally or a Moderate Amount of time (3-4 Days)	Most or All of the time (5-7 Days)
1. I was bothered by things that usually don't bother me	0	1	2	3
2. I did not feel like eating; my appetite was poor	0	1	2	3
3. I felt that I could not shake off the blues even with help from my family or friends	0	1	2	3
4. I felt that I was just as good as other people	0	1	2	3
5. I had trouble keeping my mind on what I was doing	0	1	2	3
6. I felt depressed	0	1	2	3
7. I felt that everything I did was an effort	0	1	2	3
8. I felt hopeful about the future	0	1	2	3
9. I thought my life had been a failure	0	1	2	3
10. I felt fearful	0	1	2	3
11. My sleep was restless	0	1	2	3
12. I was happy	0	1	2	3
13. I talked less than usual	0	1	2	3
14. I felt lonely	0	1	2	3
15. People were unfriendly	0	1	2	3
16. I enjoyed life	0	1	2	3
17. I had crying spells	0	1	2	3
18. I felt sad	0	1	2	3
19. I felt that people disliked me	0	1	2	3
20. I could not get "going"	0	1	2	3

A4.4 SF-36

1. In general, would you say your health is:

	(circle one)
Excellent	1
Very good	2
Good	3
Fair	4
Poor	5

2. Compared to one year ago, how would you rate your health in general now?

	(circle one)
Much better now than one year ago	1
Somewhat better now than one year ago	2
About the same as one year ago	3
Somewhat worse now than one year ago	4
Much worse now than one year ago	5

3. The following items are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

(circle one number on each line)

ACTIVITIES	**Yes, Limited A Lot**	**Yes, Limited A Little**	**No, Not Limited At All**
a. **Vigorous activities**, such as running, lifting heavy objects, participating in strenuous sports	1	2	3
b. **Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	1	2	3
c. Lifting or carrying groceries	1	2	3
d. Climbing **several** flights of stairs	1	2	3
e. Climbing **one** flight of stairs	1	2	3
f. Bending, kneeling, or stooping	1	2	3
g. Walking **more than a kilometre**	1	2	3
h. Walking **several blocks**	1	2	3
i. Walking **one block**	1	2	3
j. Bathing or dressing yourself	1	2	3

4. During the past week, have you had any of the following problems with your work or other regular daily activities as a result of your physical health?

(circle one number on each line)

	YES	**NO**
a. Cut down on the **amount of time** you spent on work or other activities	1	2
b. **Accomplished less** than you would like	1	2
c. Were limited in the kind of work or other activities	1	2
d. Had **difficulty** performing the work or other activities (for example, it took extra effort)	1	2

5. During the past week, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?

(circle one number on each line)

	YES	NO
a. Cut down the amount of time you spent on work or other activities	1	2
b. Accomplished less than you would like	1	2
c. Didn't do work or other activities as carefully as usual	1	2

6. During the past week, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?

(circle one)

Not at all	1
Slightly	2
Moderately	3
Quite a bit	4
Extremely	5

7. How much bodily pain have you had during the past week?

(circle one)

None	1
Very mild	2
Mild	3
Moderate	4
Severe	5
Very severe	6

8. During the past week, how much did pain interfere with your normal work (including both work outside the home and housework)?

(circle one)

Not at all	1
A little bit	2
Moderately	3
Quite a bit	4
Extremely	5

9. These questions are about how you feel and how things have been with you during the past week. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past week –

(circle one number on each line)

	All of the Time	Most of the Time	A Good Bit of the Time	Some of the Time	A Little of the Time	None of the Time
a. Did you feel full of pep?	1	2	3	4	5	6
b. Have you been a very nervous person?	1	2	3	4	5	6
c. Have you felt so down in the dumps that nothing could cheer you up?	1	2	3	4	5	6
d. Have you felt calm and peaceful?	1	2	3	4	5	6
e. Did you have a lot of energy?	1	2	3	4	5	6
f. Have you felt downhearted and blue?	1	2	3	4	5	6
g. Did you feel worn out?	1	2	3	4	5	6
h. Have you been a happy person?	1	2	3	4	5	6
i. Did you feel tired?	1	2	3	4	5	6

10. During the past week, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?

	(circle one)
All the time	1
Most of the time	2
Some of the time	3
A little of the time	4
None of the time	5

11. How TRUE or FALSE is each of the following statements for you?

(circle one)

	Definitely True	Mostly True	Don't Know	Mostly False	Definitely False
a. I seem to get sick a little easier than other people	1	2	3	4	5
b. I am as healthy as anybody I know	1	2	3	4	5
c. I expect my health to get worse	1	2	3	4	5
d. My health is excellent	1	2	3	4	5

12. Has there been any change in your health since the last survey? (check one box)

Worse	A very great deal worse	-7
	A great deal worse	-6
	A good deal worse	-5
	Moderately worse	-4
	Somewhat worse	-3
	A little worse	-2
	Almost the same, hardly any worse at all	-1
	No change	0
Better	Almost the same, hardly any better at all	1
	A little better	2
	Somewhat better	3
	Moderately better	4
	A good deal better	5
	A great deal better	6
	A very great deal better	7

Appendix 5: Proxy Questionnaire (administered at 6 months)

A5.1: Health Utilities Index Mark 2 and Mark 3: (HUI23P1.15Q)

For the standard version of the HUI Mark 2 and 3 self-administered, proxy-assessed "one-week" health status assessment contact:

Health Utilities Inc.,
Dundas ON, Canada
L9H 2V3,
phone (905)525-9140
url: <www.healthutilities.com>

A5.2 EQ-5D (proxy version)

By placing a check-mark in one box in each group below, please indicate which statements best describe the subject's state of health today (in your opinion).

Mobility

No problems in walking about □

Some problems in walking about □

Confined to bed □

Self-Care

No problems with self-care □

Some problems washing or dressing myself □

Unable to wash or dress myself □

Usual Activities *(e.g. work, study, housework, family or leisure activities)*

No problems with performing usual activities □

Some problems with performing usual activities □

Unable to perform usual activities □

Pain/Discomfort

No pain or discomfort □

Moderate pain or discomfort □

Extreme pain or discomfort □

Anxiety/Depression

Not anxious or depressed □

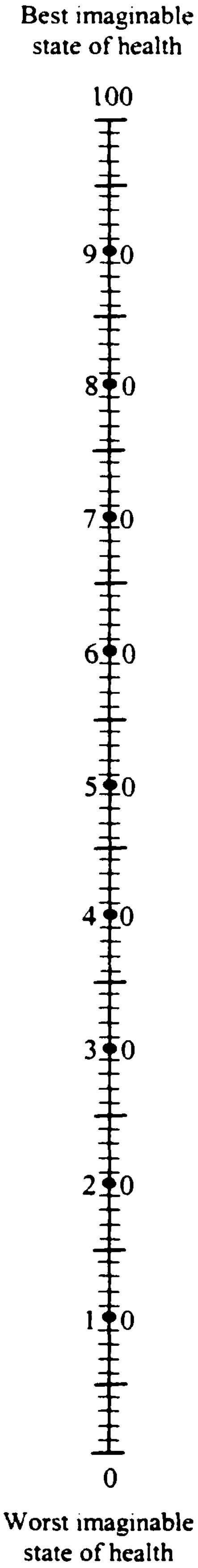
Moderately anxious or depressed □

Extremely anxious or depressed □

To help people say how good or bad their state of health is, we have drawn a scale (rather like a thermometer) on which the best state you can imagine is marked 100 and the worst state you can imagine is marked 0.

We would like you to indicate on this scale how good or bad the subject's health is today, in your opinion. Please do this by drawing a line from the box below to whichever point on the scale indicates how good or bad the subject's health is today.

Subject's
state of health
today



A5.3 CES-D Scale (adapted proxy version)

Circle the number for each statement which best describes how often the subject felt or behaved this way -- **DURING THE PAST WEEK:**	Rarely or None of the Time (Less than 1 Day)	Some or a Little of the Time (1-2 Days)	Occasionally or a Moderate Amount of time (3-4 Days)	Most or All of the time (5-7 Days)
1. Bothered by things that usually don't bother them	0	1	2	3
2. Did not feel like eating; appetite was poor	0	1	2	3
3. Could not shake off the blues even with help from family or friends	0	1	2	3
4. Felt that they were just as good as other people	0	1	2	3
5. Had trouble keeping their mind on what they were doing	0	1	2	3
6. Felt depressed	0	1	2	3
7. Felt everything was an effort	0	1	2	3
8. Felt hopeful about the future	0	1	2	3
9. Thought their life had been a failure	0	1	2	3
10. Felt fearful	0	1	2	3
11. Had a restless sleep	0	1	2	3
12. Felt happy	0	1	2	3
13. Talked less than usual	0	1	2	3
14. Felt lonely	0	1	2	3
15. People were unfriendly	0	1	2	3
16. Enjoyed life	0	1	2	3
17. Had crying spells	0	1	2	3
18. Felt sad	0	1	2	3
19. Felt that people disliked them	0	1	2	3
20. Could not get "going"	0	1	2	3

A5.4 SF-36 (Proxy Adapted Version)

For the following questions, please circle the number that best describes the subject's health. Please complete all the questions. We apologize for some of the items that ask questions similar to those you have already answered.

1. In general, would you say the subject's health is:

	(circle one)
Excellent	1
Very good	2
Good	3
Fair	4
Poor	5

2. Compared to one year ago, how would you rate the subject's health in general now?

	(circle one)
Much better now than one year ago	1
Somewhat better now than one year ago	2
About the same as one year ago	3
Somewhat worse now than one year ago	4
Much worse now than one year ago	5

3. The following items are about activities the subject might do during a typical day. Does the subject's health now limit them in these activities? If so, how much?

(circle one number on each line)

ACTIVITIES	**Yes, Limited A Lot**	**Yes, Limited A Little**	**No, Not Limited At All**
a. **Vigorous activities**, such as running, lifting heavy objects, participating in strenuous sports	1	2	3
b. **Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	1	2	3
k. Lifting or carrying groceries	1	2	3
l. Climbing **several** flights of stairs	1	2	3
m. Climbing **one** flight of stairs	1	2	3
n. Bending, kneeling, or stooping	1	2	3
o. Walking **more than a kilometre**	1	2	3
p. Walking **several blocks**	1	2	3
q. Walking **one block**	1	2	3
r. Bathing or dressing themselves	1	2	3

4. During the past week, has the subject had any of the following problems with their work or other regular daily activities as a result of their physical health?

(circle one number on each line)

	YES	**NO**
Cut down on the **amount of time** they spent on work or other activities	1	2
d. **Accomplished less** than they would like	1	2
e. Were limited in the kind of work or other activities	1	2
d. Had **difficulty** performing the work or other activities (for e.xample, it took extra effort)	1	2

5. During the past week, did the subject have any of the following problems with work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?

(circle one number on each line)

	YES	NO
d. Cut down the amount of time they spent on work or other activities	1	2
e. Accomplished less than they would like	1	2
f. Didn't do work or other activities as carefully as usual	1	2

6. During the past week, to what extent has the subject's physical health or emotional problems interfered with their normal social activities with family, friends, neighbors, or groups?

(circle one)

Not at all	1
Slightly	2
Moderately	3
Quite a bit	4
Extremely	5

7. How much bodily pain has the subject had during the past week?

(circle one)

None	1
Very mild	2
Mild	3
Moderate	4
Severe	5
Very severe	6

10. During the past week, how much did pain interfere with the subject's normal work (including both work outside the home and housework)?

(circle one)

Not at all	1
A little bit	2
Moderately	3
Quite a bit	4
Extremely	5

11. These questions are about how the subject feels and how things have been with the subject during the past week. For each question, please give the one answer that comes closest to the way they have been feeling. How much of the time during the past week –

(circle one number on each line)

	All of the Time	Most of the Time	A Good Bit of the Time	Some of the Time	A Little of the Time	None of the Time
b. Did they feel full of pep?	1	2	3	4	5	6
b. Have they been a very nervous person?	1	2	3	4	5	6
c. Have they felt so down in the dumps that nothing could cheer them up?	1	2	3	4	5	6
f. Have they felt calm and peaceful?	1	2	3	4	5	6
g. Did they have a lot of energy?	1	2	3	4	5	6
f. Have they felt downhearted and blue?	1	2	3	4	5	6
j. Did they feel worn out?	1	2	3	4	5	6
k. Have they been a happy person?	1	2	3	4	5	6
l. Did they feel tired?	1	2	3	4	5	6

10. During the past week, how much of the time has the subject's physical health or emotional problems interfered with their social activities (like visiting with friends, relatives, etc.)?

(circle one)

All the time	1
Most of the time	2
Some of the time	3
A little of the time	4
None of the time	5

11. How TRUE or FALSE is each of the following statements for the subject?

(circle one)

	Definitely True	Mostly True	Don't Know	Mostly False	Definitely False
a. They seem to get sick a little easier than other people	1	2	3	4	5
b. They are as healthy as anybody you know	1	2	3	4	5
c. They expect their health to get worse	1	2	3	4	5
e. Their health is excellent	1	2	3	4	5

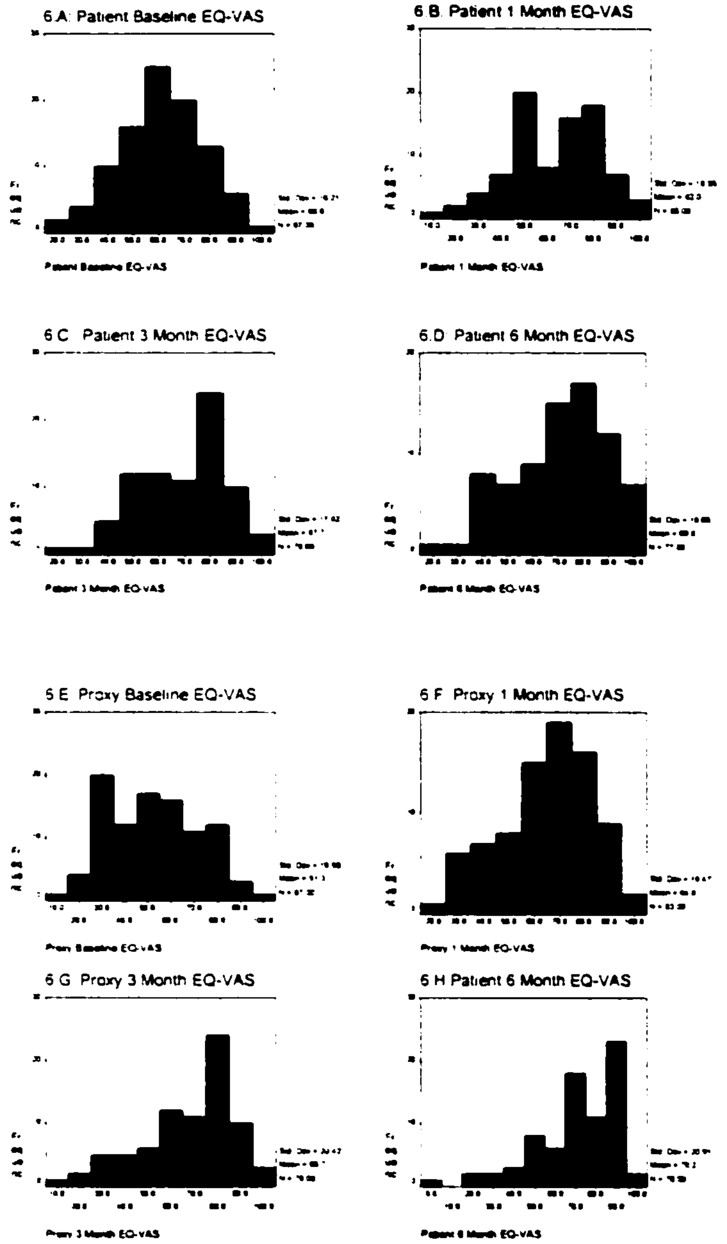
12. Has there been any change in the subject's health since the last survey?

(check one box)

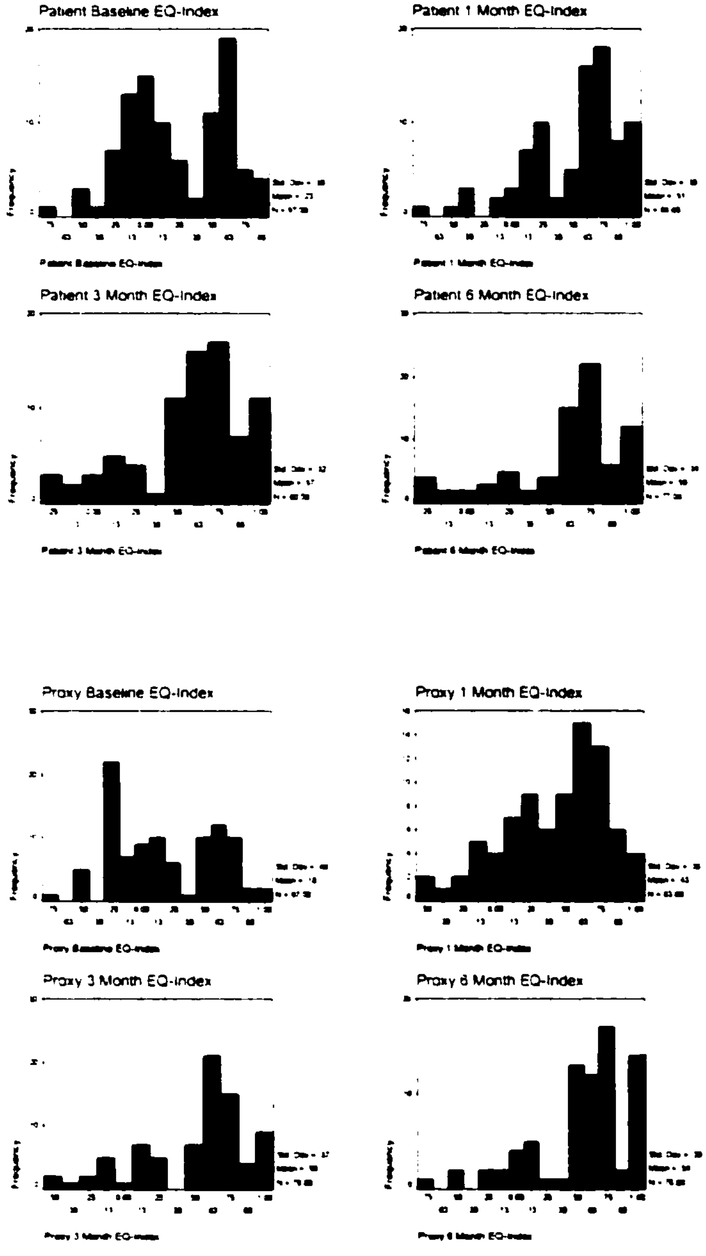
Worse	A very great deal worse	-7
	A great deal worse	-6
	A good deal worse	-5
	Moderately worse	-4
	Somewhat worse	-3
	A little worse	-2
	Almost the same, hardly any worse at all	-1
	No change	0
	Almost the same, hardly any better at all	1
Better	A little better	2
	Somewhat better	3
	Moderately better	4
	A good deal better	5
	A great deal better	6
	A very great deal better	7

Appendix 6: Frequency Distributions of Summary Scores

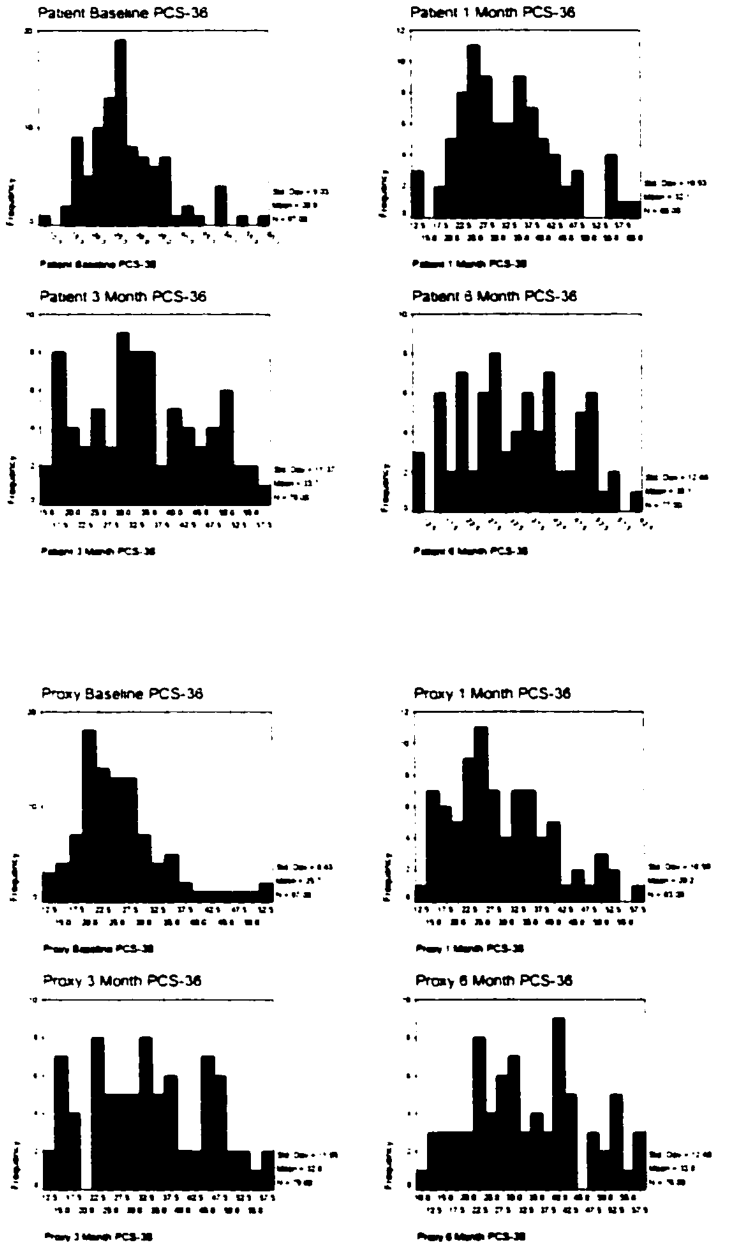
Figures 6 A to 6 H: EQ-VAS Scores



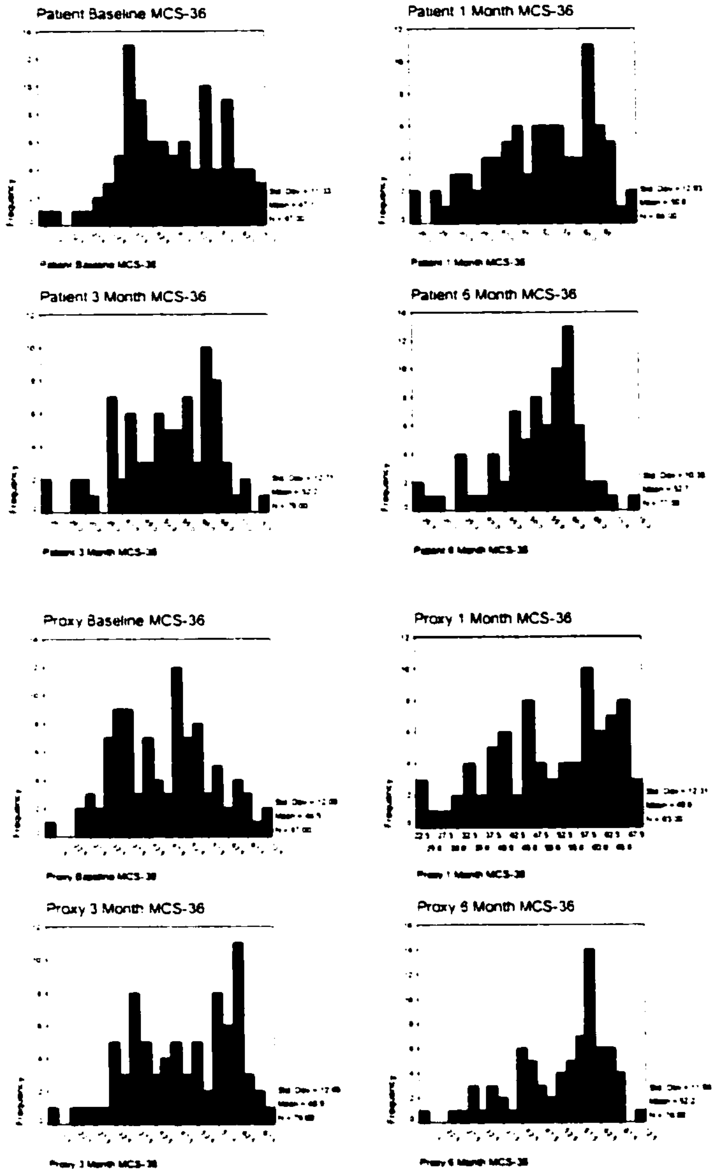
Figures 6 I to 6 P: EQ-5D Index-based Scores



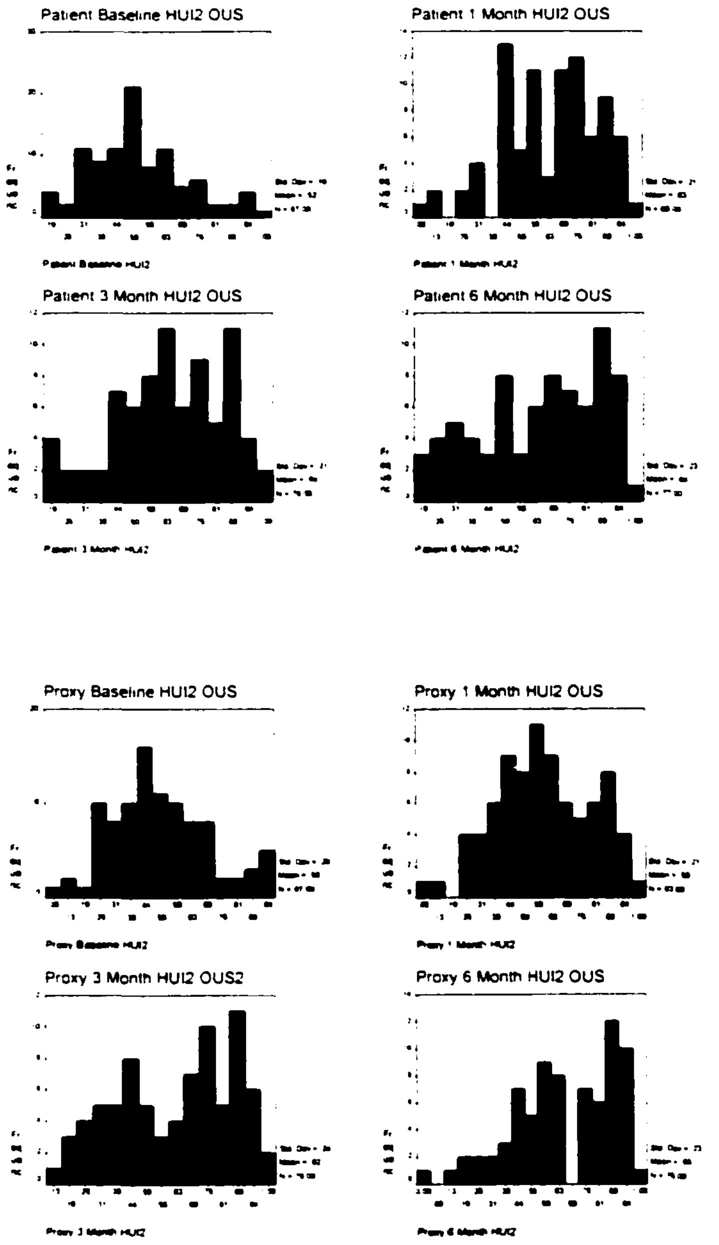
Figures 6 Q to 6 X: PCS-36 Scores



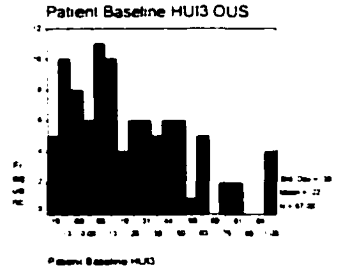
Figures 6 Y to 6 FF: MCS-36 Scores

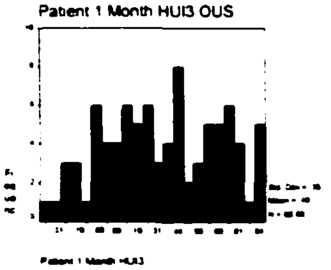


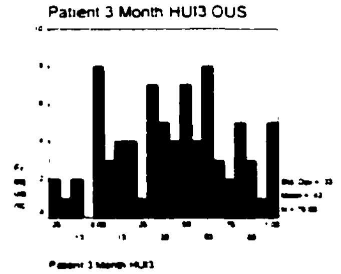
Figures 6 GG to 6 NN: HUI2 OUS

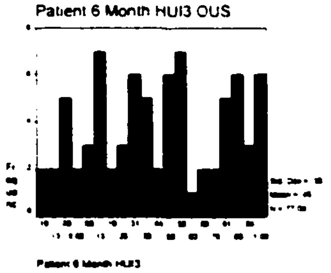


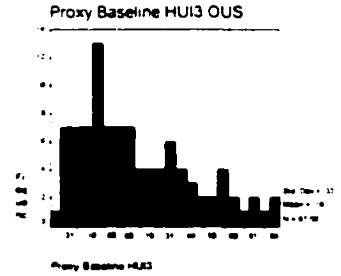
Figures 6 00 to 6 VV: HUI3 OUS

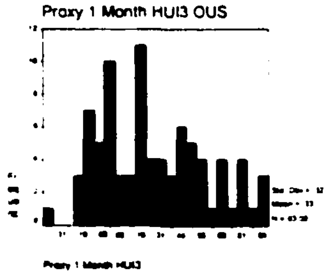


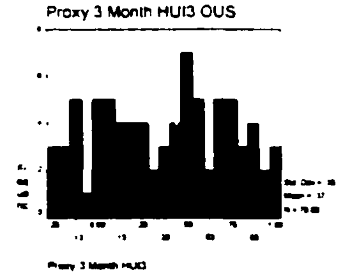


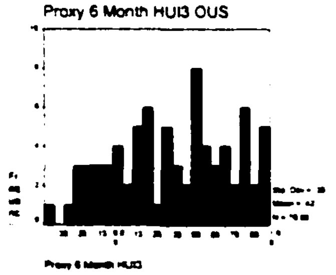












Appendix 7: Figures of Mean Summary Scores over Time

Figure 7 A: Patient EQ-VAS vs Time

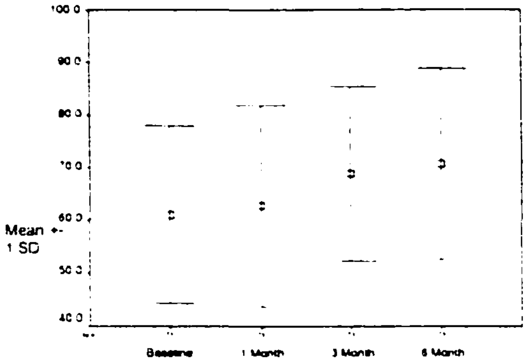


Figure 7 B: Proxy EQ-VAS vs Time

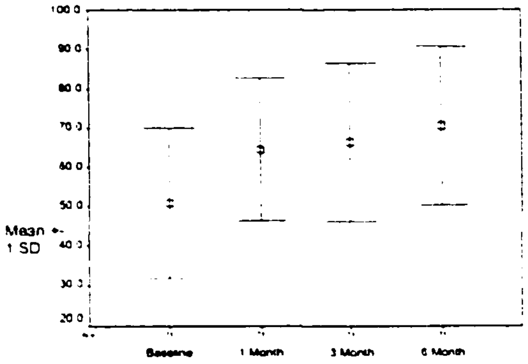


Figure 7 C: Patient EQ-Index vs Time

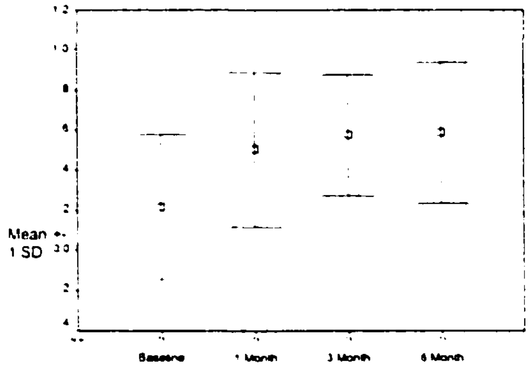


Fig 7 D: Proxy EQ-Index Scores vs Time

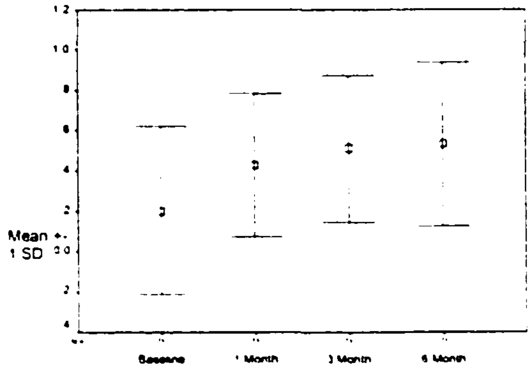


Figure 7 E: Patient PCS-36 vs Time

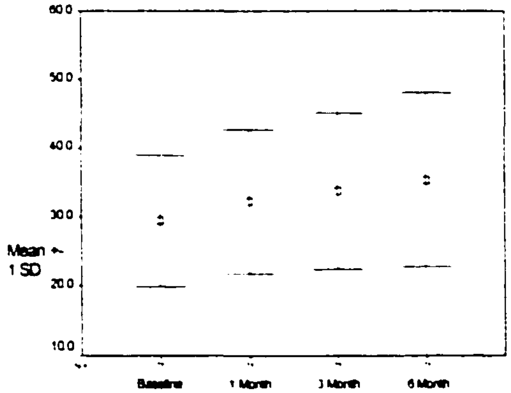


Figure 7 F: Proxy PCS-36 vs Time

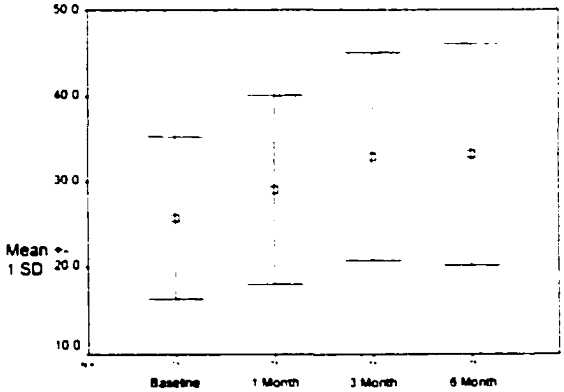


Figure 7 G: Patient MCS-36 vs Time

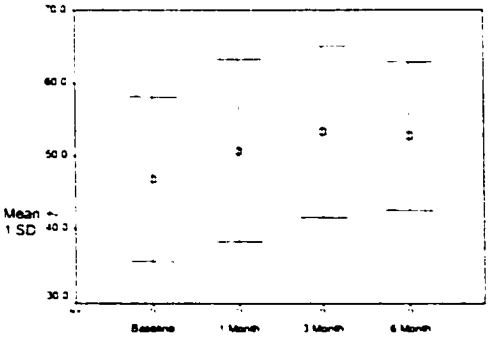


Figure 7 H: Proxy MCS-36 vs Time

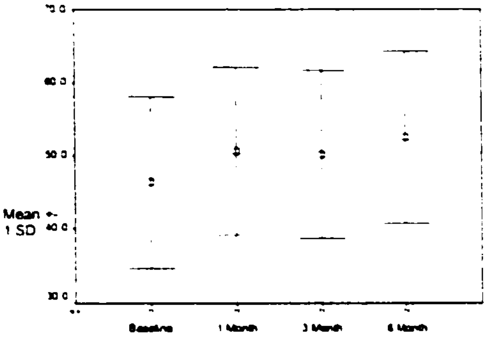


Figure 7 I: Patient HUI2 OUS vs Time

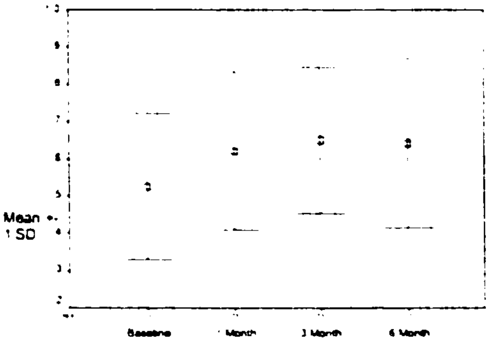


Figure 7 J: Proxy HUI2 OUS vs Time

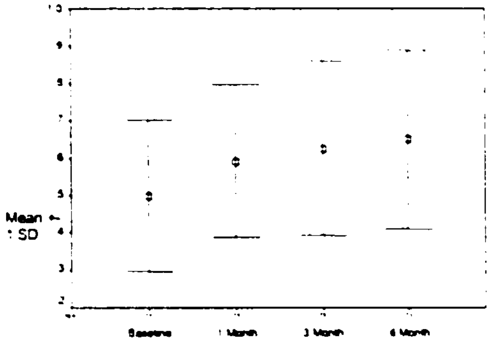


Figure 7 K: Patient HUI3 OUS vs Time

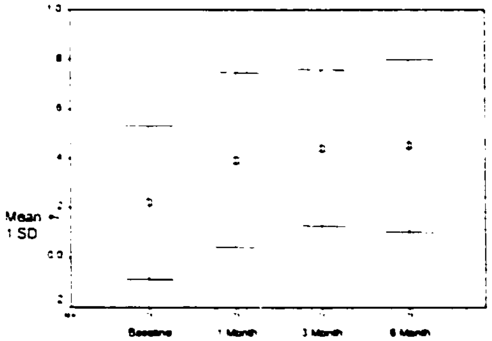
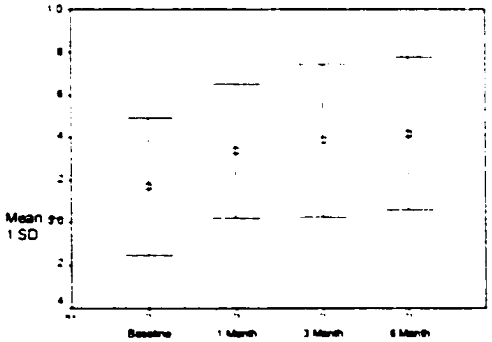


Figure 7 L:Proxy HUI3 OUS vs Time



Appendix 8: Scatterplots of Patient and Proxy Cross-Sectional Assessments

Figures 8 A to 8 D: EQ-VAS Scatterplots

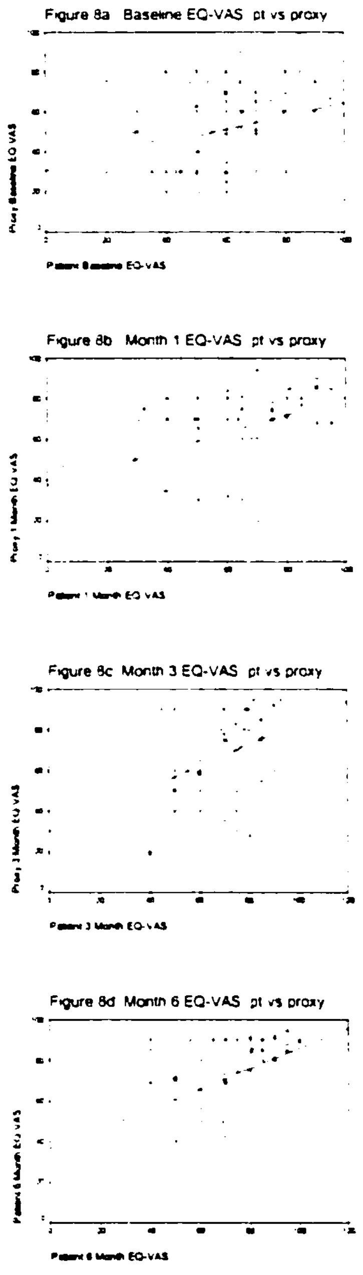


Figure 8a Baseline EQ-VAS pt vs proxy

Figure 8b Month 1 EQ-VAS pt vs proxy

Figure 8c Month 3 EQ-VAS pt vs proxy

Figure 8d Month 6 EQ-VAS pt vs proxy

Figures 8 E to 8 H: EQ-5D Index Score Scatterplots

Figure 8e Baseline EQ-Index pt vs proxy

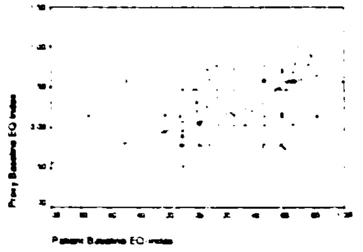


Figure 8f Month 1 EQ-Index pt vs proxy

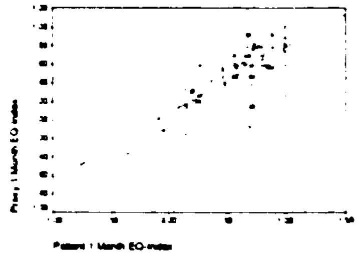


Figure 8g Month 3 EQ-Index pt vs proxy

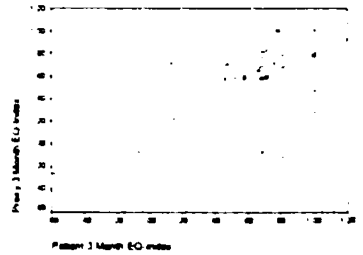
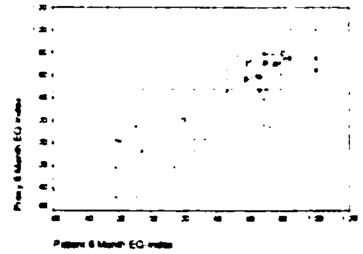


Figure 8h Month 6 EQ-Index pt vs proxy



Figures 8 I to 8 M: PCS-36 Score Scatterplots

Figure 8i Baseline PCS-36 pt vs proxy

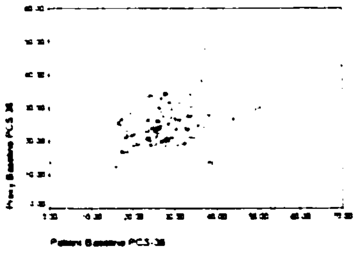


Figure 8j Month 1 PCS-36 pt vs proxy

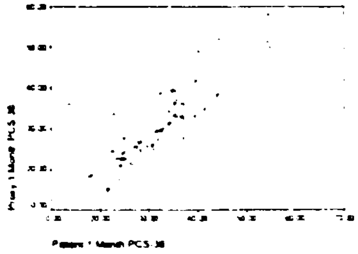


Figure 8k Month 3 PCS-36 pt vs proxy

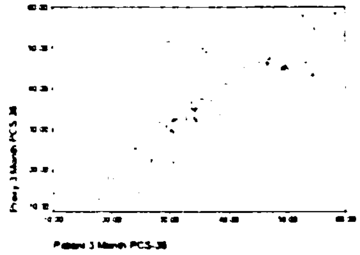
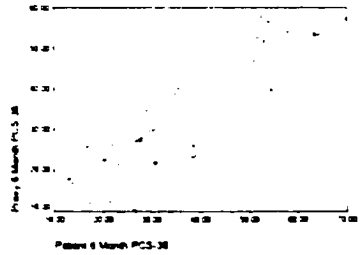


Figure 8l Month 6 PCS-36 pt vs proxy



Figures 8 M to 8 P: MCS-36 Score Scatterplots

Figure 8m. Baseline MCS-36 pt vs proxy

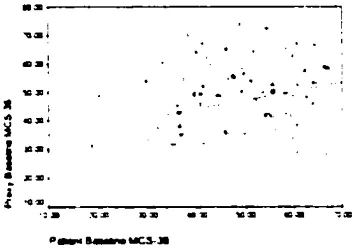


Figure 8n Month 1 MCS-36 pt vs proxy

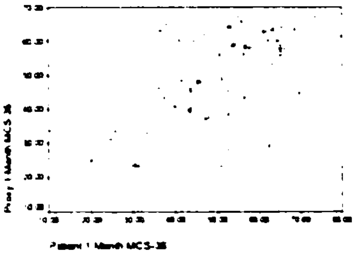


Figure 8o Month 3 MCS-36 pt vs proxy

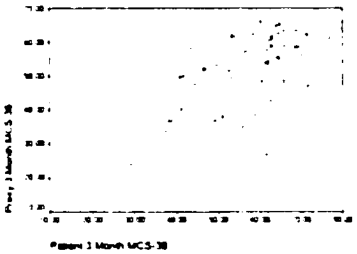
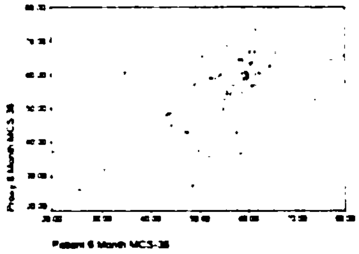


Figure 8p Month 6 MCS-36 pt vs proxy



Figures 8 I to 8 L: HUI2 OUS Scatterplots

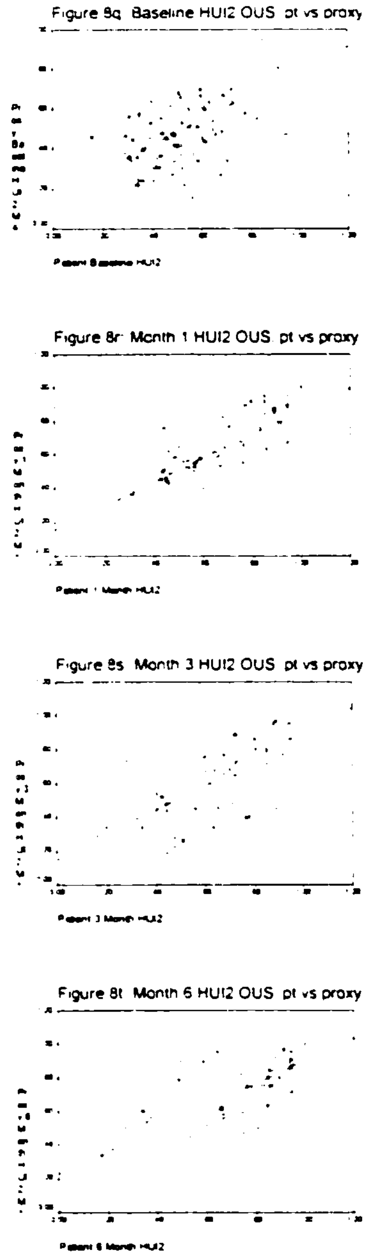


Figure 8q Baseline HUI2 OUS pt vs proxy

Figure 8r Month 1 HUI2 OUS pt vs proxy

Figure 8s Month 3 HUI2 OUS pt vs proxy

Figure 8t Month 6 HUI2 OUS pt vs proxy

Figures 8 U to 8 X: HUI3 OUS Scatterplots

Figure 8u Baseline HUI3 OUS pt vs proxy

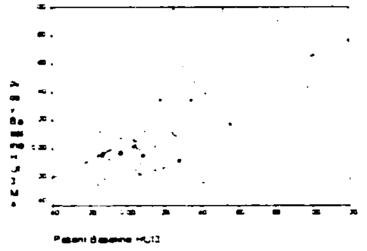


Figure 8v Month 1 HUI3 OUS pt vs proxy

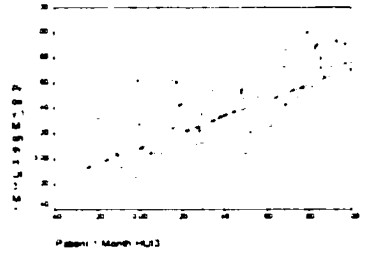


Figure 8w Month 3 HUI3 OUS pt vs proxy

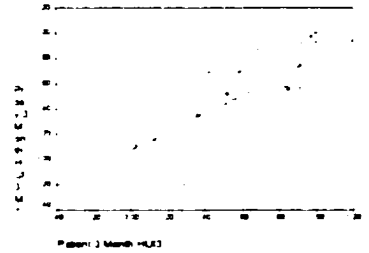
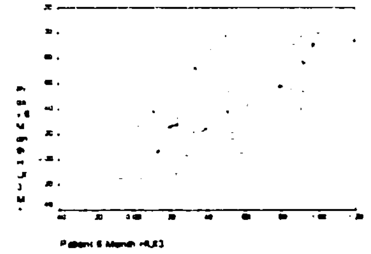
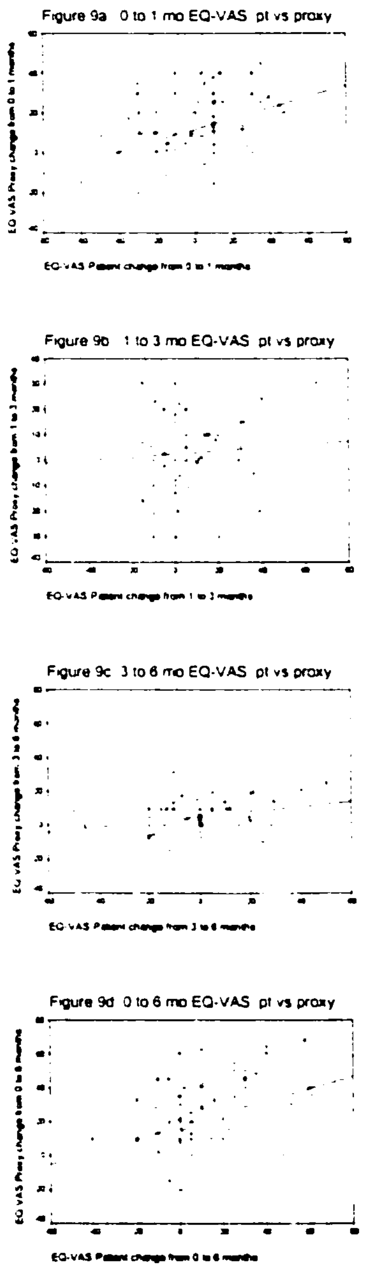


Figure 8x Month 6 HUI3 OUS pt vs proxy



Appendix 9: Scatterplots of Patient and Proxy Assessed Change Scores

Figures 9 A to 9 D: EQ-VAS Change Score Scatterplots



Figures 9 E to 9 H: EQ-Index Change Score Scatterplots

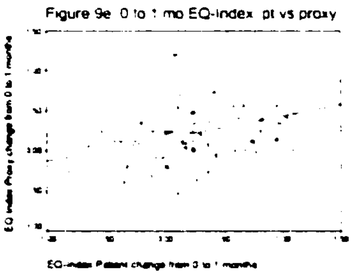


Figure 9e 0 to 1 mo EQ-index pt vs proxy

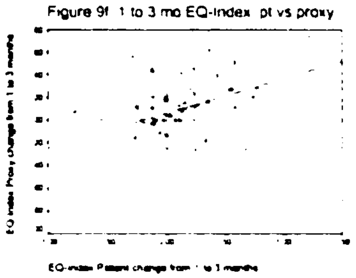


Figure 9f 1 to 3 mo EQ-index pt vs proxy

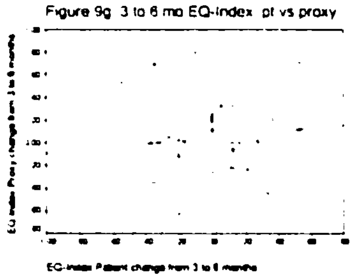


Figure 9g 3 to 6 mo EQ-index pt vs proxy

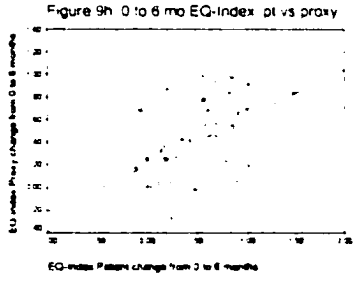
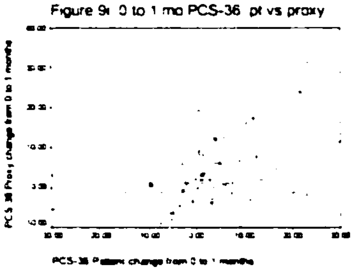
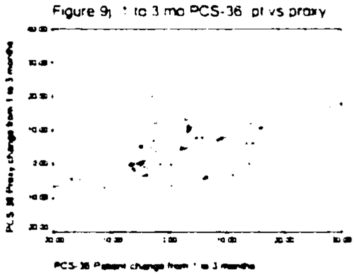
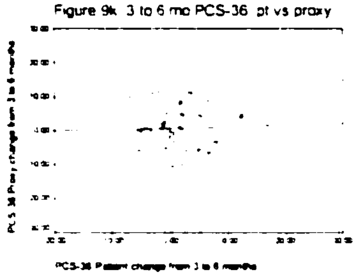


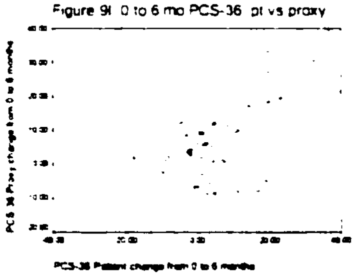
Figure 9h 0 to 6 mo EQ-index pt vs proxy

Figures 9 I to 9 L: PCS-36 Change Score Scatterplots

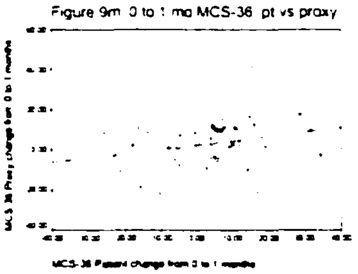


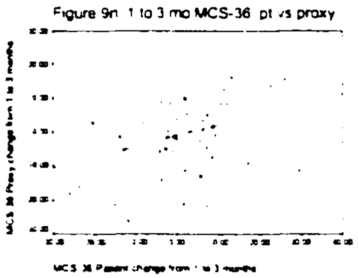


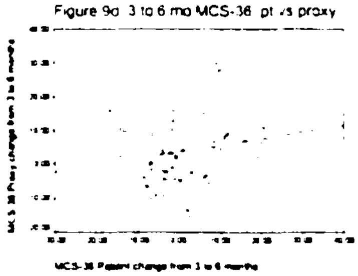


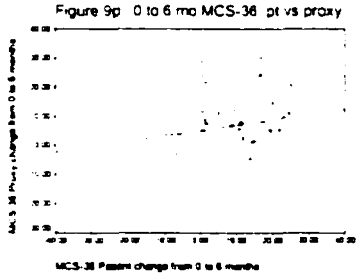


Figures 9 M to 9 P: MCS-36 Change Score Scatterplots









Figures 9 Q to 9 T: HUI2 OUS Change Score Scatterplots

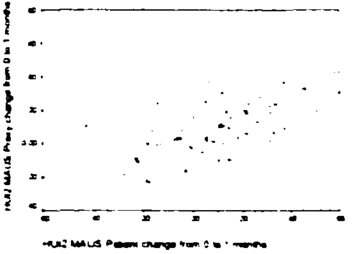


Figure 9q 0 to 1 mo HUI2 MAUS pt vs proxy

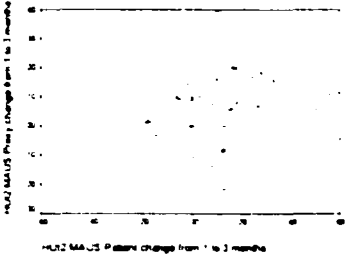


Figure 9r 1 to 3 mo HUI2 MAUS pt vs proxy

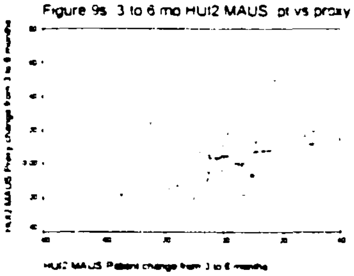


Figure 9s 3 to 6 mo HUI2 MAUS pt vs proxy

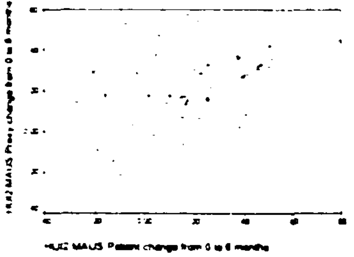


Figure 9t 0 to 6 mo HUI2 MAUS pt vs proxy

Figures 9 U to 9 X: HUI3 OUS Change Score Scatterplots

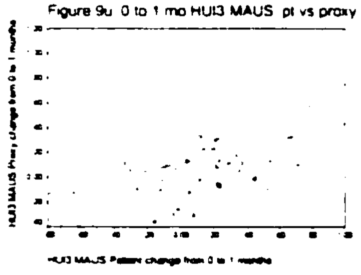


Figure 9u 0 to 1 mo HUI3 MAUS pt vs proxy

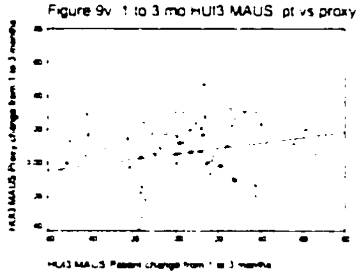


Figure 9v 1 to 3 mo HUI3 MAUS pt vs proxy

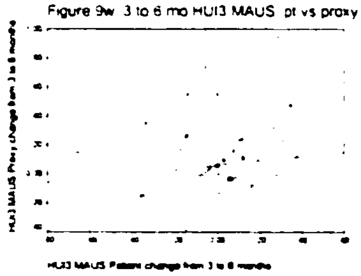


Figure 9w 3 to 6 mo HUI3 MAUS pt vs proxy

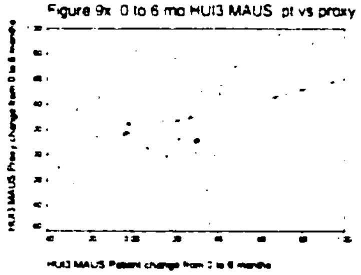


Figure 9x 0 to 6 mo HUI3 MAUS pt vs proxy

Appendix 10: Missing Data Tables

Table A10.A: Matching Variables and Bivariate Correlations for Hot Decking

HRQL Measure	Variable with missing value(s)	Primary Variable match	Bivariate correlation	Secondary Variable match (if primary unavailable)	Bivariate correlation
EQ-5D					
Mobility	W1r2eqmo	W1r2saa3	-0.778		
Self-Care	W1r2eqsc	W1r2saa3	-0.792		
Usual Activities	W1r2equa	W1r2pf	-0.801		
Pain/Discomfort	W1r2eqpd	W1r2bp	-0.742		
Anxiety/Depression	W1r2eqad	W1r2sae2	-0.710		
Self-Care	W6r1eqsc	W6r1pf	-0.665		
Pain/Discomfort	W6r2eqpd	W6r1bp	-0.777		
Anxiety/Depression	W6r2eqad	W6r2cesi	-0.573		
SF-36					
General Health	W0r2gh	W0r2vt	0.458		
General Health	W1r1gh	W1r1vt	0.612	W1r1vas	0.376
Bodily Pain	W1r1bp	W1r1sap3	0.738		
Mental Health	W1r1mh	W1r1cesi	-0.756		
Role Emotional	W1r1re	W1r1cesi	-0.536		
Social Functioning	W1r1sf	W1r1bart	-0.484		
Vitality	W1r1vt	W1r1cesi	-0.581		
General Health	W3r1gh	W3r1vit	0.687	W3r1vas	0.568
Mental Health	W3r1mh	W3r1cesi	-0.880		
Vitality	W3r1vt	W3r1gh	0.687	W3r1cesi	-0.536
HUI2/3					
Vision	W0r1sav3	W1r1sav3	0.495		
Hearing	W0r1sah3	W1r1sah3	0.514		
Speech	W0r1sas3	W1r1sas3	0.379		
Vision	W0r2sav3	W1r2sav3	0.598		
Hearing	W0r2sah3	W1r2sah3	0.693		
Emotion	W0r2sae3	W0r2sae2	0.506		
Pain	W0r2sap3	W0r2eqpd	-0.595		
Pain	W0r2sap2	W0r2eqpd	-0.555		
Vision	W1r1sav3	W3r1sav3	0.671		
Speech	W1r1sas3	W3r1sas3	0.590		
Cognition	W1r1sac3	W3r1sac3	0.411		
Emotion	W1r1sae2	W1r1sae3	0.651		
Cognition	W1r1sac2	W3r1sac2	0.426		
Self-Care	W1r1sat2	W1r1pf	0.565		
Pain	W1r1sap2	W1r1eqpd	-0.852		
Vision	W1r2sav3	W3r2sav3	0.616	W0r2sav3	0.598
Hearing	W1r2sah3	W0r2sah3	0.693		
Speech	W1r2sas3	W3r2sas3	0.552		

Hearing	W3r2sah3	W6r2sah3	0.519	W1r2sah3	0.490
Cognition	W3r2sac3	W6r2sac3	0.704		
Emotion	W3r2sae2	W3e2sae3	0.786		
Cognition	W3r2sac2	W6r2sac2	0.677		
Speech	W6r1sas3	W3r1sas3	0.508		
Ambulation	W6r1saa3	W6r1eqmo	-0.797		
Mobility	W6r1sam2	W6r1eqmo	-0.741		
Hearing	W6r2sah3	W6r1sah3	0.371		
Speech	W6r2sas3	W3r2sas3	0.454		
Cognition	W6r2sac3	W6r2sac3	0.704		
Cognition	W6r2sac2	W3r2sac2	0.677		

*w= wave; r= respondent;

e.g. W6r2sac2= 6 month proxy single attribute utility scores for HUI2

Table A10.B: Number of Assessments with Missing Items

HRQL Measure	Patient		Proxy	
EQ-5D VAS	Number of Assessments with Missing Items	Total Number of Missing Items	Number of Assessments with Missing Items	Total Number of Missing Items
Baseline	0	--	0	--
Time 1	0	--	0	--
Time 3	0	--	0	--
Time 6	0	--	0	--
EQ-5D Index				
Baseline	0	--	0	--
Time 1	0	--	2	6
Time 3	0	--	0	--
Time 6	1	1	2	3
SF-36				
Baseline	0	--	1	1
Time 1	3	6	0	--
Time 3	0	--	0	--
Time 6	0	--	0	--
HUI2*				
Baseline	0	4	1	4
Time 1	4	7	5	5
Time 3	0	--	2	4
Time 6	1	3	1	3
HUI3*				
Baseline	6	7	5	5
Time 1	3	3	5	5
Time 3	0	--	3	3
Time 6	4	4	3	3

*HUI non-item response data includes items requiring imputation due to incompatibility with scoring algorithm

Figure A10.C: PCS-36 Scores by Assessment Dropout

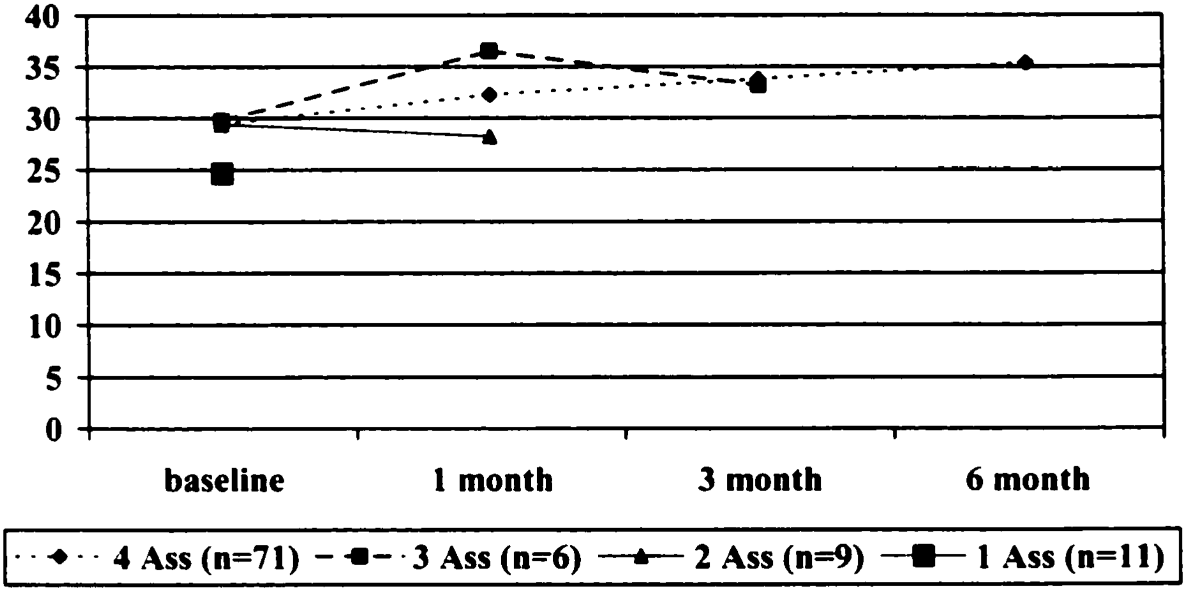


Figure A10.D: MCS-36 Scores by Assessment Dropout

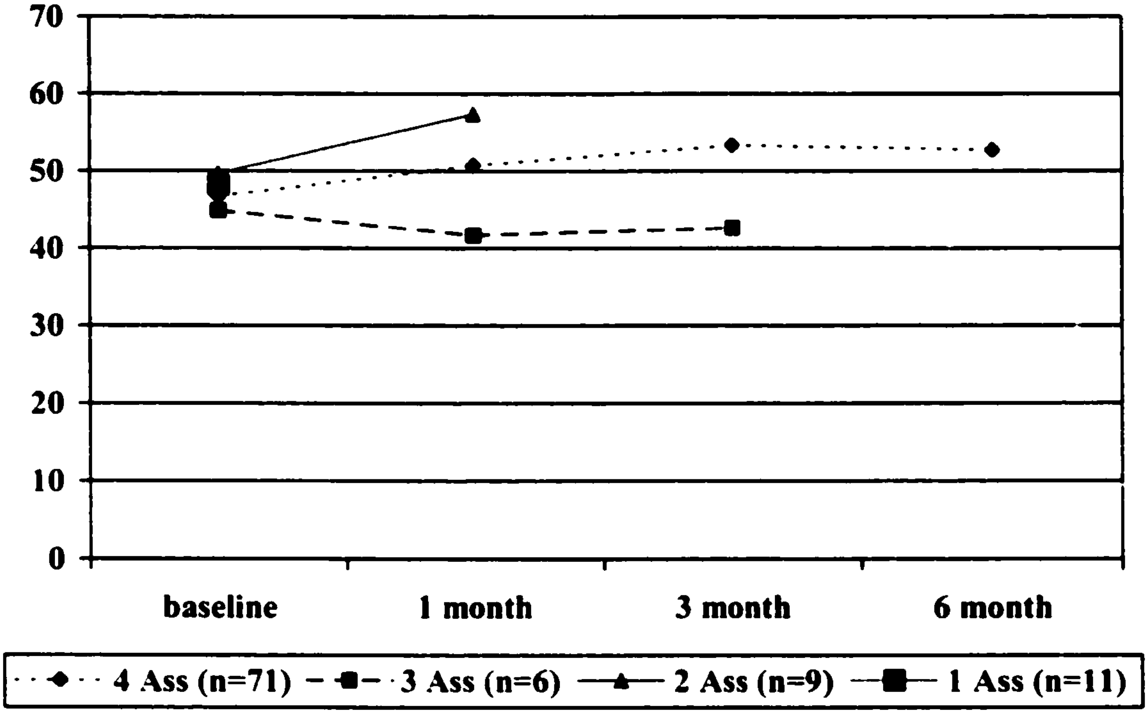


Figure A10.E: EQ-VAS Scores by Assessment Dropout

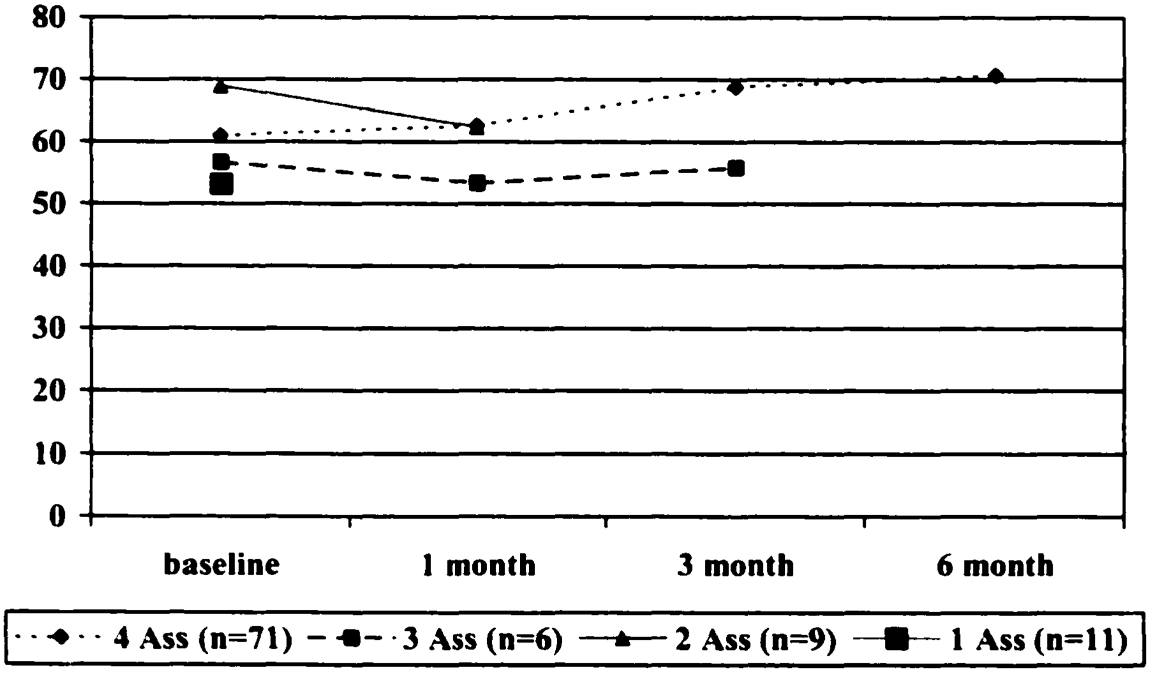


Figure A10.F: EQ-Index Scores by Assessment Dropout

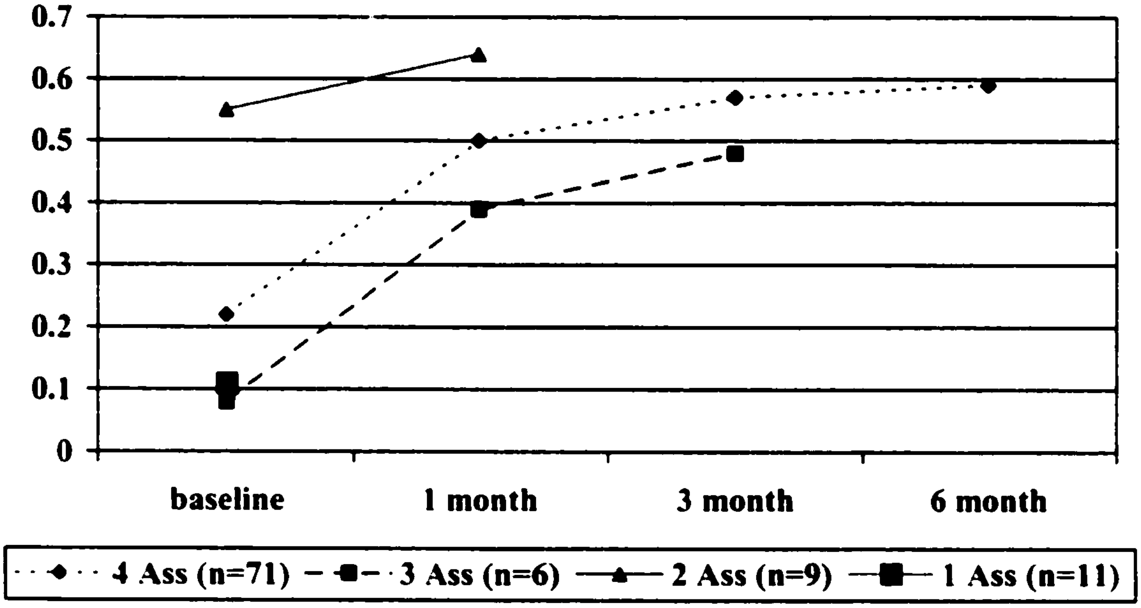


Figure A10.G: HUI2 OUS Scores by Assessment Dropout

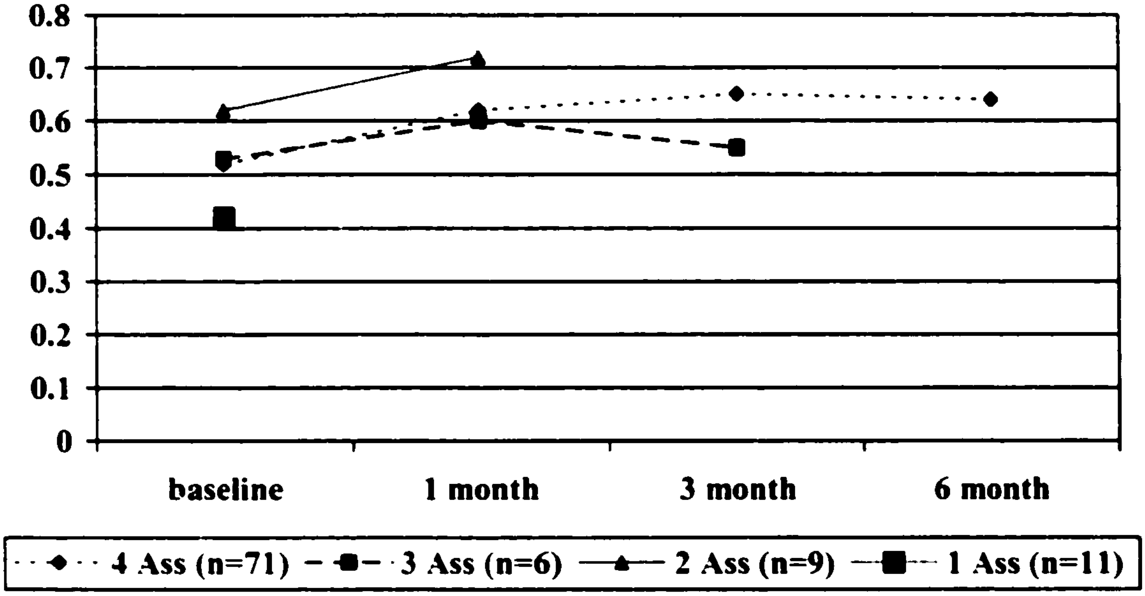
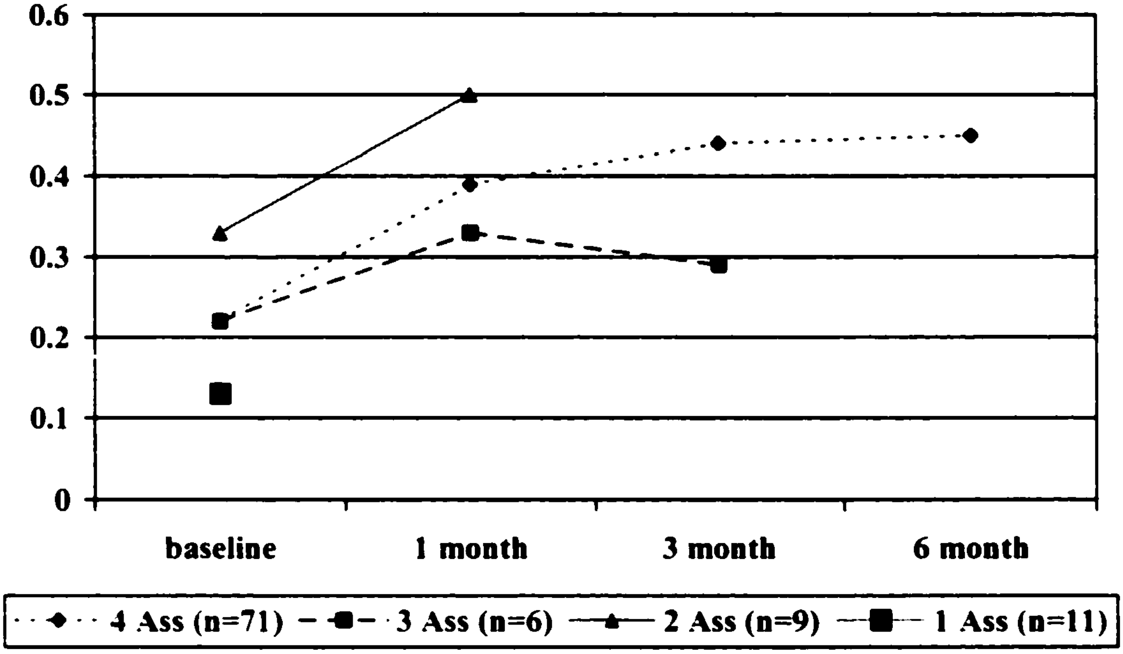


Figure A10.H: HUI3 OUS Scores by Assessment Dropout



Appendix 11:

Domain/Attribute Scores on EQ-5D, SF-36, HUI2/3

Table 11 A: SF-36 domains (Patient Self-Assessment)

	Valid N	Mean	Std Deviation	Median	Minimum	Maximum
Baseline						
Physical functioning	97	17.94	26.86	.00	.00	100.00
Role physical	97	8.51	23.90	.00	.00	100.00
Bodily pain	97	63.04	26.85	62.00	.00	100.00
General health	97	56.31	17.54	57.00	.00	90.00
Vitality	97	43.44	19.97	45.00	.00	90.00
Social Functioning	97	40.08	25.89	37.50	.00	100.00
Role Emotional	97	48.80	45.66	33.33	.00	100.00
Mental Health	97	69.69	18.63	72.00	12.00	100.00
Patient Baseline PCS-36	97	28.89	9.03	27.38	10.17	60.69
Patient Baseline MCS-36	97	47.10	11.33	46.45	18.60	67.14
Month 1						
Physical functioning	86	28.92	29.24	18.33	.00	100.00
Role physical	86	17.15	33.93	.00	.00	100.00
Bodily pain	86	69.05	28.23	73.00	.00	100.00
General health	86	61.44	23.21	62.00	15.00	100.00
Vitality	86	48.64	25.67	45.00	.00	100.00
Social Functioning	86	58.87	32.87	62.50	.00	100.00
Role Emotional	86	58.14	44.92	83.33	.00	100.00
Mental Health	86	74.49	22.22	82.00	16.00	100.00
Patient 1 Month PCS-36	86	32.15	10.53	31.08	12.84	60.02
Patient 1 Month MCS-36	86	50.76	12.93	52.58	20.12	73.51
Month 3						
Physical functioning	79	36.95	29.85	35.00	.00	100.00
Role physical	79	26.27	36.45	.00	.00	100.00
Bodily pain	79	66.23	30.72	72.00	.00	100.00
General health	79	61.77	23.81	67.00	5.00	100.00
Vitality	79	50.82	22.96	55.00	.00	95.00
Social Functioning	79	66.30	29.17	75.00	.00	100.00
Role Emotional	79	68.78	42.48	100.00	.00	100.00
Mental Health	79	74.13	23.30	80.00	4.00	100.00
Patient 3 Month PCS-36	79	33.67	11.37	32.91	14.56	58.29
Patient 3 Month MCS-36	79	52.18	12.71	53.75	19.35	76.64
Month 6						
Physical functioning	77	42.71	31.63	40.00	.00	100.00
Role physical	77	33.44	39.86	25.00	.00	100.00
Bodily pain	77	71.06	29.92	80.00	.00	100.00
General health	77	59.19	21.98	62.00	5.00	100.00
Vitality	77	52.55	22.62	55.00	.00	100.00
Social Functioning	77	62.50	31.15	62.50	12.50	100.00
Role Emotional	77	71.00	42.35	100.00	.00	100.00
Mental Health	77	79.58	16.61	84.00	32.00	100.00
Patient 6 Month PCS-36	77	35.06	12.44	34.52	12.23	63.34
Patient 6 Month MCS-36	77	52.66	10.36	54.78	24.24	73.84

Table 11 B: SF-36 domains (Proxy-Assessment)

	Valid N	Mean	Std Deviation	Median	Minimum	Maximum
Baseline						
Physical functioning	97	13.14	21.30	5.00	.00	100.00
Role physical	97	5.67	17.86	.00	.00	100.00
Bodily pain	97	50.55	28.72	42.00	.00	100.00
General health	97	51.18	22.11	50.00	5.00	97.00
Vitality	97	34.95	21.73	35.00	.00	80.00
Social Functioning	97	40.34	32.10	37.50	.00	100.00
Role Emotional	97	47.08	47.57	33.33	.00	100.00
Mental Health	97	67.38	19.67	72.00	.00	100.00
Patient Baseline PCS-36	97	25.66	8.43	24.01	11.79	53.20
Patient Baseline MCS-36	97	46.47	12.09	47.39	17.30	73.73
Month 1						
Physical functioning	83	26.59	28.85	15.00	.00	95.00
Role physical	83	11.75	24.49	.00	.00	100.00
Bodily pain	83	60.30	29.38	54.00	12.00	100.00
General health	83	52.65	22.46	52.00	10.00	100.00
Vitality	83	44.52	20.68	45.00	5.00	90.00
Social Functioning	83	54.67	28.67	50.00	.00	100.00
Role Emotional	83	59.44	45.40	100.00	.00	100.00
Mental Health	83	70.84	20.53	76.00	8.00	100.00
Patient 1 Month PCS-36	83	29.22	10.50	26.97	11.42	58.01
Patient 1 Month MCS-36	83	49.84	12.31	52.59	23.08	67.34
Month 3						
Physical functioning	79	32.66	31.54	20.00	.00	100.00
Role physical	79	30.70	39.42	.00	.00	100.00
Bodily pain	79	60.87	27.50	62.00	.00	100.00
General health	79	53.92	24.65	52.00	.00	100.00
Vitality	79	45.99	21.89	45.00	.00	90.00
Social Functioning	79	61.87	28.93	62.50	.00	100.00
Role Emotional	79	54.43	47.79	66.67	.00	100.00
Mental Health	79	71.00	20.71	72.00	.00	100.00
Patient 3 Month PCS-36	79	32.81	11.95	32.22	13.02	58.37
Patient 3 Month MCS-36	79	48.92	12.45	49.64	16.34	68.79
Month 6						
Physical functioning	76	34.61	31.76	25.00	.00	100.00
Role physical	76	36.84	42.13	25.00	.00	100.00
Bodily pain	76	63.43	28.56	62.00	.00	100.00
General health	76	55.19	24.49	52.00	.00	97.00
Vitality	76	48.22	24.23	47.50	.00	100.00
Social Functioning	76	66.61	29.82	75.00	.00	100.00
Role Emotional	76	67.98	40.53	100.00	.00	100.00
Mental Health	76	75.34	20.13	84.00	.00	100.00
Patient 6 Month PCS-36	76	33.01	12.48	31.05	10.71	57.77
Patient 6 Month MCS-36	76	52.16	11.94	56.32	17.34	73.16

Table 11 C: HUI 2 Single Attribute Scores (Patient Self-Assessment)

	Valid N	Mean	Std Deviation	Median	Minimum	Maximum
Baseline						
sensation	97	.70	.26	.65	.00	1.00
mobility	97	.51	.30	.61	.00	1.00
emotion	97	.90	.13	.86	.37	1.00
cognition	97	.92	.09	.86	.66	1.00
self care	97	.40	.46	.00	.00	1.00
pain	97	.91	.16	.95	.42	1.00
HUI2 OUS	97	.52	.19	.50	.16	1.00
Month 1						
sensation	86	.81	.15	.87	.00	1.00
mobility	86	.67	.29	.61	.00	1.00
emotion	86	.90	.14	1.00	.37	1.00
cognition	86	.90	.16	.86	.00	1.00
self care	86	.64	.45	.85	.00	1.00
pain	86	.89	.21	.95	.00	1.00
HUI2 OUS	86	.63	.21	.67	.06	1.00
Month 3						
sensation	79	.80	.15	.87	.00	1.00
mobility	79	.71	.25	.61	.34	1.00
emotion	79	.90	.19	1.00	.00	1.00
cognition	79	.90	.10	.86	.66	1.00
self care	79	.74	.41	1.00	.00	1.00
pain	79	.86	.22	.95	.00	1.00
HUI2 OUS	79	.64	.21	.63	.16	1.00
Month 6						
sensation	77	.68	.31	.87	.00	1.00
mobility	77	.74	.28	.92	.00	1.00
emotion	77	.93	.11	1.00	.37	1.00
cognition	77	.91	.08	.86	.66	1.00
self care	77	.65	.46	1.00	.00	1.00
pain	77	.90	.18	.95	.00	1.00
HUI2 OUS	77	.64	.23	.66	.18	1.00

Table 11 D: HUI 2 Single Attribute Scores (Proxy Assessment)

	Valid N	Mean	Std Deviation	Median	Minimum	Maximum
Baseline						
Sensation	97	.74	.14	.65	.00	1.00
Mobility	97	.55	.25	.61	.00	1.00
Emotion	97	.85	.17	.86	.00	1.00
Cognition	97	.87	.17	.86	.00	1.00
self care	97	.36	.44	.00	.00	1.00
Pain	97	.84	.21	.95	.00	1.00
HUI2 OUS	97	.50	.20	.47	.08	.95
Month 1						
Sensation	83	.79	.12	.87	.65	1.00
Mobility	83	.65	.26	.61	.00	1.00
Emotion	83	.88	.13	.86	.37	1.00
Cognition	83	.88	.17	.86	.00	1.00
self care	83	.61	.43	.85	.00	1.00
Pain	83	.87	.19	.95	.00	1.00
HUI2 OUS	83	.59	.21	.58	.08	1.00
Month 3						
Sensation	79	.79	.17	.87	.00	1.00
Mobility	79	.72	.25	.61	.34	1.00
Emotion	79	.88	.18	.86	.00	1.00
Cognition	79	.85	.18	.86	.00	1.00
self care	79	.66	.44	.85	.00	1.00
Pain	79	.86	.21	.95	.00	1.00
HUI2 OUS	79	.62	.24	.67	.11	1.00
Month 6						
Sensation	76	.76	.23	.87	.00	1.00
Mobility	76	.75	.25	.92	.00	1.00
Emotion	76	.90	.17	1.00	.00	1.00
Cognition	76	.87	.18	.86	.00	1.00
self care	76	.74	.39	1.00	.00	1.00
Pain	76	.86	.20	.95	.00	1.00
HUI2 OUS	76	.65	.23	.64	-.02	1.00

Table 11 E: HUI 3 Single Attribute Scores (Patient Self-Assessment)

	Valid N	Mean	Std Deviation	Median	Minimum	Maximum
Baseline						
Vision	97	.91	.16	.95	.00	1.00
Hearing	97	.87	.28	1.00	.00	1.00
Speech	97	.83	.19	.82	.00	1.00
Ambulation	97	.40	.37	.36	.00	1.00
Dexterity	97	.54	.35	.45	.00	1.00
Emotion	97	.85	.19	.91	.33	1.00
Cognition	97	.86	.18	.92	.32	1.00
Pain	97	.85	.24	.92	.00	1.00
HUI3 OUS	97	.22	.30	.15	-.22	1.00
Month 1						
Vision	86	.89	.18	.95	.00	1.00
Hearing	86	.91	.19	1.00	.32	1.00
Speech	86	.95	.11	1.00	.67	1.00
Ambulation	86	.58	.36	.67	.00	1.00
Dexterity	86	.71	.35	.88	.00	1.00
Emotion	86	.89	.21	1.00	.00	1.00
Cognition	86	.87	.20	.92	.00	1.00
Pain	86	.85	.25	.92	.00	1.00
HUI3 OUS	86	.40	.35	.40	-.29	.97
Month 3						
Vision	79	.87	.18	.95	.00	1.00
Hearing	79	.92	.17	1.00	.32	1.00
Speech	79	.96	.11	1.00	.67	1.00
Ambulation	79	.65	.32	.67	.00	1.00
Dexterity	79	.78	.31	.88	.00	1.00
Emotion	79	.88	.22	1.00	.00	1.00
Cognition	79	.84	.21	.92	.32	1.00
Pain	79	.84	.22	.92	.00	1.00
HUI3 OUS	79	.42	.33	.43	-.27	1.00
Month 6						
Vision	77	.90	.17	.95	.38	1.00
Hearing	77	.78	.36	1.00	.00	1.00
Speech	77	.93	.13	1.00	.67	1.00
Ambulation	77	.68	.34	.83	.00	1.00
Dexterity	77	.79	.31	1.00	.00	1.00
Emotion	77	.94	.13	1.00	.33	1.00
Cognition	77	.87	.16	.92	.32	1.00
Pain	77	.87	.19	.92	.00	1.00
HUI3 OUS	77	.45	.35	.46	-.19	1.00

Table 11 F: HUI 3 Single Attribute Scores (Proxy Assessment)

	Valid N	Mean	Std Deviation	Median	Minimum	Maximum
Baseline						
Vision	97	.86	.19	.95	.38	1.00
Hearing	97	.93	.16	1.00	.48	1.00
Speech	97	.85	.18	1.00	.00	1.00
Ambulation	97	.38	.35	.36	.00	1.00
Dexterity	97	.60	.34	.73	.00	1.00
Emotion	97	.82	.19	.91	.33	1.00
Cognition	97	.77	.25	.86	.00	1.00
Pain	97	.67	.36	.77	.00	1.00
HUI3 OUS	97	.18	.33	.08	-.29	.97
Month 1						
Vision	83	.87	.19	.95	.38	1.00
Hearing	83	.92	.17	1.00	.32	1.00
Speech	83	.93	.13	1.00	.67	1.00
Ambulation	83	.54	.35	.67	.00	1.00
Dexterity	83	.71	.32	.88	.00	1.00
Emotion	83	.87	.17	.91	.33	1.00
Cognition	83	.82	.23	.86	.00	1.00
Pain	83	.82	.19	.92	.48	1.00
HUI3 OUS	83	.33	.32	.28	-.29	1.00
Month 3						
Vision	79	.90	.14	.95	.38	1.00
Hearing	79	.90	.22	1.00	.00	1.00
Speech	79	.94	.12	1.00	.67	1.00
Ambulation	79	.61	.34	.67	.00	1.00
Dexterity	79	.80	.24	.88	.20	1.00
Emotion	79	.87	.20	.91	.00	1.00
Cognition	79	.77	.27	.86	.00	1.00
Pain	79	.77	.29	.92	.00	1.00
HUI3 OUS	79	.37	.36	.43	-.27	1.00
Month 6						
Vision	76	.89	.18	.95	.38	1.00
Hearing	76	.87	.27	1.00	.00	1.00
Speech	76	.91	.17	1.00	.00	1.00
Ambulation	76	.67	.34	.83	.00	1.00
Dexterity	76	.80	.26	.88	.00	1.00
Emotion	76	.88	.20	.91	.00	1.00
Cognition	76	.80	.25	.92	.00	1.00
Pain	76	.81	.24	.92	.00	1.00
HUI3 OUS	76	.42	.36	.43	-.36	1.00

Table 11 G: EQ-5D Domain Responses (Patient Self-Assessment)

	Mobility		Usual Activities		Self Care		Pain/ Discomfort		Anxiety/ Depression	
Baseline	Count	%	Count	%	Count	%	Count	%	Count	%
No problems	14	14.4	17	17.5	9	9.3	44	45.4	35	36.1
Some problems	48	49.5	42	43.3	41	42.3	49	50.5	58	59.8
Extreme problems	35	36.1	38	39.2	47	48.5	4	4.1	4	4.1
Month 1										
No problems	28	32.6	46	53.5	25	29.1	35	40.7	50	58.1
Some problems	50	58.1	28	32.6	39	45.3	45	52.3	33	38.4
Extreme problems	8	9.3	12	14.0	22	25.6	6	7.0	3	3.5
Month 3										
No problems	27	34.2	42	53.2	23	29.1	29	36.7	43	54.4
Some problems	49	62.0	32	40.5	43	54.4	47	59.5	32	40.5
Extreme problems	3	3.8	5	6.3	13	16.5	3	3.8	4	5.1
Month 6										
No problems	30	39.0	42	54.5	29	37.7	38	49.4	47	61.0
Some problems	38	49.4	29	37.7	34	44.2	37	48.1	30	39.0
Extreme problems	9	11.7	6	7.8	14	18.2	2	2.6		

Table 11 H: EQ-5D Domain Responses (Proxy Assessment)

	Mobility		Usual Activities		Self Care		Pain/ Discomfort		Anxiety/ Depression	
Baseline	Count	%	Count	%	Count	%	Count	%	Count	%
No problems	11	11.3	19	19.6	8	8.2	29	29.9	27	27.8
Some problems	47	48.5	32	33.0	35	36.1	62	63.9	64	66.0
Extreme problems	39	40.2	46	47.4	54	55.7	6	6.2	6	6.2
Month 1										
No problems	20	24.1	30	36.1	18	21.7	25	30.1	40	48.2
Some problems	53	63.9	38	45.8	39	47.0	55	66.3	42	50.6
Extreme problems	10	12.0	15	18.1	26	31.3	3	3.6	1	1.2
Month 3										
No problems	24	30.4	37	46.8	19	24.1	23	29.1	39	49.4
Some problems	49	62.0	34	43.0	42	53.2	51	64.6	37	46.8
Extreme problems	6	7.6	8	10.1	18	22.8	5	6.3	3	3.8
Month 6										
No problems	25	32.9	35	46.1	24	31.6	30	39.5	36	47.4
Some problems	45	59.2	31	40.8	38	50.0	43	56.6	36	47.4
Extreme problems	6	7.9	10	13.2	14	18.4	3	3.9	4	5.3

Appendix 12: Bivariate Correlations for Construct Validity

Table 12 A: Pearson Correlations for Patient Baseline Scores

Correlations

	PF	RP	BP	GH	VT	SF	RE	MH	PCS 36	MCS 36	sav3	sah3	sas3	saa3	sad3	sae3	sac3	sap3	OUS3	sas2	sam2	sae2	sac2	sat2	sap2	OUS2	EQ VAS	EQ Index	MRH	NIH SS	SSS 48	Bart Indx	Ces D
PF	1 0	31**	13	17	39**	36**	-10	07	75**	-12	08	09	19	60**	48**	13	-10	16	52**	19	60**	-03	00	63**	07	60**	41**	42**	-54**	-28**	47**	57**	-23
RP	31**	1 0	16	19	23	15	-04	-02	58**	-11	00	16	19	13	15	08	-08	02	18	17	12	-11	-07	16	-02	16	23	-03	-06	06	02	05	-14
BP	13	16	1 0	21	20	20	11	19	49**	09	02	15	05	06	-03	00	08	37**	14	13	00	22	06	01	51**	27**	22	11	-14	07	05	11	-20
GH	17	19	21	1 0	48**	27**	27**	36**	31**	38**	22	22	18	19	19	07	26	13	33**	26	15	26	25	19	14	33**	33**	19	-16	-10	14	21	-31**
VT	39**	23	20	48**	1 0	35**	28**	31**	34**	44**	11	16	05	25	10	10	-01	05	21	13	24	08	05	33**	-08	27**	30**	21	-18	-09	15	29**	-45**
SF	36**	15	20	27**	35**	1.0	21	21	27**	43**	13	-05	-02	39**	26**	13	-07	05	23	04	34**	16	02	39**	24	41**	37**	36**	-35**	-25	29**	42**	-38**
RE	-10	-04	11	27**	28**	21	1 0	41**	-35**	83**	-05	-11	09	-10	-09	09	16	06	05	-02	-08	18	15	-10	10	04	15	-06	-03	10	-11	00	-44**
MH	07	-02	19	36**	31**	21	41**	1 0	-17	74**	-09	-17	11	04	04	26**	20	20	16	-03	00	47**	20	-01	24	17	15	15	-03	00	00	04	-65**
PCS-36	75**	58**	49**	31**	34**	27**	-35**	-17	1 0	-39**	14	30**	17	46**	36**	02	-08	21	43**	29**	43**	-07	-03	48**	17	51**	38**	30**	-39**	-16	36**	42**	-06
MCS-36	-12	-11	09	38**	44**	43**	83**	74**	-39**	1 0	-03	-17	03	-04	-06	18	17	07	05	-06	-05	35**	18	-05	14	08	17	06	-01	02	-07	04	-60**
sav3	08	00	02	22	11	13	-05	-09	14	-03	1 0	10	17	12	18	-06	26**	-02	26	34**	14	04	20	06	-09	22	02	02	-13	-19	10	09	-20
sah3	09	16	15	22	16	-05	-11	-17	30**	-17	10	1 0	11	14	13	-09	-01	02	31**	73**	12	-21	-02	05	-15	24	02	02	-11	-01	08	06	-16
sas3	19	19	05	18	05	-02	09	11	17	03	17	11	1 0	12	32**	22	10	06	44**	65**	10	11	04	16	04	37**	06	05	-19	-24	21	15	-12
saa3	60**	13	06	19	25	39**	-10	04	46**	-04	12	14	12	1 0	65**	21	-23	08	74**	20	92**	-03	-16	68**	13	77**	45**	64**	-71**	-53**	75**	77**	-18
sad3	48**	15	-03	19	10	26**	-09	04	36**	-06	18	13	32**	65**	1.0	23	-04	05	76**	29**	69**	06	-06	58**	13	66**	30**	51**	-60**	-58**	63**	61**	-12
sae3	13	08	00	07	10	13	09	26**	02	18	-06	-09	22	21	23	1.0	04	18	41**	07	20	39**	06	14	28**	31**	18	23	-09	-17	14	18	-24
sac3	-10	-08	08	26	-01	-07	16	20	-08	17	26**	-01	10	-23	-04	04	1 0	-02	19	00	-21	18	91**	-17	-01	-04	-14	-13	18	00	-16	-20	-09
sap3	16	02	37**	13	05	05	06	20	21	07	-02	02	06	08	05	18	-02	1 0	29**	06	05	18	-02	12	41**	25	17	10	-12	07	03	02	-13
MAUS3	52**	18	14	33**	21	23	05	16	43**	05	26	31**	44**	74**	76**	41**	19	29**	1 0	47**	70**	10	20	54**	20	83**	39**	45**	-63**	-54**	63**	62**	-34**
sas2	19	17	13	26	13	04	-02	-03	29**	-06	34**	73**	65**	20	29**	07	00	06	47**	1 0	15	-03	-02	15	-08	46**	12	03	-21	-15	19	15	-30**
sam2	60**	12	00	15	24	34**	-08	00	43**	-05	14	12	10	92**	69**	20	-21	05	70**	15	1.0	-03	-15	66**	11	77**	43**	59**	-69**	-55**	70**	72**	-17
sae2	-03	-11	22	26	08	16	18	47**	-07	35**	04	-21	11	-03	06	39**	18	18	10	-03	-03	1 0	16	-03	33**	24	09	11	-05	-06	-13	-03	-28**
sac2	00	-07	06	25	05	02	15	20	-03	18	20	-02	04	-16	-06	06	91**	-02	20	-02	-15	16	1 0	-14	-02	00	-10	-06	19	-03	-11	-17	-18
sat2	63**	16	01	19	33**	39**	-10	-01	48**	-05	06	05	16	68**	58**	14	-17	12	54**	15	66**	-03	-14	1 0	08	71**	45**	49**	-64**	-42**	56**	68**	-14
sap2	07	-02	51**	14	-08	24	10	24	17	14	-09	-15	04	13	13	28**	-01	41**	20	-08	11	33**	-02	08	1.0	43**	10	24	-21	-12	16	25	-15
MAUS2	60**	16	27**	33**	27**	41**	04	17	51**	08	22	24	37**	77**	66**	31**	-04	25	83**	46**	77**	24	00	71**	43**	1 0	46**	53**	-73**	-53**	63**	72**	-40**
EQ-VAS	41**	23	22	33**	30**	37**	15	15	38**	17	02	02	06	45**	30**	18	-14	17	39**	12	43**	09	-10	45**	10	46**	1.00	37**	-39**	-24	39**	42**	-31**
EQ-Index	42**	-03	11	19	21	36**	-06	15	30**	06	02	02	05	64**	51**	23	-13	10	45**	03	59**	11	-06	49**	24	53**	37**	1 00	-52**	-41**	56**	60**	-14
MRHS	-54**	-06	-14	-16	-18	-35**	-03	-03	-39**	-01	-13	-11	-19	-71**	-60**	-09	18	-12	-63**	-21	-69**	-05	19	-64**	-21	-73**	-39**	-52**	1 0	64**	-78**	-86**	24
NIHSS	-28**	06	07	-10	-09	-25	10	00	-16	02	-19	-01	-24	-53**	-58**	-17	00	07	-54**	-15	-55**	-06	-03	-42**	-12	-53**	-24	-41**	64**	1 00	-77**	-68**	12
SSS-48	47**	02	05	14	15	29**	-11	00	36**	-07	10	08	21	75**	63**	14	-16	03	63**	19	70**	-13	-11	56**	16	63**	39**	56**	-78**	-77**	1.0	86**	-16
Bart Indx	57**	05	11	21	29**	42**	00	04	42**	04	09	06	15	77**	61**	18	-20	02	62**	15	72**	-03	-17	68**	25	72**	42**	60**	-86**	-68**	86**	1 00	-22
cesd	-23	-14	-20	-31**	-45**	-38**	-44**	-65**	-06	-60**	-20	-16	-12	-18	-12	-24	-09	-13	-34**	-30**	-17	-28**	-16	-14	-15	-40**	-31**	-14	24	12	-16	-22	1.00

**. correlation is significant at the 0.01 level (2-tailed), *correlation is significant at the 0 05 level (2-tailed) .

Correlations

	PF	RP	BP	GH	VT	SF	RE	MH	PCS 36	MCS 36	sav3	sah3	sas3	saa3	sad3	sae3	sac3	sap3	OUS3	sas2	sam2	sae2	sac2	sat2	sap2	OUS2	EQ VAS	EQ Index	MRH	NIH SS	SSS 48	Bart Indx	Ces D
PF	1 0	55**	32	19	.33**	32**	11	14	74**	.04	19	12	22	60**	46**	31	14	18	58**	19	60**	21	16	57**	25	63**	39**	63**	-66**	-42**	57**	64**	-37
RP	55**	1 0	40	20	32	44	19	14	64**	17	15	14	13	47	23	24	11	21	46	10	47	14	11	.31	19	44	32	34	-41	-19	37	39	-32
BP	32	40	1 0	29	44	36	23	29	64**	29	24	-04	10	33	29	26	02	50**	40	03	27	11	01	19	44**	.35**	33	30	-29	-06	.20	23	-32
GH	.19	20	29	1 0	33**	.31**	23**	32**	43**	34**	29	24	03	17	25	26	33	14	36**	11	15	22	26	12	30	34**	46**	24	-19	-08	08	17	-39**
VT	33**	32	44	33**	1 0	.41**	29**	42**	35**	55**	22	14	16	42	33	40	20	21	49	27	38	25	25	22**	31	47**	45**	38	-40	-27	32	40**	-47**
SF	.32**	44	36	31**	41**	1 0	19	50	33**	59**	29	19	11	40**	34**	45	24	33	56	10	35**	28	19	.37**	32	48**	55**	49**	-28**	-.15	.28**	33**	-49**
RE	.11	19	23	23**	.29**	.19	1 0	39**	-15**	.75**	-05	05	29	19	12	31	24	20	34	17	15	30	20	14	.09	30	30	25	-.14	-.13	.07	10	-47**
MH	14	14	29	32**	.42**	.50	.39**	1 0	-07	82**	.13	13	29	28	30	58**	30	29	49	24	25	62**	27	33	18	49	42	37	-.12	-.10	.08	12	-76**
PCS-36	.74**	64**	64**	43**	35**	.33**	-.15**	-07	1 0	-17**	.29	09**	02	47**	37**	15	04	25	44**	05**	46**	-01	05	34**	38	45**	35**	41**	-53**	-24	44**	.50**	-17
MCS-36	04	17	29	34**	55**	59**	.75**	82**	-17**	1.0	.13	15	28	26	24	.54	35	29	50	23	21	50**	30	23	20	43	48	36	-12	-11	.08	13	-71**
sav3	.19	15	24	29	.22	.29	-05	19	.29	13	1.0	09	12	12	16	07	19**	23	34	33**	11	06	13	17	19	24	40	33	-14	-.17	.10	.11	-30
sah3	12	14	-04	.24	.14	.19	05	13	09**	15	.09	1 0	04	14	-06	14	27	-04	26**	.21**	.09	08	.16	08	-05	.13	08	05	00	04	.00	.01	-25
sas3	.22	13	10	03	.16	.11	29	29	02	.28	.12	04	1 0	26	24**	14	11	08	.33**	80**	27	09	09	25	10	41**	06	22	-23	-.38	26	23	-37
saa3	.60**	47	33	17	.42	.40**	19	28	47**	26	.12	14	26	1 0	54**	.44	07	43	80**	24	94**	15	13	.55**	40	.77**	37**	68**	-66**	-.49**	.70**	65**	-45
sad3	46**	23	29	25	.33	.34**	.12	30	37**	24	.16	-06	24**	54**	1 0	34	19	29	68**	.17**	.52**	21	20	61**	48	64**	40**	56**	-49**	-48**	63**	53**	-45
sae3	.31	24	26	26	.40	.45	.31	58**	15	.54	07	14	14	44	34	1 0	31	32	61**	.13	38	51**	23	.32	26**	.55**	40	45	-27	-31	.32	28	-65
sac3	.14	11	02	33	.20	24	.24	30	04	.35	.19**	27	11	07	19	31	1.0	08	48	22	10	.44	.91**	18	12	.46	29	17	-13	-.15	.06	.10	-43
sap3	.18	21	50**	14	21	33	.20	.29	25	.29	23	-04	08	43	29	32	.08	1 0	56**	-04	39	.11	08	13	.30**	36	34	44	-13	-07	.14	13	-30
MAUS3	58**	46	40	36**	.49	.56	.34	49	44**	50	.34	26**	33**	80**	68**	61**	.48	56**	1 0	.36**	75**	39	44	.61**	.43	88**	58**	71**	-.55**	-47**	58**	.55**	-67**
sas2	.19	10	03	.11	.27	.10	.17	24	.05**	23	33**	21**	80**	24	17**	13	.22	-.04	.36**	1.0	.23	07	23	24	.00	40**	15	17	-.22	-34	.22	.21	-32**
sam2	60**	47	27	15	.38	.35**	15	25	.46**	.21	.11	09	27	94**	52**	38	.10	39	75**	23	1 0	.15	19	54**	41	.78**	29**	63**	-64**	-.48**	.68**	.64**	-44
sae2	.21	14	11	.22	.25	.28	.30	62**	-01	.50**	.06	08	09	13	21	.51**	44	.11	39	07	15	1 0	38	.17	02**	.46	30	25	-08	-09	04	12	-58**
sac2	.16	11	01	26	.25	.19	.20	27	05	.30	.13	.16	09	13	20	23	91**	08	.44	.23	19	38	1.0	.21	.11	.47	26	.18	-.18	-.15	.13	16	-36
sat2	.57**	31	19	.12	.22**	.37**	.14	33	.34**	23	.17	08	25	55**	61**	32	18	13	61**	24	54**	17	21	1.0	.24	72**	33**	59**	-.57**	-39**	.61**	.58**	-37
sap2	.25	.19	44**	30	.31	.32	.09	.18	38	.20	.19	-05	.10	40	48	26**	.12	30**	.43	00	.41	02**	.11	24	1 0	.57**	27	40	-32	-.27	.37	.30	-37
MAUS2	.63**	44	35**	34**	.47**	.48**	.30	49	.45**	43	.24	13	41**	77**	.64**	55**	46	36	88**	40**	.78**	46	47	72**	.57**	1.0	47**	67**	-63**	-.48**	.61**	62**	-66**
EQ-VAS	.39**	32	.33	46**	.45**	.55**	30	42	.35**	48	.40	08	06	37**	40**	.40	29	34	.58**	15	.29**	30	26	33**	27	47**	1 00	63**	-.49**	-.35	.43**	51**	-48**
EQ-Index	.63**	34	30	24	.38	49**	.25	37	.41**	36	33	05	22	68**	56**	45	17	44	71**	17	.63**	25	.18	59**	40	67**	63**	1 00	-.59**	-48**	.67**	70**	-55
MRHS	-66**	-41	-29	-.19	-.40	-.28**	-14	-12	-53**	-12	-.14	00	-23	-66**	-49**	-27	-13	-.13	-55**	-22	-64**	-08	-18	-57**	-.32	-63**	-49**	-59**	1.0	64**	-78**	-86**	28
NIHSS	-.42**	-19	-06	-08	-.27	-.15	-.13	-10	-24	-11	-.17	.04	-38	-49**	-.48**	-31	-15	-.07	-47**	-.34	-46**	-09	-.15	-39**	-27	-.48**	-35	-48**	.64**	1.00	-.77**	-68**	23
SSS-48	.57**	37	.20	08	.32	.28**	07	08	44**	.08	.10	00	26	70**	63**	32	06	.14	58**	22	.68**	04	.13	.61**	37	61**	43**	67**	-.78**	-.77**	1.0	.86**	-30
Bart Indx	64**	39	23	.17	.40**	.33**	.10	12	50**	.13	11	01	23	65**	53**	28	10	13	55**	21	64**	12	16	58**	30	62**	51**	.70**	-86**	-68**	.86**	1 00	-33
cesd	-37	-32	-32	-39**	-.47**	-.49**	-47**	-76**	-.17	-71**	-30	-25	-37	-45	-45	-65	-43	-30	-67**	-32**	-44	-58**	-36	-37	-.37	-66**	-48**	-55	28	23	-.30	-33	1 00

**. correlation is significant at the 0 01 level (2-tailed), *correlation is significant at the 0 05 level (2-tailed)

Table 12 B: Pearson Correlations for Proxy Baseline Scores

Table 12 C: Correlations (Spearman's rho) for Self-Assessed Baseline Scores

	EQ-MO	EQ-SC	EQ-UA	EQ-PD	EQ-AD
PF	-.63	-.63	-.62	-.07	-.20
RP	-.12	-.19	-.18	-.19	-.08
BP	-.06	.00	-.20	-.55	-.23
GH	-.18	-.17	-.15	-.29	-.16
VT	-.22	-.22	-.28	-.18	-.15
SF	-.31	-.42	-.39	-.15	-.22
RE	.11	.00	-.01	-.11	-.09
MH	.05	-.05	-.01	-.20	-.44
PCS	-.45	-.37	-.47	-.32	-.15
MCS	.11	-.03	.01	-.16	-.24
SAV3	-.03	.04	.07	-.04	-.28
SAH3	-.17	-.08	-.08	-.03	.23
SAS3	-.23	-.14	-.13	.03	-.12
SAA3	-.77	-.76	-.60	-.06	-.03
SAD3	-.65	-.59	-.61	.04	-.08
SAE3	-.13	-.22	-.15	.00	-.37
SAC3	.15	.15	.14	-.01	-.19
SAP3	-.01	-.01	-.20	-.45	-.20
OUS3	-.61	-.59	-.55	-.11	-.15
SAS2	-.26	-.17	-.16	.03	-.08
SAM2	-.74	-.71	-.56	-.01	-.02
SAE2	.15	.07	.03	-.05	-.47
SAC2	.12	.10	.10	-.03	-.27
SAT2	-.63	-.57	-.57	-.09	-.12
SAP2	-.07	-.08	-.22	-.56	-.21
OUS2	-.65	-.59	-.58	-.17	-.16
EQMO	1.00	.73	.67	.07	.08
EQSC	.73	1.00	.67	.10	.12
EQUA	.67	.67	1.00	.11	.16
EQPD	.07	.10	.11	1.00	.12
EQAD	.08	.12	.16	.12	1.00
EQ-VAS	-.39	-.39	-.39	-.21	-.20
EQ-IDX	-.72	-.71	-.69	-.21	-.22
MRH	.71	.64	.57	.13	.04
NIHSS	.54	.52	.46	-.09	.23
SSS	-.71	-.66	-.58	-.05	-.11
BI	-.72	-.72	-.62	-.14	-.09
CESD	.12	.12	.25	.18	.43

Table 12 D: Correlations (Spearman's rho) for Proxy Baseline Scores

	EQ-MO	EQ-SC	EQ-UA	EQ-PD	EQ-AD
PF	-.67	-.71	-.58	-.25	-.21
RP	-.47	-.33	-.43	-.30	-.21
BP	-.26	-.10	-.11	-.58	-.13
GH	-.09	-.09	-.05	-.29	-.26
VT	-.33	-.23	-.29	-.27	-.30
SF	-.37	-.40	-.42	-.18	-.40
RE	-.11	-.11	-.11	-.09	-.34
MH	-.18	-.20	-.22	-.17	-.54
PCS	-.37	-.24	-.22	-.48	.02
MCS	-.17	-.20	-.23	-.12	-.53
SAV3	-.09	-.19	-.14	-.15	-.19
SAH3	-.10	.00	.05	.01	-.12
SAS3	-.20	-.21	-.18	-.11	-.21
SAA3	-.73	-.54	-.63	-.31	-.27
SAD3	-.44	-.50	-.48	-.30	-.17
SAE3	-.39	-.28	-.32	-.26	-.59
SAC3	.03	-.12	-.04	-.02	-.37
SAP3	-.32	-.18	-.24	-.59	-.23
OUS3	-.55	-.53	-.54	-.39	-.49
SAS2	-.13	-.21	-.24	-.02	-.19
SAM2	-.67	-.54	-.61	-.27	-.28
SAE2	-.15	-.14	-.13	-.04	-.61
SAC2	.00	-.15	-.10	-.05	-.40
SAT2	-.52	-.70	-.54	-.14	-.16
SAP2	-.36	-.21	-.24	-.55	-.18
OUS2	-.57	-.57	-.55	-.29	-.44
EQMO	1.00	.71	.65	.29	.25
EQSC	.71	1.00	.70	.19	.24
EQUA	.65	.70	1.00	.22	.26
EQPD	.29	.19	.22	1.00	.23
EQAD	.25	.24	.26	.23	1.00
EQ-VAS	-.51	-.53	-.46	-.23	-.37
EQ-IDX	-.84	-.79	-.78	-.50	-.50
MRH	.66	.60	.59	.11	.13
NIHSS	.52	.49	.46	.09	.19
SSS	-.70	-.69	-.62	-.18	-.19
BI	-.71	-.74	-.63	-.16	-.21
CESD	.38	.37	.34	.27	.64