OUTCOME MEASURES IN ECONOMIC EVALUATIONS
OF RHEUMATOID ARTHRITIS

by

CARLO ARMANDO MARRA

B.Sc. (Pharm.), University of British Columbia, 1992
Pharm.D., University of British Columbia, 1995

A THESIS SUBMITTED IN THE PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY in

THE FACULTY OF GRADUATE STUDIES

Department of Health Care and Epidemiology

We accept this thesis as conforming to the required standard

Aslam H. Anis

John M. Esdaile

Jacek Kopec

Anthony E. Boardman

THE UNIVERSITY OF BRITISH COLUMBIA

May 06, 2004

© Carlo Armando Marra, 2004
The work presented in this thesis was conceived, conducted and disseminated by the doctoral candidate. The co-authors of the manuscripts that comprise part of this dissertation made contributions only as is commensurate with a dissertation committee or as experts in a specific area as it pertains to the work. The co-authors provided direction and support. The co-authors reviewed each manuscript prior to submission for publication and offered critical evaluations; however, the candidate was responsible for the writing and the final content of these manuscripts.

Aslam H. Anis, Ph.D., Chair, Supervisory Committee
ABSTRACT

Objectives: The primary objectives of this study were to: 1) compare the properties of commonly utilized indirect utility assessment instruments (the Health Utilities Index Mark 2 and 3 [HUI2 and HUI3], the EuroQol [EQ-5D], and the Short Form 6-D [SF-6D] in terms of feasibility, reliability, construct validity and longitudinal construct validity (responsiveness) in rheumatoid arthritis (RA); and 2) determine if, when utilized to act as quality weights in the estimation of quality adjusted life years (QALYs) in an economic evaluation, the application of scores from the different instruments would result in different incremental cost per QALY ratios. The primary hypotheses of this study were that there would be differences between these instruments in terms of their properties and that using their scores to estimate QALYs in an economic evaluation of an intervention for RA would result in significantly different estimates.

Methods: Three hundred and twenty patients between 19 and 90 years of age diagnosed with RA residing in the Greater Vancouver Regional District or rural Okanagan region of British Columbia were recruited. Patients were administered a questionnaire containing the HUI2, HUI3, EQ-5D, SF-6D, a disease-specific instrument (the Rheumatoid Arthritis Quality of Life [RAQoL] questionnaire, and a disability index (the Health Assessment Questionnaire [HAQ]). In addition, questions were asked regarding RA management (including drug use and toxicity), RA severity (including swollen and tender joints, pain visual analogue scale, RA duration, patient global assessment of disease activity VAS, and self-perceived RA severity and control), socio-economic status, and RA health utilization. Questionnaires were administered at baseline, three and six months thereafter. In a subset of patients, an
additional questionnaire was administered within five weeks of the three month questionnaire to determine reliability.

Results: Scores obtained with the HUI2, HUI3, EQ-5D, and the SF-6D were significantly different, had low agreement, and appeared to be measuring mostly physical function and pain. All the instruments displayed cross-sectional construct validity and were able to discriminate between different levels of severity of RA. However, when their scores were used to estimate QALYs in an economic evaluation of RA, there was a two fold difference between the lowest (using the HUI3) and highest (using the SF-6D) incremental cost per QALY ratios. Further examination revealed that the scores achieved with the indirect utility assessment instruments were influenced by annual household income despite adjustment for RA severity and other chronic diseases. Finally, in longitudinal analyses, the disease-specific RAQoL displayed the highest reliability and sensitivity to change with the HUI3 and SF-6D scores being the most responsive of the indirect utility assessment instruments in measuring positive change.

Conclusions: Although all indirect utility assessment measures appear to be able to assess generic HRQL in RA, when used as quality weights to estimate QALYs in an economic evaluation, they yielded vastly different estimates of the incremental cost-effectiveness ratio that could result in different policy recommendations. The scores of these instruments could also be influenced by income leading to possible bias in cost-effectiveness analyses. The HUI3 and SF-6D were responsive to positive changes in RA. The RAQoL displayed excellent properties and is a suitable disease-specific HRQL instrument for RA.
TABLE OF CONTENTS

ABSTRACT .......................................................................................................................... II

TABLE OF CONTENTS ........................................................................................................ IV

LIST OF TABLES ................................................................................................................ IX

LIST OF FIGURES ............................................................................................................. XII

CHAPTER 1: INTRODUCTION ........................................................................................... 1

1.1 RHEUMATOID ARTHRITIS: EPIDEMIOLOGY, ECONOMIC BURDEN AND ECONOMIC EVALUATION .............................................................................................................. 1

1.2 RESEARCH NEEDS AND STUDY JUSTIFICATION ..................................................... 8

1.3 STUDY HYPOTHESIS, OBJECTIVES, AND THESIS ORGANIZATION ..................... 11

1.4 SUMMARY .................................................................................................................. 13

1.5 REFERENCES ............................................................................................................. 15

CHAPTER 2: BACKGROUND ............................................................................................. 25

2.1 COST-EFFECTIVENESS ANALYSIS AND THE QALY .............................................. 25

2.2 PREFERENCE-BASED, INDIRECT UTILITY ASSESSMENT MEASURES .................. 31

2.2.1. The Health Utilities Index (HUI) Mark 2 and 3 ..................................................... 31

2.2.2. The EuroQol (EQ-5D) ....................................................................................... 35

2.2.3. The Short Form 6D (SF-6D) .............................................................................. 39

2.3 EMPIRIC COMPARISONS BETWEEN THE INDIRECT UTILITY ASSESSMENT INSTRUMENTS ......................................................................................................................... 40

2.3.1. Comparisons between the Health Utilities Index Mark 2 and Mark 3................. 40

2.3.2. Comparisons across Indirect Utility Assessment Instruments Outside of Musculoskeletal Diseases .................................................................................................................. 44

2.3.3. Comparisons across Indirect Utility Assessment Instruments within Musculoskeletal Diseases .................................................................................................................. 54

2.4 QUALITY WEIGHTINGS IN THE ESTIMATION OF QALYS IN COST-UTILITY ANALYSES IN RA: WHAT ARE INVESTIGATORS USING? ................................................. 58

2.5 SUMMARY .................................................................................................................. 60
2.6 REFERENCES ................................................................................................................. 62

CHAPTER 3: A COMPARISON OF FOUR INDIRECT METHODS OF ASSESSING UTILITY VALUES IN RHEUMATOID ARTHRITIS ................................................................................. 74

3.1 FOREWORD ..................................................................................................................... 74

3.2 INTRODUCTION .............................................................................................................. 74

3.3 METHODS ...................................................................................................................... 76

3.3.1 Measures .................................................................................................................. 77

3.3.2 Data Analysis ......................................................................................................... 78

3.4 RESULTS ....................................................................................................................... 80

3.4.1 Comparison of Utility Scores .................................................................................. 80

3.4.2 Analysis of Agreement ......................................................................................... 82

3.4.3 Exploratory Factor Analysis .................................................................................. 82

3.5 DISCUSSION ................................................................................................................. 83

3.6 REFERENCES .............................................................................................................. 88


4.1 FOREWORD ..................................................................................................................... 106

4.2 INTRODUCTION .............................................................................................................. 106

4.3 METHODS ...................................................................................................................... 110

4.3.1 Sample ..................................................................................................................... 110

4.3.2 Measures ................................................................................................................ 111

4.3.3 Data Analysis ......................................................................................................... 113

4.4 RESULTS ....................................................................................................................... 115

4.4.1 Sample ..................................................................................................................... 115

4.4.2 Description of Global and Single-Attribute Utilities .............................................. 116

4.4.3 Construct Validity .................................................................................................. 117

4.5 DISCUSSION ................................................................................................................. 119
LIST OF TABLES

TABLE 2.1: SOURCE OF PREFERENCES USED FOR QALY WEIGHTS IN ECONOMIC EVALUATIONS OF RA .......................................................... 73

TABLE 3.1: COMPARISON OF THE INDIRECT UTILITY ASSESSMENT INSTRUMENTS .......................................................... 91

TABLE 3.2: CLINICAL CHARACTERISTICS OF THE STUDY PARTICIPANTS .......................................................... 92

TABLE 3.3: OVERALL MEAN AND MEDIAN UTILITY SCORES FROM THE INSTRUMENTS IN THE SAMPLE OF RA PATIENTS .......................................................... 93

TABLE 3.4: INTRACLASS CORRELATIONS AND 95% CONFIDENCE INTERVALS BETWEEN INSTRUMENTS .......................................................... 94

TABLE 3.5: ROTATED FACTOR PATTERN MATRIX .......................................................... 95

TABLE 3.6: FACTOR CORRELATION MATRIX .......................................................... 96

TABLE 3.7: RELATIVE PRATT INDEX SCORES ASSESSING RELATIVE CONTRIBUTION OF EACH FACTOR TO THE MODEL'S ADJUSTED R2 .......................................................... 97

TABLE 4.1: OVERVIEW OF MAUT INSTRUMENT PROPERTIES .......................................................... 127

TABLE 4.2: CHARACTERISTICS OF THE STUDY PARTICIPANTS .......................................................... 128

TABLE 4.3: MULTI-ATTRIBUTE AND SINGLE ATTRIBUTE UTILITY SCORES FROM THE MAUT INSTRUMENTS .......................................................... 129

TABLE 4.4: DOMAIN RESPONSES FOR THE MAUT INSTRUMENTS .......................................................... 130

TABLE 4.5: RELATIONSHIP BETWEEN RA SEVERITY AND CONTROL AND THE GLOBAL UTILITY SCORES FOR EACH OF THE MAUT INSTRUMENTS .......................................................... 132

TABLE 4.6: DICHOTOMOUS MEASURES OF RA SEVERITY .......................................................... 133

TABLE 4.7: CORRELATIONS (SPEARMAN'S RHO) FOR MULTI-ATTRIBUTE AND SELECT SINGLE ATTRIBUTE UTILITY SCORES WITH RA SEVERITY .......................................................... 134

TABLE 4.8: SIMPLE LINEAR REGRESSION ANALYSES FOR OVERALL INSTRUMENT SCORES AND HAQ .......................................................... 135

TABLE 5.1: OBSERVED TRANSITION PROBABILITY MATRICES FOR METHOTREXATE FROM THE ATTRACT TRIAL (FROM WEEK 30 TO WEEK 54) .......................................................... 162
TABLE 7.3: INTRACLASS CORRELATION COEFFICIENT VALUES FOR GENERIC AND DISEASE-
SPECIFIC HRQL MEASURES FOR THOSE REPORTING NO CHANGE IN THEIR RHEUMATOID
ARTHRITIS BETWEEN 0 AND 6 MONTHS ..........................................................252

TABLE 7.4: MINIMALLY IMPORTANT DIFFERENCES REPORTED IN THE LITERATURE AND
DERIVED FROM THE SAMPLE USING ANCHOR-BASE APPROACHES ..................253

TABLE 7.5: CORRELATIONS BETWEEN THE TRANSITION QUESTION AND CHANGES IN
RHEUMATOID ARTHRITIS OUTCOME VARIABLES FROM 0 TO 6 MONTHS .................254

TABLE 7.6: DIFFERENCES AND RESPONSIVENESS STATISTICS FROM BASELINE TO 6 MONTHS
STRATIFYING THE SAMPLE BY THE TRANSITION QUESTION ...........................255

TABLE 7.7: DIFFERENCES AND RESPONSIVENESS STATISTICS FROM BASELINE TO 6 MONTHS
STRATIFYING THE CATEGORIES CREATED FROM PATIENT GLOBAL ASSESSMENT OF
DISEASE SEVERITY VAS ..................................................................................256

TABLE 7.8: RANKINGS OF RESPONSIVENESS OF MEASURES ACCORDING TO THE
RESPONSIVENESS STATISTIC AND THE EXTERNAL CRITERIA OF CHANGE (EITHER
RESPONSES TO THE PATIENT TRANSITION QUESTION OR TO THE PATIENT GLOBAL
ASSESSMENT OF DISEASE ACTIVITY VAS) .......................................................257

TABLE 7.9: ASSOCIATIONS BETWEEN INSTRUMENT UNWEIGHTED DOMAINS / SINGLE
ATTRIBUTE SCORE CHANGES AND SELF-REPORTED CHANGE FROM 0 TO 6 MONTHS.....258
LIST OF FIGURES

FIGURE 3.1: DISTRIBUTIONS OF GLOBAL UTILITY VALUES ACROSS THE MAUT INSTRUMENTS
................................................................................................................................. 98
FIGURE 3.2: BLAND-ALTMAN PLOT OF DIFFERENCE BETWEEN THE HUI2 AND HUI3 VS. THE
AVERAGE SCORE WITHIN PATIENTS ........................................................................... 100
FIGURE 3.3: BLAND-ALTMAN PLOT OF DIFFERENCE BETWEEN THE HUI3 AND THE SF-6D VS. THE
AVERAGE SCORE WITHIN PATIENTS ........................................................................... 101
FIGURE 3.4: BLAND-ALTMAN PLOT OF DIFFERENCE BETWEEN THE HUI3 AND EQ-5D VS. THE
AVERAGE SCORE OF THESE TWO INSTRUMENTS WITHIN PATIENTS ................. 102
FIGURE 3.5: BLAND-ALTMAN PLOT OF THE DIFFERENCE BETWEEN THE EQ-5D AND SF-6D VS.
THE AVERAGE SCORE WITHIN PATIENTS ............................................................... 103
FIGURE 3.6: BLAND-ALTMAN PLOT OF DIFFERENCE BETWEEN THE HUI2 AND THE SF-6D VS. THE
AVERAGE SCORE WITHIN PATIENTS ................................................................. 104
FIGURE 3.7: BLAND-ALTMAN PLOT OF DIFFERENCE BETWEEN THE HUI2 AND EQ-5D VS. THE
AVERAGE SCORE OF THESE TWO INSTRUMENTS WITHIN PATIENTS ............. 105
FIGURE 4.1: BOX PLOT OF MAUT INSTRUMENT GLOBAL UTILITY SCORES ............. 136
FIGURE 5.1: A SCHEMATIC REPRESENTATION OF THE MARKOV, HAQ-BASED MODEL USED FOR
THE COST-EFFECTIVENESS ANALYSIS ................................................................. 171
FIGURE 5.2: KAPLAN-MEIER SURVIVAL CURVES FROM THE 100,000 MONTE CARLO
SIMULATIONS ........................................................................................................... 172
FIGURE 5.3: INCREMENTAL COSTS AND QALYS FROM 1000 2ND ORDER MONTE CARLO
SIMULATIONS ........................................................................................................... 173
FIGURE 5.4: COST-UTILITY ACCEPTABILITY CURVES FOR EACH INDIRECT UTILITY MEASURE
.................................................................................................................................. 174
FIGURE 6.1: GENERIC HRQL BY SELF-REPORTED ANNUAL INCOME (HUI3 AND SF-6D) ....215
FIGURE 6.2: GENERIC HRQL BY SELF-REPORTED ANNUAL INCOME (HUI2 AND EQ-5D) ......216
FIGURE 6.3: RAQOL SCORE AND HAQ DISABILITY INDEX BY SELF-REPORTED INCOME ....217
FIGURE 7.1: AGREEMENT BETWEEN THE PATIENT TRANSITION QUESTION AND CHANGES USING MID CUTOFFS FOR THE GENERIC AND DISEASE-SPECIFIC INSTRUMENTS ...........259
FIGURE 7.2: SCATTERPLOT OF HUI2 UTILITY SCORES OVER TIME STRATIFIED BY THE RESULTS OF THE COLLAPSED TRANSITION QUESTION.................................................................262
FIGURE 7.3: SCATTERPLOT OF HUI3 UTILITY SCORES OVER TIME STRATIFIED BY THE RESULTS OF THE COLLAPSED TRANSITION QUESTION.................................................................263
FIGURE 7.4: SCATTERPLOT OF EQ-5D UTILITY SCORES OVER TIME STRATIFIED BY THE RESULTS OF THE COLLAPSED TRANSITION QUESTION.................................................................264
FIGURE 7.5: SCATTERPLOT OF SF-6D UTILITY SCORES OVER TIME STRATIFIED BY THE RESULTS OF THE COLLAPSED TRANSITION QUESTION.................................................................265
FIGURE 7.6: RESULTS OF THE MULTI-RESPONSE MODEL OF THE ASSOCIATION BETWEEN A CHANGE IN THE HUI2 AND THE TRANSITION QUESTION .............................................................266

FIGURE 7.15: RESULTS OF THE MULTI-RESPONSE MODEL OF THE ASSOCIATION BETWEEN A CHANGE IN THE SF-6D AND THE PATIENT GLOBAL ASSESSMENT OF DISEASE ACTIVITY


ACKNOWLEDGEMENTS

A large debt of gratitude is owed to my co-supervisors, Aslam Anis and John Esdaile, for providing me with invaluable mentorship and support. From the outset, they provided sound advice, amazing opportunities and lots of good ideas. I would have been lost without them.

Most sincere thanks to my committee members: Stephen Marion who shared so unselfishly of his time imparting a small fraction of his immense knowledge to me and Jacek Kopec for mentorship and counsel that has benefited me greatly and improved the quality of my work. The support of dedicated research assistants made this work possible: Barbara Vinduska, Amir Adel Rashidi, and Janet Pursell. Also, much thanks to the rheumatologists who facilitated recruitment: Robert Offer, Andrew Chalmers, Kamran Shojania, Barry Koehler, Graham Reid, Dan Macleod, Alice Klinkhoff, John Kelsall, Milton Baker, and Diane Lacaille. Best wishes and heartfelt thanks to all of the study participants. Special thanks to all of the people who shared their time and knowledge with me along the way especially: Daphne Guh, Chris Richardson, and John Woolcott. Also, Larry Lynd, who forged the path ahead of me, motivated me to embark on this career path, and left very large footprints that were difficult to fill.

Most of all, I want thank my amazing family; especially, my wife Fawziah for pushing me to follow my goals and for providing me with unyielding support and advice from the beginning that gave me the necessary strength and motivation to begin and complete this work. Without her many compromises, dedication and commitment, none of this would have been possible. To her, I dedicate this thesis. Finally, Yasmin and Noah, thank you both for the constant reminders about the things that are truly important in life and keeping me grounded in reality.

This research was generously supported by grants from the Canadian Arthritis Network. Thanks to the Canadian Institutes of Health Research, the Arthritis Society and the Michael Smith Foundation for Health Research for their fellowship support.
CHAPTER 1

INTRODUCTION

1.1 RHEUMATOID ARTHRITIS: EPIDEMIOLOGY, ECONOMIC BURDEN
AND ECONOMIC EVALUATION

Rheumatoid arthritis (RA) is a chronic, progressive, inflammatory disease that afflicts approximately 300,000 Canadians. The cause of this condition is unknown and there is no cure. This disease affects the physical functioning of patients as well as their psychological and social health and eventually progresses to substantial disability through the loss of mobility, increased morbidity, and premature mortality. The incidence of death from cardiovascular disease, infection and cancer are significantly higher than those experienced in the general population.

RA can occur at any age but its onset peaks between the ages of 40 and 60 years. The prevalence of RA is approximately 0.5 to 1% of the adult population and the incidence appears to be decreasing over the past 4 decades (from 61.2 per 100,000 in 1955 to 1964 to 32.7/100,000 in 1985 to 1994). However, this finding is from one population-based, longitudinal study in a specific geographic area (Rochester, Minnesota) and may not be generalizable to areas with different ethnicity (this sample was 96% white) or environmental exposures. Other prevalence data is difficult to find and is often not population-based. Using administrative databases in British Columbia, a population-based estimate of 27,710
RA cases (mean (SD) age of 64.1(17), 67% women) was identified translating into a prevalence rate of 0.76%.\textsuperscript{16} Although the incidence of RA may be decreasing, results from epidemiological evaluations show that the premature mortality associated with this condition has not changed over the last several decades despite the development of more effective interventions.\textsuperscript{2,17,18} However, a recent analysis has determined that the use of methotrexate (MTX) is associated with a substantial survival benefit in patients with RA despite having had worse prognostic factors for mortality prior to being treated with this agent.\textsuperscript{9} After adjusting for confounding by indication (specifically, in this case, patients with more severe disease having a higher probability of being prescribed MTX), this beneficial affect on mortality was demonstrated in comparisons with patients who were taking other disease modifying antirheumatic drugs (DMARDs) or patients taking no DMARDs (the adjusted hazard ratio of MTX compared to no DMARDs was 0.2, 95% confidence interval (CI) 0.1 to 0.7). The mortality hazard ratio for comparisons between those using MTX and those with no MTX use (i.e. other DMARDs) was 0.4 (95% CI 0.2 to 0.8) for all cause mortality, 0.3 (95% CI 0.2 to 0.7) for cardiovascular mortality, and 0.6 (95% CI 0.2 to 1.2) for non-cardiovascular mortality. Thus, from the results of this analysis, it would appear that the application of an effective treatment such as MTX is associated with survival benefits.

In another study examining survival in RA,\textsuperscript{7} Wolfe et al. demonstrated that the strongest predictor of mortality in this disease group was longitudinal changes in the Health Assessment Questionnaire (HAQ). A one standard deviation increase in the HAQ (a higher HAQ represents more severe disease) resulted in a 26.2% greater increase in the odds ratio for mortality compared to the next most powerful predictor of mortality, the patient
completed global severity index. Interestingly, changes from the fourth quartile to the first quartile for both of these measures would have an estimated reduction in mortality by 50% for the HAQ and 33% for the patient completed global severity index. Thus, since the HAQ and other self-reported measures are the most highly predictive of future mortality, lowering the HAQ with effective drug therapy (as shown by Wolfe et al.\(^9\)) or improving self-rated health should result in improved survival in RA. There are many other examples in the literature of the effect of drug therapy on the HAQ in patients with RA; however, these patients have not been followed for a long enough period of time to determine if this reduction in HAQ translates into a reduction in mortality.\(^{19-23}\) Thus, one can only postulate whether the reduction in HAQ due to drug treatment (other than MTX) observed in these trials results in a lower mortality rate.

The severity of RA has been shown to be significantly related to a reduction in health-related quality of life (HRQL)\(^{24,25}\) and HRQL in RA has been shown to be worse than in other forms of arthritis.\(^{26}\) The finding that lower HRQL is associated with higher RA severity has been shown using both disease-specific measures such as the Rheumatoid Arthritis Quality of Life Questionnaire (RAQoL)\(^{27,28}\) and the Arthritis Impact Management Scale (AIMS),\(^{29,30}\) and, to a lesser extent, preference-based measures of HRQL such as the EuroQol (EQ-5D).\(^{31}\)

As with the humanistic burden that RA exerts in terms of mortality and reduction in HRQL, the economic burden of RA to society is substantial and is thought to rival that of coronary artery disease.\(^{32}\) Since RA usually starts in the 4\(^{th}\) or 5\(^{th}\) decade of life\(^{33}\) and the disease cannot be cured, the costs attributable to this condition are often compounded over several years. There are a number of studies in the literature which attempt to examine the
direct and indirect costs associated with RA from a variety of perspectives.\textsuperscript{34-46} The findings of these analyses show a great degree of variability in the estimated costs, partly due to differences in the costing methodologies employed\textsuperscript{47} (and some utilized charges instead of costs), which variables were included in the ascertainment of direct medical costs, and methodological problems in the determination of cost-of-illness in RA.\textsuperscript{48} Despite these limitations, each study concluded that the costs to manage RA were substantial, and when assessed, the indirect costs were a large portion of the overall costs to manage RA.

Several reviews examining studies that investigate the costs to manage RA have been published.\textsuperscript{48-50} Most of these studies included in these reviews were conducted in an era when new, expensive drug therapies for RA, such as the tumour necrosis factor (TNF) alpha blockers and newer non-steroidal anti-inflammatory drugs, were not yet available. Specifically, in a review by Lubeck,\textsuperscript{49} hospitalizations generally accounted for \( \geq 60\% \) whereas drug costs generally accounted <25\% of direct costs in RA. Pugner et al.\textsuperscript{50} found that the mean annual direct cost to manage RA was $5,425 (1998 U.S. dollars) and that the median percentage of this total due to hospitalization and drugs was 47\% and 16\%, respectively. Few studies have attempted to quantify indirect costs but, of those that have,\textsuperscript{36,39,40,42,43} these have ranged from $1082 to $37501 (1996 U.S. dollars) per patient. However, again in the determination of indirect costs, there was a lack of clarity in the studies on how the results were determined and/or important methodological issues that makes comparisons across the studies difficult.

The only Canadian study that was published in this earlier era of RA drug treatment was by Clarke et al. who examined cohorts of individuals with rheumatoid arthritis in Saskatchewan.\textsuperscript{42,43} In this longitudinal study of almost 1000 patients with RA, annual direct
and indirect costs were determined to be $4656 and $1597, respectively (1994 Canadian dollars). The authors determined resource utilization and direct medical costs from assessing the number of physician visits, medications, diagnostic tests, and inpatient care. Inpatient care was associated with almost two-thirds of the direct medical costs. Indirect costs were assessed through productivity loss using the human capital approach. Since most of the patients were > 60 years of age and no longer considered themselves to be in the work force, the authors did not include these individuals in the calculation of indirect costs resulting in small estimates.

However, the results of these studies are no longer accurate because of the introduction and increasing use of biological drug therapy (infliximab, etanercept and adalimumab) as well as a new class of nonsteroidal anti-inflammatory drugs (the cyclooxygenase [COX] 2 specific inhibitors)\textsuperscript{51}\textsuperscript{a} in the past few years which have caused medication costs to manage RA to skyrocket.\textsuperscript{35} These biological agents are effective but are extremely expensive with annual acquisition costs from $12,000 to over $20,000 per patient. A recent analysis examining costs in the biological era was published by Michaud et al..\textsuperscript{35} In a sample of 7,527 patients with RA answering semi-annual questionnaires from January 1999 to December 2001, direct medical costs (calculated from physician and other health professional visits, radiologic examinations, laboratory and other tests, outpatient surgeries, hospitalizations and medications) were determined. In the entire sample, the mean direct cost was $9,519 (2001 U.S. dollars) of which 66% was due to drug costs and 16% and 17% were due to hospital costs and outpatient costs, respectively. In those receiving biological

\textsuperscript{a} Reference 51 was authored by the candidate during the tenure of the doctoral program and has been inserted as Appendix I
agents, the annual direct cost was $19,016 compared to $6,164 (2001 U.S. dollars) in those not receiving these agents.

The use of these agents may have had an impact on indirect costs as measured by productivity losses as well. In a study by Yelin et al., the association between the use of etanercept and employment outcomes were investigated in a sample of 497 RA patients. In order to ensure eligibility for employment, only patients between 18 and 64 were included in the study. In structured telephone interviews, patients were asked questions regarding their employment status in the year of diagnosis (75% in the etanercept group vs. 77% in the non- etanercept group) and in the study year (71% in the etanercept group vs. 55% in the non-etanercept group). After adjusting for demographics, overall health status, duration of RA, RA status, and occupation type, the difference increased to 20%, 95% CI 9% to 32% difference (53% vs. 73% employed in the non-etanercept and etanercept groups, respectively). Thus, it would appear that, at least among those of working age, etanercept has the potential to reduce the indirect costs associated with RA.

With respect to the impact that these new biological agents make on HRQL and HAQ scores, a few studies have shown benefit. Using the Short Form 36 (SF-36), both infliximab and etanercept have been shown to improve HRQL (at least in the short term) over MTX in randomized controlled trials. In a recently published observational study of patients either being treated with infliximab or with stable RA, the responsiveness of the SF-36, the EQ-5D, the standard gamble, and the Short Form 6-D (SF-6D) were evaluated. In the group treated with infliximab, large responsiveness indices (>0.80 effect sizes) were observed for the SF-36 physical component score, and relevant domains in the SF-36 (bodily pain, physical functioning, role physical, social functioning and vitality). For preference-
based measures, the SF-6D was highly responsive (effect size of 1.40), the EQ-5D was moderately responsive (effect size of 0.67) and the standard gamble was poorly responsive (effect size of 0.49). However, the sample size of this study was small (60 patients on infliximab and 24 patients with stable RA) so the results were not conclusive.

Although economic evaluations of pharmacotherapies are not new in RA, there has been an explosion of published cost-effectiveness analyses since the introduction of leflunomide (a new DMARD), new biological agents, pharmacogenetic technologies and COX-2 specific inhibitors. Although there are many shortcomings of several of these analyses, a detailed discussion of these is beyond the scope of this chapter. A critical review pertaining to the cost-effectiveness literature of biological agents will soon be available. However, a couple of limitations of these analyses have direct relevance to this thesis - namely, the use of randomized controlled data to estimate outcomes and the attempt (or lack thereof) to incorporate HRQL data into the outcome variables. Wolfe et al. make a compelling case that the short-term efficacy data derived from randomized controlled trials are not suitable to extrapolate to long-term cost-effectiveness results and that observational drug-treatment databases should be utilized. This finding is based on the evidence that treatment outcomes derived from observational databases can often be different than those derived from randomized controlled trials in RA. The second major limitation of many of

---

b Of note, the candidate authored one of the cost-effectiveness analyses on the pharmacogenetics technologies (reference 66) during his tenure as a doctoral student and it has been included as Appendix II.

c Of note, the candidate authored a paper (reference 73) during the tenure of his doctoral program that showed that the efficacy and toxicity of cyclosporin in an observational database was different than reported in randomized controlled trials and this has been included as Appendix III.
these cost-effectiveness analyses is the lack of an attempt to integrate preference-based generic HRQL measures into the economic evaluations or the application of instruments/techniques that have not undergone appropriate testing in RA. For example, many of the analyses report either costs alone or cost-effectiveness ratios using naturalistic units (such as reduction in swollen joints or in proportions of patients improving using standard criteria). Other analyses utilized direct preference elicitation techniques such as a visual analogue scale (VAS), Time Trade Off (TTO) or standard gamble (SG) although the TTO/SG have been shown to be poorly responsive and poorly correlated with clinical outcomes in patients with rheumatoid arthritis. The most commonly applied method to obtain preference scores was the EQ-5D in the recent cost-effectiveness analyses of biological agents which has been shown to be both responsive and valid in rheumatoid arthritis.

Other instruments utilized to derive preference-based scores that are commonly utilized in cost-utility analyses have not yet been applied in assessing the cost-effectiveness of treatments for rheumatoid arthritis. These instruments include the Health Utilities Index Mark 2 (HUI2) and Mark 3 (HUI3), and the SF-6D. However, research in RA and other disease states suggests that scores obtained with these systems are not interchangeable and could have a profound impact on the estimation of incremental cost-effectiveness ratios.

1.2 RESEARCH NEEDS AND STUDY JUSTIFICATION

The appropriate and most efficient use of health care resources has resulted in the need to conduct economic evaluations for new and existing treatments in order to inform
decision-making. For diseases such as RA that are chronic and incurable with a documented impact on HRQL, the need to integrate HRQL data into treatment assessment is critical. Compounding this point is the evolving field of RA treatment, which has brought about several new, effective but very expensive agents in the past few years. Thus, through the cost-utility analysis framework, preference-based measures of HRQL can be used to inform resource allocation decisions in health care. This is done through the calculation of the quality adjusted life year (QALY) which is commonly used in the denominator of the incremental cost-utility ratio calculation.

As originally described by Weinstein et al., “the quality adjusted life year approach assigns to each period of time a weight, ranging from 0 to 1, corresponding to the health-related quality of life during that period, where a weight of 1 corresponds to optimal health, and a weight of 0 corresponds to a health state judged to be equivalent to death”. Thus, QALYs relating to a health outcome are expressed as the value (weighting) given to a particular health state multiplied by the time spent in that state. The weightings used in the calculation of QALYs are derived from preferences for health states, which can be measured directly through the application of various methodologies such as the standard gamble (SG), time trade off (TTO), person trade-off, and rating scales. However, due to the expense and inconvenience associated with administering many of the direct approaches, generic, preference-based questionnaires have been developed which integrate health into a single index (where death is anchored at zero and perfect health at one). These questionnaires typically consist of a health classification system with an associated scoring formula that assigns preference-weighted values to the health states defined by the classification system and integrates the different aspects of health into a single index. The questionnaires that are
most commonly utilized are the Health Utilities Index Mark 2 and 3 (HUI2 and HUI3), the EuroQol (EQ-5) and the Short Form 6D (SF-6D).\textsuperscript{79,80} However, despite their widespread use, there have been few comparative studies addressing their strengths, weaknesses and interchangeability.\textsuperscript{80}

Empirical assessments of instruments designed to measure or value HRQL involves examining feasibility, reliability, validity and responsiveness.\textsuperscript{83} Feasibility refers to the ability of the instrument to be used in practice and accepted by respondents.\textsuperscript{81} Reliability refers to stability of responses if the conditions under examination remain unchanged.\textsuperscript{83} Validity is the extent to which an instrument measures the property it is intended to measure.\textsuperscript{84} Much of the literature on validity of HRQL measures focuses on the discriminative properties of instruments which is also the technique commonly employed for preference-based instruments.\textsuperscript{85} Another essential property of HRQL instruments is the ability to detect change over time and, the extent to which this change is important or meaningful.\textsuperscript{86} Finally, it is not clear if these instruments, which all purport to assess the same construct, namely a single index score of HRQL, are interchangeable and, if used as the weights for QALYs in an economic evaluation of pharmacotherapy for RA, would result in comparable outcomes and potentially similar policy decisions.

This study was conceived based on the need to compare the properties of the most commonly utilized indirect, preference-based measures in terms of their cross-sectional construct validity, differences in aspects of health that they assess, potential biases in their scores in terms of the effects of income, and their sensitivity to change and responsiveness. Chapter 2 provides a detailed review of these preference-based, indirect utility instruments,
studies that compare their properties, and the use of preference-based measures as QALY weights in economic evaluations of RA.

1.3 STUDY HYPOTHESIS, OBJECTIVES, AND THESIS ORGANIZATION

The overall aim of this study was to evaluate and compare the different properties of the four indirect utility instruments (HUI2, HUI3, SF-6D, EQ-5D) and to assess whether using the scores generated by their different systems in the same cost-effectiveness framework would result in different outcomes. The primary hypothesis of this study was that quality adjusted life year (QALY) estimates obtained using these instruments would be different and would result in different incremental cost-utility ratios and, therefore, potentially different policy decisions.

The first objective of this study was to determine if, on a cross-sectional basis, the indirect utility instruments would yield similar utility values in patients with RA and, if not, were the assessed domains of health similar among the instruments.

The second objective was to determine if these indirect utility assessment instruments displayed cross-sectional, construct validity in the assessment of patients with RA and how well they compared, in this regard, to the disease-specific RAQoL and to a disability status measure, the HAQ.

The third objective was to determine if the utilization of the different utility values generated by the indirect utility instruments in a cost-utility analysis of a new drug therapy (infliximab plus MTX) compared to usual therapy (MTX alone) for rheumatoid arthritis would result in different estimates of the incremental cost per QALY gained.
The fourth objective was to determine if the results generated by the indirect utility instruments are influenced by socioeconomic status and, if so, could therefore bias the results of cost-utility analyses.

The fifth objective was to determine the longitudinal validity of the instruments in rheumatoid arthritis in terms of their ability to be responsive to changes that patients experience in their RA.

This thesis is comprised of eight chapters, organized chronologically, addressing each of the objectives in order. This first chapter provides a brief introduction to: 1) the epidemiology of RA; 2) the effect of RA on mortality; 3) direct and indirect costs of RA; 4) the rapidly evolving field of pharmacotherapy for RA; 5) the impact on new strategies on work productivity and HRQL; 6) published cost-effectiveness analyses in RA; and 7) the use of preference-based measure scores as weighting factors for QALYs in economic evaluations of RA.

Chapter 2 provides a detailed literature review of indirect utility instruments as weightings for QALYs in the economic evaluation of interventions for RA. Specifically, the indirect utility measures, their properties, their use in the calculation of quality adjusted life years, how their properties have compared in other disease states, and the application of these instruments in cost-utility analyses in RA are reviewed in detail. Chapters 3 and 4 present the results of the cross-sectional analysis from the baseline results of a sample of RA patients who participated in our longitudinal study. Chapter 5 provides a comparison of how the application of these utility instruments in a decision-analytic, Markov model for a new pharmacotherapy in RA results in vastly different incremental cost per QALY ratios. Chapter 6 provides an examination of how these indirect utility instruments are influenced by annual
income and how this could potentially bias economic evaluation of therapies for RA. Chapter 7 presents the longitudinal validity analyses of these instruments in terms of responsiveness. Chapters 3 through 7, and Appendix I, II, III are each stand-alone manuscripts, which have either been published, are in press, or are under review by a major, peer-reviewed, scholarly journal. The work presented in this thesis was conceived, conducted, and disseminated by the doctoral candidate as has been declared by the co-supervisors of the candidate (Appendix IV).

The final chapter provides a summary of the research findings and outlines the strengths, limitations and the unique contributions and potential impact of the findings of this study.

1.4 SUMMARY

Approximately 300,000 Canadians have been diagnosed with RA. Due to its chronic, debilitating nature, the direct costs associated with the management of this condition and indirect costs due to lost employment are substantial. Functional status and HRQL have been shown to be reduced in patients with RA. New therapies, specifically biological DMARDs, have the potential to improve HRQL, functional status, and offset some of the indirect costs associated with productivity and potentially some of the direct costs of management (such as hospitalizations) although their acquisition costs are large. Therefore, cost-effectiveness analyses of these new agents that integrate appropriate preference-based measures of HRQL into years of life are required. Often, due to the convenience and availability, instruments that estimate society’s preferences for health states are used to accomplish this task. However, preliminary evidence suggests that the application of these instruments in
economic evaluations could result in vastly different cost-effectiveness outcomes. Therefore, further research is required to compare the scores and properties of these instruments in patients with RA.

This study focused on the comparison of the scores and properties of four indirect utility assessment instruments (the HUI2, HUI3, the EQ-5D, and the SF-6D) in patients with RA. The evaluation of these instruments in this population required recruitment of a sample of patients with RA for direct comparison of scores and the evaluation of their longitudinal properties. In addition, a HAQ-based, Markov model was created to test the hypothesis that incremental cost per QALY ratios would be different using the various indirect utility instrument scores as weightings for QALYs.
1.5 REFERENCES


36. Meenan RF, Yelin EH, Heke CJ, et al.. The costs of rheumatoid arthritis: A patient- 

37. Lubeck DP, Spitz PW, Fries JF, et al.. A multicentre study of annual health service 

38. Yelin E, Wanke LA. An assessment of the annual and long-term direct costs of 


40. Gabriel SE, Crowson CS, Campion ME, et al.. Direct medical costs unique to people 

41. Yelin E. The costs of rheumatoid arthritis – absolute, incremental, and marginal 

42. Clarke AE, Zowall H, Levinton C et al.. Direct and indirect medical costs incurred by 
Canadian patients with rheumatoid arthritis: A 12 year study. J Rheumatol 
1997;24:1051-1060.

short term direct medical costs incurred by patients with rheumatoid arthritis. J 
Rheumatol 1999; 26:1068-1075.

44. Liang MH, Larson M, Thompson M, et al.. Costs and outcomes in rheumatoid 

45. van Jaarvsveld CHM, Jacobs JWG, Schrijvers AJP, et al.. Direct cost of rheumatoid 
arthritis during the first six years: A cost-of-illness study. Br J Rheumatol 


69. Lee KK, You JH, Ho JT, Suen BY, Yung MY, Lau WH, Lee WV, Sung JY, Chan FK. Economic analysis of celecoxib versus diclofenac plus omeprazole for the


2.1 COST-EFFECTIVENESS ANALYSIS AND THE QALY

In recent years, cost-effectiveness analysis has emerged as the preferred technique for economic evaluation in health care. Cost-effectiveness analysis shows the relationship between the incremental net resources used (costs) and the net health benefits generated (effects) between a specific intervention and an alternative strategy. As such, the incremental cost-effectiveness ratio (ICER) can be calculated, which is simply the ratio of the difference between two interventions' costs and the difference between their effectiveness as follows:

$$ICER = \frac{\Delta Cost}{\Delta Effect}$$

Rather than express the outcomes in cost-effectiveness analysis in terms of naturalistic units (such as number of tender joints reduced), analysts have sought outcome measures that permit comparisons across conditions. This framework would inform societal decision-making such that competing interventions that produce the greatest gain in health for the resources expended could be identified. One potential way to permit cross-indication comparisons (i.e. comparisons of cost-effectiveness across disease states) would be to utilize life-expectancy as the measured outcome. However, this approach would not consider the health-related quality of life associated with various interventions and would bias funding decisions against those interventions imparting mainly HRQL while favoring only those
interventions that result in improvements in survival. As such, diseases such as rheumatoid arthritis (RA), where improvements in survival due to interventions are small (when compared to cancer therapy or HIV pharmacotherapy) but improvements in HRQL are paramount, would be hard-pressed to compete for scarce health resources. Therefore, an outcome measure that integrates both years of life and HRQL into a single metric provides a solution to this problem.\textsuperscript{1,6}

The use of quality adjusted life years (QALYs) is an attempted solution to incorporate both potential life prolongation and improvement in HRQL.\textsuperscript{1} Neumann et al. stated that "QALYs represent the benefit of a health intervention in terms of time in a series of quality-weighted health states, in which the quality weights reflect the desirability of living in the state, typically from perfect health (weighted 1.0) to dead (weighted 0.0)." Therefore, once the quality weights are obtained for each health state experienced by an individual, they are multiplied by the duration of time spent in the health state. The products of these calculations are then summed to obtain the total number of QALYs for that person in the following manner:

\[
\text{Total QALYs} = \sum_{i=1}^{T} u_i(q_i) D_i
\]

Where: \( u_i(q_i) \) = the quality of life in period \( i \) (measured by utilities);

\( t = \) the time interval of period in terms of years;

\( D_i = \) discount factor of period \( i \)

As such, the incremental cost-effectiveness ratio becomes:

\[
\text{ICER} = \frac{\Delta \text{Cost}}{\Delta \text{QALY}}
\]
The QALY approach is not without controversy and other competing methods have been suggested such as the Healthy-Years Equivalents (HYEs), disability adjusted life years (DALYs), and saved young life equivalents (SAVE).\textsuperscript{1,2,6} Each of these techniques has their relative advantages and disadvantages which are beyond the scope of this chapter; nonetheless, QALYs have remained the outcome measure of choice in the health economic literature.\textsuperscript{1,6} However, the QALY approach has several assumptions on which it is based which include, but are not limited to: 1) utility independence; 2) constant proportionality trade-off; and 3) risk attitude over life years.\textsuperscript{1,6}

However, assuming that the QALY approach is correct and assumptions are met, there is still the issue of what should be utilized as the source of weights for the health states in the QALY calculation. Certain conditions for these weights must be met which include: 1) that they be based on preferences for health states; 2) that they be measured on an interval scale; and 3) that they be anchored at perfect health (1.0) and death (0.0).\textsuperscript{2} For the latter anchoring requirement, health states can be valued to be worse than death and have negative weights associated with them.\textsuperscript{1,2} Also, the terminology used to describe these weights can be problematic as researchers have used the terms “utility”, “value” and “preference” interchangeably.\textsuperscript{1} However, as Drummond et al.\textsuperscript{2} describe, the term “utility” is reserved for preferences that are measured under conditions of uncertainty that satisfy the axioms of expected utility theory (the standard gamble [SG]). “Values” are preferences that are measured under conditions of certainty and thus include rating scale and time-tradeoff elicited scores, whereas “preferences” encompass both and is a general term to describe the desirability of a set of outcomes.
Sources for weights which meet the aforementioned assumptions come from both directly elicited techniques and indirectly elicited techniques.\(^1\)\(^2\)\(^7\) Although an in-depth discussion of techniques to directly elicit preference values to be used as weightings for QALYs is beyond the scope of this chapter, a brief description is provided.

Directly elicited techniques encompass the methods to elicit values and utilities as described above (namely the SG, the TTO, and the RS). Within this framework, it is recommended to use choice-based techniques (SG or TTO) over scaling methods (RS).\(^1\)\(^2\) The SG is grounded in von Neumann Morgenstern expected utility theory (EUT) and, in its usual form, asks respondents to choose between a particular, intermediate health state with certainty and a gamble involving a probability of a better or worse outcome than the certain outcome.\(^6\)\(^7\) The goal with the SG approach is to find the probability \(p\) in the gamble at which the respondent is indifferent between the certain and uncertain alternatives.\(^2\) Although long considered to be the preferred method due to its strong ties to EUT, a recent qualitative study investigated what thought processes respondents invoke in formulating their SG responses and found that some respondents were incorporating inappropriate information into their choices.\(^8\) In addition, Dolan argues that since there is evidence that people violate the assumptions of EUT, much of the appeal of the SG is lost.\(^6\)

The TTO technique involves asking a respondent to make tradeoffs between a shorter life span in perfect health versus a longer life span in the health state in question. The time in full health is varied until the respondent is indifferent between the two alternatives.\(^1\)\(^2\) The TTO choice is not made under uncertainly so the values that it elicits are not considered utilities, at least under EUT.\(^2\)\(^6\) Recently, a study appeared that has cast doubt on the ability
to use the TTO as weighting for QALYs due to the violation of the constant proportional
time trade-off assumption.9

Finally, the RS technique asks respondents to indicate ratings for health states (or
their own health state) on a scale (usually a vertical or horizontal line) with endpoints of
"worst" and "best" health states usually represented by 0 and 100. To allow for the
possibility for health states worse than death, the line is often anchored at the "worst" and
"best" imaginable health states.6,7

In comparing directly elicited preference scores, application of the various techniques
result in different preference weights.1 The SG approach almost always generates scores that
are higher than the TTO method, and both are usually greater than RS scores.1,2,7 Since
people are risk averse, they are less willing to accept the gamble outcome presented in the
SG and more willing to accept the certainty. As well, since people have positive time
preference and value years of life in the near future more than years of life in the distant
future, they would be more willing to give up years of life at the end of a profile as in the
TTO.6 Both of these assumptions would lead to higher SG scores than TTO values.

Indirect preference or utility assessment techniques involve the use of generic health
classification systems in the form of a questionnaire.1,2 Through completion of the health
classification system, respondents are assigned a health state which, in turn, is valued using a
scoring function that applies preference weights from another population (i.e. society). Due
to their relative ease and low cost to administer when compared to the SG or TTO
techniques, these questionnaires are widely applied.1,2 These instruments commonly utilize
multi-attribute utility theory (MAUT) to combine many attributes into a single utility value.2
As it is beyond the scope of the chapter to describe MAUT in detail, the reader is referred to reviews for a complete description and assumptions involved with this theory.\textsuperscript{1,2}

The most common examples of these questionnaires include the Health Utilities Index Mark 2 and Mark 3 (HUI2 and HUI3), the EuroQol (EQ-5D) and the Short-Form 6D (SF-6D). Each of these systems assesses different domains of health and relies on different scoring functions/methods to determine preference scores.\textsuperscript{10} Other preference-based measures that have been less commonly applied are the Quality of Well-Being scale, the Finnish 15-D, and the Assessment of Quality of Life (AQoL).\textsuperscript{2,10}

In the Canadian Coordinating Office of Health Technology Assessment’s Guidelines for the Economic Evaluation of Pharmaceuticals,\textsuperscript{11} it is stated that due to the lack of head-to-head comparisons of these systems, there is little information to guide users in the choice of instrument. As such, it advises users “to study the alternative systems, to select in advance the one that best suits the study objectives, to justify the selection in the study protocol, and to stick with it. It is not appropriate to try a variety of approaches and simply pick the one that puts the product in the best light”.\textsuperscript{11} Dolan echoes this advice and states that the “evidence on the responsiveness of one measure relative to another is in short supply” and that further within patient comparisons are necessary.\textsuperscript{6} Neumann et al. voice concerns about the potential lack of sensitivity to important changes in particular disease states that might be experienced in the application of these systems.\textsuperscript{1} Finally, Hawthorne et al. surmise that sensitivity of these instruments might be context specific with some instruments being more sensitive to health states in some diseases when compared to others.\textsuperscript{10}
In the next sections of this chapter, I will review the four most commonly utilized indirect utility assessment instruments, the comparative data that exists and the choice of QALY weightings used in the published cost-utility analyses of treatments for RA.

2.2 PREFERENCE-BASED, INDIRECT UTILITY ASSESSMENT MEASURES

2.2.1. The Health Utilities Index (HUI) Mark 2 and 3

The HUI Mark 2 and 3 systems (HUI2 and HUI3) are generic preference-based measures which, when used together, describe almost 1,000,000 unique health states.12 The HUI2 and HUI3 health classification systems were designed to directly link the multi-attribute health status classification system used to describe health status with preference-based, multi-attribute utility functions. The preference-based scoring functions convert the descriptive health classifications into values for each attribute and a single value for overall HRQL.12-15

The HUI2 was originally developed to assess the global morbidity burden of childhood cancer. The content of this instrument was based on a study in which lay raters ranked 15 attributes of health according to importance.16 The HUI3 was developed to improve upon the definitions used in the HUI2, be applicable in both clinical and general population health studies, and to have structural independence among its attributes (i.e. such that all combinations of levels in the system are possible).12 Since its creation, the HUI3 has been used in a variety of clinical studies and in five major population health surveys in Canada.14
The HUI2 and HUI3 were developed with the intention of capturing 'within the skin' attributes of health status.\textsuperscript{15} The HUI2 classification system originally consisted of 7 attributes: Sensation (vision, hearing, speech), Mobility, Emotion, Cognition, Self-care, Pain, and Fertility. Although fertility was initially included to assess sub-fertility and infertility sequelae associated with childhood cancer and its treatment, this attribute has been dropped from the current HUI questionnaires. The 8 attributes of the HUI3 classification system are: Vision, Hearing, Speech, Ambulation, Dexterity, Emotion, Cognition, and Pain. Although certain attributes have the same names across the classification systems, they have different underlying constructs.\textsuperscript{12} The Pain attribute assesses severity of pain in the HUI3 system whereas, in the HUI2 system, the frequency of pain and its control are considered. Similarly, for the Emotion attributes, in the HUI3 system there is a focus on happiness versus depression but in the HUI2 system, distress and anxiety are assessed. Finally, the Cognition attribute in the HUI2 system concentrates on learning whereas in the HUI3, the ability to solve day to day problems is assessed. Therefore, studies sometimes combine the HUI2 and HUI3 to take advantage of both of their properties.\textsuperscript{17-19} However, the creators recommend that the HUI3 be specified as the measure for primary analysis due to its larger descriptive system (972,000 possible health states vs. 24,000 possible health states in the HUI2), structural independence, and availability for comparison with population norms.\textsuperscript{12} The authors also suggest that adding the HUI2 to the HUI3 also provides an efficient source to conduct sensitivity analysis on the utility values.

Therefore, the HUI2 and HUI3 are often administered together and have been formatted for interviewer-administration (face-to-face or by telephone) or self-completion.\textsuperscript{12} There are two versions for the viewpoints of the questionnaires: a self-assessment version
where information is collected from people about their own health and a proxy assessment version, where health status information is collected from people other than the subjects. Also, there are different health status assessments periods that can be captured by the HUI systems which are classified as “current” or “usual”. There are three “current” versions which assess periods of the past 1-week, 2-weeks and 4-weeks. The “usual” version does not specify time recall periods. The “current” versions are recommended for clinical or health economic evaluations whereas the “usual” version is recommended for use in population health surveys. 

Both the HUI2 and HUI3 scoring systems are based upon multiplicative multi-attribute utility functions. This facilitates calculation of HRQL index 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 HUI2 and HUI3. Both systems allow for calculation of utility values less than zero (health states considered worse than death) with the lowest possible scores being -0.03 for the HUI2 and -0.36 for the HUI3. The utility scores are assumed to have interval scale properties, whereas the attribute levels do not have interval scale properties. HUI utility scores represent mean community preferences, and the scores have been calculated from preference scores measured in accordance with von Neumann- Morgenstern expected utility theory and extensions of this theory to accommodate multiattribute utility functions. The investigators obtained preferences (using the SG in four marker states and a rating scale for the remainder in a random sample of the population living around Hamilton, Ontario) for single-deficit states, including “corner” states in which the deficit is set to the worst level. Rating scale value scores were
converted to utility scores using a power function determined from the relationship between
the SG and rating scale value scores in the marker states.\textsuperscript{13,14}

A number of the attributes of the HUI2 and HUI3 are specifically relevant to the
study of rheumatoid arthritis, including Mobility, Emotion (from both systems), Self-Care,
Pain (from both systems), Ambulation, and Dexterity. 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’, may be
important to patients with rheumatoid arthritis such as social functioning, work/productivity,
and energy.\textsuperscript{23-27} However, the scoring functions of the HUI2 and HUI3 may indirectly
capture some of these ‘outside the skin’ attributes.

There has been some work to characterize what would represent the minimally
important difference (MID) for the HUI2 and the HUI3.\textsuperscript{12} Grootendorst et al. concluded that
differences on the HUI3 of 0.03 or more should be considered to be clinically important\textsuperscript{28},
whereas Samsa et al.\textsuperscript{29} determined, from a small random sample of 160 patients from a
Veteran’s Administration hospital, that 0.02 (95% confidence interval 0.01 to 0.05) was a
clinically meaningful difference. Based upon these results and the fact that the smallest
difference between utility scores between levels of an HUI attribute is 0.05, the creators of
the HUI systems recommend that a difference of 0.05 (and possibly smaller) is likely
meaningful. However, further research is required to substantiate these recommendations in
variety of different diseases including RA.

To date, no published studies could be identified that examine the properties of the
HUI2 or HUI3 specifically in RA. However, a few studies examine the properties of the
HUI3 system (but not the HUI2) in people with one of several musculoskeletal diseases of
which RA was included.\textsuperscript{26-28, 30} Since most of these are comparative studies with other preference-based measures, they will be discussed in detail in section 2.3.

Analysis of an arthritis patient sub-sample from a population health survey using the HUI3 found the greatest burden of morbidity in pain with very small, but significant (due to large sample sizes), differences in ambulation and cognition compared to a reference group without stroke or arthritis.\textsuperscript{28} However, the major limitation of this analysis was that the diagnosis of arthritis was self-reported and could have represented one of a number of conditions including osteoarthritis, rheumatoid arthritis or other rheumatic conditions. This limitation is substantiated by the relatively high mean HUI3 utility scores for patients reporting "arthritis" in this sample (0.84) which was significantly lower (but not as low as expected if the sample had been limited to only RA) than the mean control score (those without arthritis or stroke) of 0.92.

\subsection*{2.2.2. The EuroQol (EQ-5D)}

The EQ-5D was designed as a cardinal index of health for describing and valuing HRQL.\textsuperscript{31} The main objective in creating this instrument was to develop a standardized measure for describing and valuing HRQL that could be used to generate cross-national comparisons of health status.\textsuperscript{22} The dimensions of the instrument were selected after a detailed review of several, generic health status measures.

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 or utility score), and a VAS-based self-rated assessment of HRQL. The EQ-5D was intended for self-completion and the recall period refers to the present (today).

The scoring algorithm typically applied to the descriptive health classification system is the UK-based York scoring system. This scoring system was generated from interviews of a sample of the general UK population. Respondents were asked to rank and then value hypothetical EQ-5D health states using the Time Trade-Off approach. 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. 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 all domains other than Usual Activities.

Scores on the EQ-5D range from -0.56 to 1.00 where negative values represent health states worse than death. This range is the widest of all utility values determined by any of the preference-based instruments. In addition, the brevity of the EQ-5D has been considered to be a strength in the study of HRQL. In a large (sample size was over 1100 patients in each group), randomized comparison of the two instruments, the response rate of the EQ-5D was shown to be higher than the SF-6D in severely disabled persons after stroke (66% with no missing data vs. 55%, p<0.0001, respectively).
However, the advantage of brevity of the EQ-5D leads to a major limitation in its health classification system – the few domains of health assessed and the small number of health states described. Of all the preference-based instruments, the EQ-5D has the smallest number of health states described (243) compared to the HUI2 (24,000), HUI3 (972,000) and the SF-6D (18,000). In addition, it has the fewest domains of health that are assessed by its system. Therefore, in studies of chronic conditions such as RA, these might not be sufficient to accurately describe impairments in HRQL. For example, the EQ-5D lacks dimensions of HRQL that may be impacted by RA such as dexterity, social functioning and vitality. 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. Also, many researchers have found that there are many gaps in the distribution of scores achieved by the EQ-5D, especially in the mid-utility range (between 0.30 to 0.5) with a clustering of scores in the upper-utility area possibly leading to ceiling effects.30,38

Another concern about the EQ-5D relating to the small number of health states described is its ability to detect responsiveness and sensitivity to changes. Floor and ceiling effects have also been observed to a greater degree in the EQ-5D than in the Short Form-12 (SF-12), an abbreviated version of the Short Form-36 (SF-36).39 In studies comparing the EQ-5D to the SF-36, the EQ-5D index score was less responsive to change and less able to discriminate between groups than the SF-36.40-42 A recent study comparing the EQ-5D to the SF-6D (a preference-based measure derived from the SF-36 – see below for more details), the authors found that the SF-6D was more sensitive than the EQ-5D in detecting small changes in patients who had undergone liver transplantation.43
Specifically, in RA, there have been two studies that have examined the construct validity, responsiveness and reliability of the EQ-5D. Hurst et al. first reported on the validity of the EQ-5D in terms of its ability to measure both current health status and change in health status in a small sample of 55 patients with RA. At baseline, the EQ-5D index scores were significantly correlated with other condition-specific measures including loss of function, joint pain, joint tenderness and mood. In addition, EQ-5D change scores were correlated with changes in these measures. In a larger study with 233 RA patients, these results were repeated. In addition, the investigators found that the test-retest reliability of the EQ-5D utility score (ICC=0.78, 95% CI 0.60-0.96) after two weeks which was higher than all other clinical measures except the Health Assessment Questionnaire (HAQ) scores. Therefore, it appears that the EQ-5D has demonstrated reliability and cross-sectional and longitudinal construct validity in RA. However, from these results, it was unclear how these properties for the EQ-5D would compare to other preference-based instruments in the assessment of RA.

In a study comparing the responsiveness of generic health status measures in patients with RA who were either receiving infliximab or not receiving infliximab, SF-6D scores were found to be consistently higher than EQ-5D scores. These authors found lower test-retest reliability for the EQ-5D (ICC = 0.66) compared to the SF-6D (ICC=0.72). In addition, these authors found mean changes almost two-fold greater in the EQ-5D than the SF-6D in those patients receiving infliximab. However, effect sizes as a measure of responsiveness were larger in the SF-6D due primarily to the smaller standard deviation at baseline (0.07 for the SF-6D vs. 0.30 for the EQ-5D). Other studies comparing the EQ-5D to other preference-
based measures have been conducted outside of RA and will be discussed in detail below.\textsuperscript{30,43}

No studies were identified that identified what is thought to be the minimally important difference (MID) in index scores of the EQ-5D. It has been hypothesized that the MID is 0.03 since this represents the smallest change in the utility values that can occur as a result of a one category change in a single dimension. Further research is necessary to further characterize the MID for the EQ-5D.

\subsection*{2.2.3. The Short Form 6D (SF-6D)}

The most recent of the preference-based, indirect utility assessment instruments, the SF-6D, was created by Brazier et al. in an effort to derive a scoring algorithm to derive preference-based scores from the SF-36.\textsuperscript{46,47} The SF-36 is one of the most widely utilized HRQL measures and contains 36 questions assessing these eight dimensions: physical functioning, role limitation due to physical health, social functioning, vitality, bodily pain, mental health, role limitation due to emotional problems, and general health.\textsuperscript{48} The SF-6D revised the SF-36 into a six-dimensional health state classifications system assessing physical functioning, role limitations, social functioning, pain, mental health, and vitality.

The SF-6D health classification system defines health states by a respondent selecting one level from each of the six dimensions. Each dimension has four to six levels and thus, 18,000 possible health states are defined in this manner.\textsuperscript{22,47} To assess preferences for the multi-attribute health states defined by the SF-6D system, the creators used an interviewer administered SG in a representative sample from the UK.\textsuperscript{47} The boundaries of the SF-6D utility scores are from 0.30 to 1.00 with a score of 1.00 being indicative of "full health". The
MID of the SF-6D, based upon a meta-analysis of seven longitudinal studies, was determined to be the 0.033 (95% CI 0.029 to 0.037).49

Due to the newness of this preference-based, indirect utility assessment instrument, there have been few published studies in which it has been utilized. However, the use of this measure is increasing, and with the availability of many SF-36 datasets that could be converted into preference-based measures, it is anticipated that the application of this measure will continue to grow.26,30,43,49-51 Of note, the creators of the SF-6D state that, when compared to the EQ-5D, the fact that the SF-6D has a much larger descriptive system may result in greater sensitivity.47 Specific studies comparing results obtained with the SF-6D and other preference-based measures will be described in detail below.

2.3 EMPIRIC COMPARISONS BETWEEN THE INDIRECT UTILITY ASSESSMENT INSTRUMENTS

2.3.1. Comparisons between the Health Utilities Index Mark 2 and Mark 3

One of the first studies published in the literature examining differences achieved with scores achieved from the HUI2 and HUI3 was authored by Neumann et al..52 These investigators compared scores achieved with the two HUI systems in a cross-sectional sample of 679 patients with Alzheimer's disease (AD) and their caregivers. In addition, the investigators utilized the scores obtained by the two systems in a decision-analytic, Markov-model based, economic evaluation of a new drug to determine what the impact using the different utility values would be on the incremental cost-effectiveness ratios. When patients completed the questionnaires, their mean (SD) utility scores were lower on the HUI3
(0.22[0.26]) than on the HUI2 (0.53 [0.21]). However, when caregivers completed the questionnaires as proxies for the patients, similar results were found between the two systems (mean score [SD] on the HUI3 of 0.87 [0.14] and HUI2 0.87 [0.11]). Both systems appeared to have construct validity in terms of their ability to discriminate between severity levels of AD. For the HUI3, patient scores ranged from 0.47(0.24) for questionable AD to -0.23 (0.08) for terminal AD, compared with a range of 0.73 (0.15) to 0.14 (0.07) for the HUI2. In the results of the cost-effectiveness analysis, the results were more economically attractive when the scores for the HUI3 were used as compared to the HUI2.

Maddigan et al. examined the construct validity of the two HUI systems in 394 patients with type 2 diabetes in rural communities in Alberta and subsequently compared the scores of the two systems and examined reasons for their differences. The mean score of the HUI2 was higher (0.78, SD 0.18) than the mean score of the HUI3 (0.64, SD 0.30). Using the “known groups” approach to the assessment of construct validity, the investigators found that the HUI2, HUI3 and the RAND-12 all discriminated across subgroups of individuals representative of more and less advanced diabetes or differing levels of disease severity. For example, disease severity measures were associated with impairment on the vision ambulation, dexterity and pain attributes on the HUI3 and impairments on self-care and mobility attributes of the HUI2. Overall scores were lower in those above the median duration of diabetes than those below and in those whose diabetes was managed using insulin compared to diet alone.

In the paper comparing the HUI2 to the HUI3 scores and examining the extent to which each of these systems detect differences associated with varying levels of type 2 diabetes severity or disease advancement, 372 individuals were available for analysis. Specifically,
differences were investigated in single attribute and overall HUI2 and HUI3 utility scores of groups with presumed differences in disease severity or stability of control. Severity of type 2 diabetes was defined based upon those receiving insulin therapy (most severe) to those treated with diet alone (least severe). Stability of control was defined based upon absenteeism from work, emergency room visits, and hemoglobin A1c values. Relative to HUI2 scores, larger differences were seen in HUI2 scores for individuals defined as having more advanced type 2 diabetes. Both the pain and emotion attributes of the HUI3 categorized a larger proportion of the sample as having moderate to severe impairment than corresponding attributes in the HUI2 system. These observations prompted the authors to conclude that, due to the greater range of possible scores (including the wide range of states valued as worse than death) and its superior ability (relative to the HUI2) to discriminate between those with moderate or severe impairment as compared to mild or no impairment, the HUI3 may be a better instrument to utilize in type 2 diabetes.

The responsiveness of the HUI systems was recently compared to the SF-36 and disease-specific measures (the Harris Hip Scale, Western Ontario and McMaster University Osteoarthritis Index (WOMAC), and McMaster-Toronto Arthritis Patient Preference Disability Questionnaire (MACTAR)) in patients undergoing hip arthroplasty. Feeny et al. evaluated the responsiveness of these questionnaires in 90 patients (out of a possible 553 patients who had been initially referred for hip disease). Questionnaires were applied prior to surgery and post-surgery facilitating comparisons between pairs of measures for each patient. The responsiveness statistics utilized were the effect size (ES), the standardized response mean (SRM), the relative efficiency statistic (RE) and the paired t-test. Some form of improvement was detected by the overall/summary scores for all of the instruments and in
many of the domain scores of the SF-36 and the single-attributes for the HUI systems. All of
the overall scores had large ES statistics including the HUI2 and HUI3. For example, for the
SF-36, improvements were observed in the physical functioning, bodily pain and vitality
domains as well as the physical component summary score; for the HUI2, improvements
were observed in the pain and self-care attributes; and finally, for the HUI3, improvements
were observed in the pain and ambulation attributes. Surprisingly, the mobility attribute for
the HUI2 was not responsive. Overall, as hypothesized, the disease-specific measures were
most responsive but the generic measures yielded acceptable responsiveness statistics and
would be suitable to be used in this context.

Another publication has resulted from this data set which examines differences between
community based preferences (based on responses to the HUI2 and HUI3) and individual
preferences (based upon SG utilities). The investigators examined agreement (as measured
by intraclass correlation coefficients [ICC]) between the different utility assessment
techniques and compared the mean scores between instruments. Mean scores were
statistically higher for the SG when compared to the HUI3 (0.62, SD 0.31 vs. 0.52, SD 0.21)
but not the HUI2 (0.62, SD 0.31 vs. 0.62, SD 0.19). However, the ICCs were low ranging
from 0.06 (agreement between the SG and HUI2) to 0.09 (agreement between the SG and the
HUI3). Thus, the authors concluded that the HUI2 was a good proxy for directly measured
SG at the group level; however, this conclusion cannot be applied at the individual level (as
evidenced by the low agreement).

Feeny et al. conducted another study examining the relationships between the HUI2,
HUI3 and directly-measured SG scores at both the individual- and group-level in a sample of
140 teenage survivors of extremely low birthweight (ELBW) and 124 control group teens.
Again, mean SG scores were compared to HUI2 and HUI3 scores and agreement was assessed between the SG and the HUI2 and HUI3 using the ICC. For the ELBW group, the SG, HUI2 and HUI3 mean (SD) scores were 0.90 (0.20), 0.89 (0.14), and 0.80 (0.22) compared to the control scores of 0.93 (0.11), 0.95 (0.09) and 0.89 (0.13) respectively. The differences between the ELBW and control HUI2 and HUI3 scores were significantly different (p<0.0001 and p<0.0002) and clinically important. However, no such differences were observed between the SG in the sample and the controls. Also, although there were no differences between mean SG and HUI2 scores, mean SG and HUI2 were significantly different (p<0.001) than HUI3 scores with the latter being systematically lower. In the assessment of agreement, the ICCs between the SG and the HUI2 or HUI3 were very low indicating poor agreement at the individual level. Agreement between the HUI2 and HUI3 was moderate at 0.63 (95% 0.01 to 0.73) in the sample and control groups combined. Again, the authors conclude that, at the group level, results from the HUI2 and the SG are interchangeable but this relationship did not hold up at the individual level.

2.3.2. Comparisons across Indirect Utility Assessment Instruments Outside of Musculoskeletal Diseases

The first study comparing preference based scores derived from indirect utility assessment instruments arose within the framework of a randomized clinical trial of 561 patients being treated with tirilazad mesylate or placebo for aneurysmal subarachnoid hemorrhage. The EQ-5D VAS, the HUI2 utility scores and a rating scale were used as measures of patient preferences. The scoring function of the EQ-5D was not considered as the authors stated that they were assessing patient preferences (rather than societal
preferences). The measures of preferences tended to have higher agreement at lower levels of functioning and poor agreement at higher levels of functioning. Since this study did not directly compare the scores that are typically utilized as QALY weights, little can be concluded from this research. However, this study likely sparked the interests of other investigators to conduct comparisons between the indirect utility assessment instruments and was the forerunner to a body of research.

The creation of the Assessment of Quality of Life (AQoL) questionnaire led investigators to compare its properties with those of the HUI3, EQ-5D, SF-6D and a Finnish preference-based measure, the 15-D. Of note, the SF-6D scores were calculated by an older algorithm which has since been changed. The investigators administered these instruments to residents in Victoria, Australia. The sample was selected to provide a heterogeneous, representative sample of community members weighted by socioeconomic status, chronically ill patients attending outpatient clinics in two of Melbourne’s largest hospitals, and inpatients from three hospitals. The response rate was 58% (n=396), 43% (n=334) and 58% (n=266) for the community, outpatient and inpatient samples, respectively for a total number of 976 respondents.

The investigators found that the distributions of the scores of the five instruments were quite different with AQoL, HUI3 and EQ-5D having a greater range of scores and lower values than the SF-6D and the 15-D. However, when broken down by sample type (community, outpatient, inpatient) and by age (16-35, 36-50, 51-65, and >66), all the instruments displayed a monotonic, decreasing relationship between sample and age-groupings. Spearman’s correlations between each pair of instruments scores revealed high (>0.60) correlation coefficients for all comparisons. The AQoL and the 15-D had the highest
correlation (0.80) whereas the EQ-5D and the HUI-3 had the lowest (0.64). Finally, in a more detailed analysis of patterns of agreement between the instruments, it was revealed that a change in the average score in the SF-6D and the 15-D corresponded to a much greater change in the scores predicted by the other three instruments.

In the determination of ceiling effects (where scores cluster at the highest ends of the scale), scores of the other instruments were plotted when the score of the instrument of interest was at a maximum value of 1.00. By examining the dispersion in the other scores, the investigators determined the ability of the other instruments to detect differences in health states when the instrument of interest was at its ceiling. The results showed that the dispersion of scores for the other instruments when the AQoL, 15D, or SF-6D were at 1.00 was minimal suggesting that these instruments had a relatively high ceiling. However, when a utility value of 1.00 was achieved with the HUI3 and the EQ-5D, there was significant dispersion of scores in the other instruments suggesting a possible low ceiling effect. Thus, it would appear that despite having a wider range, both the HUI3 and the EQ-5D display ceiling effects that are not experienced by the SF-6D.

Bosch and Hunink compared the HUI3 and the EQ-5D in 88 patients treated for intermittent claudication in the Netherlands. These patients completed the HUI3, EQ-5D, RAND 36-Item Health Survey 1.0, TTO, SG and rating scale before revascularization and at follow-up at 1 month after the procedure. After revascularization, improvements were mostly noted in the HUI3 attributes of pain and ambulation compared to mobility, usual activities and pain/discomfort domains in the EQ-5D system.

It was hypothesized that since TTO scores are usually lower than SG scores, the mean TTO and the EQ-5D (which uses the TTO in its scoring function) scores would be...
lower than the mean SG and the HUI3 (which uses the SG in its scoring function as described above) scores. Prior to treatment, the EQ-5D mean (SD) score (0.57 (0.25)) was significantly lower than the HUI3 mean (SD) score (0.66 (0.20), \( p < 0.01 \)). Also, as hypothesized, the TTO mean (SD) scores (0.82 (0.17)) were lower than the SG mean (SD) scores (0.91 (0.14)). However, at 1 month after the procedure, there were no differences between the HUI3 and the EQ-5D scores (0.77 (0.21) vs. 0.79 (0.23), respectively). To investigate agreement at the individual level, the investigators determined the ICC values between the HUI3 and EQ-5D at baseline (0.49) and at 1 month after the procedure (0.66). The ICC between the changes in the HUI3 and EQ-5D score was poor (0.30). The authors concluded that studies utilizing the mean values of these systems (such as in cost-utility analysis) would conclude a lower impact on HRQL due to revascularization if the HUI3 was used instead of the EQ-5D due to the smaller changes in utility scores.

Longworth and Bryan conducted a comparison of the EQ-5D and the SF-6D in liver transplant patients in 524 patients (90% response rate for at least one questionnaire) in the UK.\(^{43}\) Investigators administered the HRQL questionnaires from point of listing on the transplant list and then in 3 month intervals until transplantation. After transplantation, HRQL questionnaires were given at 3, 6, 12 and 24 months. At the conclusion of the study, there were 1462 data pairs (at two consecutive time points) to compare the two indirect utility assessment instruments.

When results of the mean scores of the two instruments were compared at baseline (listing time) to 12 months post-transplantation, the EQ-5D detected a significant improvement in HRQL (mean score increased from 0.52 to 0.61, mean change of 0.09, 95\%
CI 0.03 to 0.14); however, the SF-6D did not show a significant change (mean score increased from 0.61 to 0.62, mean change of 0.01, 95% CI -0.04 to 0.05).

In pre-transplantation measurements, the SF-6D was found to have a much narrower spread and symmetrical distribution when compared to the EQ-5D. Due to its relatively high lower bound of 0.30, scores could not dip lower than this value. Conversely, no patients were scored as 1.00 (full health) by the SF-6D system. On the other hand, the EQ-5D had a sizeable proportion of respondents classified as health states worse than dead (negative values) at any time point prior to transplantation and a number of patients scoring full health at all time points prior to transplantation. In post-transplantation measurements, the results were similar with less of the EQ-5D scores being in the “worse than dead” range but more at the full health point (1.00). The distribution of the SF-6D scores was similar as those achieved prior to transplantation with a few patients reporting full health (1.00). Of note, there were also gaps in the distribution of the EQ-5D scores, with the most noticeable being between 0.37 to 0.50 and 0.88 to 1.00.

Although the correlations between the EQ-5D and SF-6D scores were high (0.76, p<0.001), there was a large amount of variation in the scores across the measures. In the examination of ceiling effects, when the EQ-5D was scored as 1.00 (a total of 237 paired observations), only 22 SF-6D also were at full health. The remaining SF-6D scores ranged from 0.57 to 0.99 with a mean score of 0.82. Thus, it would appear that towards the higher range of utility scores, the SF-6D showed greater sensitivity. However, the reverse was true when floor effects were examined. More respondents indicated the lowest levels on the SF-6D than the EQ-5D domains. For example, 42% and 21% of respondents indicated the most severe levels on the role limitation and vitality domain, respectively on the SF-6D.
questionnaire with the largest proportion indicating the worst level on any EQ-5D domain being usual activities at 14%.

Therefore, from the results of this analysis, it would appear that despite having better properties at the upper end of the utility range, the SF-6D displays floor effects. This finding is likely limited to disease states where the burden of disease is large as in organ transplantation. As such, the use of the SF-6D over instruments such as the EQ-5D may underestimate the magnitude of HRQL improvements in these types of conditions and undervalue treatment in cost-utility analysis. This finding is somewhat in agreement with the statement by Brazier et al. in that "any greater sensitivity [of the SF-6D] would be most likely in groups experiencing mild to moderate health problems and in those expected to experience comparatively small changes or where small differences are expected between interventions."47

O’Brien et al. examined the level of agreement between the SF-6D utility algorithm and the HUI3 in patients at increased risk of sudden cardiac death participating in a randomized trial of implantable defibrillator therapy.51 The SF-6D and the HUI3 questionnaires were completed at baseline by 246 patients generating cross-sectional scores. The mean values from the HUI3 (0.61, 95% CI 0.60 to 0.63) and the SF-6D (0.58, 95% CI 0.54 to 0.62, p<0.03). As shown with other studies, the range of the HUI3 scores were much greater (-0.21 to 1.00) as compared to the SF-6D scores (0.30 to 0.95). The distributions of the scores of the two systems again were quite different with the SF-6D passing the Kolmogorov-Smirnoff test for normality. The distribution for the HUI3 scores failed statistical tests for normality and followed a skewed, bimodal pattern. Agreement, assessed using the intraclass correlation coefficient (ICC), was low (ICC 0.42, 95% CI 0.31 to 0.52).
In their discussion, the authors raise several interesting points on potential reasons for the differences observed in the scores from the two instruments. Firstly, the SF-6D considers different domains of health while the HUI3 is based on “within the skin” attributes. Secondly, although their scores are both based on the SG, SF-6D health states were valued directly while the HUI3 health states were directly valued by RS and converted to SG scores by a statistical power function. Finally, although the absolute scores for these instruments were different in a cross-sectional study, it is not clear from their results if difference scores would vary to the same extent in a longitudinal analysis.

Siderowf et al.\textsuperscript{60} compared the scores of three preference-based instruments, the EQ-5D, the Disability and Distress Index (DDI), and the HUI2 in 100 patients with idiopathic Parkinson’s disease (PD). The DDI contains four functional domains: general mobility, usual activities, self-care and social and person relations. Responses on these domains are combined with an overall rating for the dimension “distress” and the entire system is scored between -1.486 and 1.0. While the DDI appears to be preference-based, it does not provide utility values. Construct validity of the three preference-based instruments was determined by comparing their scores with the total Unified PD Rating Scale [UPDRS] (a widely used disease-specific, symptom severity rating system), the Hoehn and Yahr scale (disease severity scale in PD) and the Beck Depression Inventory. The three instruments’ discriminative ability was tested by dividing the study sample into upper and lower halves and quartiles based on the UPDRS.

Overall, the mean (SD) of the three scores were: EQ-5D 0.59 (0.27), HUI2 0.75 (0.18) and DDI 0.93 (0.17) which were significantly different (\(p<0.001\) for pairwise comparisons). Only the EQ-5D supplied scores that were negative (health states valued as
worse than death). The scores between the instruments were moderately to strongly correlated with Pearson correlations coefficients of 0.74 (HUI2 with the EQ-5D), 0.62 (EQ-5D with the DDI) and 0.56 (DDI with the HUI2). All three instruments were significantly correlated with disease-specific measures. Generally, the DDI had lower correlations with these measures than the HUI2 and EQ-5D. In terms of discriminative ability, the HUI2 and EQ-5D were superior to the DDI in their ability to distinguish between severities of PD. None of the instruments were able to distinguish between subjects with and without motor fluctuations or drug-induced dyskinesia. In their discussion of the research findings the authors raise an interesting point – namely, that because all the instruments yielded scores that were correlated with the disease-specific measures, they might be measuring functional status much more than preferences. In order to test this hypothesis, further studies examining correlations and agreement with directly elicited preference techniques such as the SG and TTO.

Using 36 clinical experts to score the HUI2, HUI3 and the EQ-5D classification systems according to literature reports on eight sequelae associated with childhood meningitis, scores obtained using these three health classification systems were compared. The sequelae chosen in the valuation exercise were deafness, minor hearing loss, epilepsy, mild mental retardation, severe mental retardation combined with tetraplegia, paresis of the leg, and mild mental retardation combined with epilepsy and paresis of the leg. For each of the sequelae, the investigators constructed a short, structured synopsis that reported on relevant domains. In general, scores on the HUI2 and the EQ-5D were comparable except for the severe retardation and tetraplegia sequelae which was scored, on average, to be -0.15 (0.13) with the EQ-5D and 0.12 (0.03) with the HUI2. Interestingly, with the same health
state, the score on the HUI3 was -0.33 (0.02). All health states were scored significantly lower using the HUI3 than the other systems (p<0.05 for all). However, the HUI2 and HUI3 had the same ranking for the health states. Rankings with the EQ-5D system were similar except it ranked epilepsy lower than mild hearing loss and leg paresis lower than deafness in contrast to the HUI systems. Using various measures of agreement, there were significant differences across all three of the instruments for each of the sequelae suggesting that they were not interchangeable. From their results, the authors concluded that sensitivity analyses of QALY weightings must be employed in cost-utility analysis in order to account for the observed differences in scores.

Lubetkin et al. examined the relationship between the SF-12, the EQ-5D and the HUI3 for overall scores and in analogous domains of health in a convenience sample of 301 participants (77% participation rate) at an inner-city community health centre in New York City. Participants were mainly from ethnic minorities, had low annual incomes (90% earned less than $30K), and low education (47% had high school graduation or less). Using Pearson’s correlation coefficients, correlation between the overall scores ranged from 0.41 (SF-12 with EQ-VAS) to 0.69 (HUI3 with EQ-5D index). Considering just the preference-based measures, correlations between similar domains (using Kendall’s tau for ordinal variables) were 0.59 (between the HUI3 ambulation attribute and the EQ-5D mobility domain), 0.58 (between the HUI3 pain attribute and the EQ-5D pain domain, and 0.55 (between the HUI3 emotion attribute and the EQ-5D anxiety/depression domain). Areas of impairment most frequently detected by the HUI3 were pain, vision, cognition and emotion, whereas, for the EQ-5D pain/discomfort and anxiety/depression were impaired most often. The authors concluded that despite differences in the structure of these systems, correlations
between related aspects were moderate to strong and participants demonstrated consistency in responses across analogous items.

From a population health perspective, there have been comparisons of the HUI3 and the EQ-5D both in the UK and in Canada.\textsuperscript{63,64} In the comparison in the UK, the EQ-5D, a modified version of the HUI3 and the SF-12 were compared within a general population sample.\textsuperscript{63} The modified version of the HUI3 that was used was an eight item questionnaire (one for each domain) that is available at no charge from the developers. The authors claimed that they could not afford the fees associated with using the standard HUI3 questionnaire as they are substantial.

The three instruments were evaluated in terms of their feasibility, coverage (such that there should be a broad range of responses across its items), and discrimination (ability to discriminate between individuals based upon self-rated health status, measurable morbidity, and socioeconomic status). All instruments showed feasibility in that there were low non-response rates (less than 6% for all the items across all questionnaires). The SF-12 had a broad distribution of scores across its items although there was still heavy skewing towards responses indicative of good health. However, the HUI3 and the EQ-5D scores were highly skewed on all dimensions with the majority reporting full health. In the EQ-5D, respondents were least likely to report problems on the self-care domain but most likely to report pain/discomfort. Forty-nine percent of respondents indicated no problems on all five domains of the EQ-5D. For the HUI3, respondents were least likely to report problems on the speech dimension but most likely to report decrements in the pain dimension. In the sample, there were 35 distinct health states (out of the possible 243) described by the EQ-5D system compared to 126 distinct health states defined by the HUI3. Finally, in terms of their ability to
discriminate between self-reported health states and socioeconomic status, all three instruments had acceptable levels of performance. The SF-12 summary scores could not discriminate among people with different education levels. The authors concluded that, despite the differences in their descriptive systems and scoring functions, overall there was no discerning feature to pick one over the other as a population health measure.

The Canadian study attempted to assess the relationship between the HUI3 and the EQ-5D at both the descriptive and scoring level. The analysis is performed on answers given by 1,477 respondents of Statistics Canada 1998 National Population Survey pilot study. Both the HUI and the EQ-5D mean scores declined with increasing age and with decreasing self-perceived health. People who report chronic conditions had lower scores and people with more severe problems a larger change in the scores. The HUI3 and EQ-5D scores were moderately correlated (0.58 for Spearman correlation coefficient) as were answers to self-related health questions (coefficients varying between 0.48 and 0.56). The instruments' mean scores had reasonable agreement in less healthy respondents and respondents of younger age (16-34) but, in healthier respondents, the mean scores were less similar. Thus, based on these results, it was decided to continue to use the HUI3 for the Canadian National Population Health Surveys.

2.3.3. Comparisons across Indirect Utility Assessment Instruments within Musculoskeletal Diseases

Recently, three studies comparing the various properties of some of the indirect utility instruments in samples with rheumatologic conditions were published. Of these, in only one of the studies did all patients have RA while in the other two studies, patients had
a mixture of musculoskeletal diseases. The study conducted exclusively in RA, authored by Russell et al., examined the reliability and responsiveness of the SF-36, SF-6D, EQ-5D, SG, the modified HAQ, and a pain VAS in two groups of RA patients (Group 1 consisted of 24 patients with stable RA and Group 2 consisted of 60 patients beginning infliximab therapy). Patients in group 2 were assessed prior to being initiated on infliximab therapy and after 14 weeks of infliximab treatment.

Test-retest reliability was estimated for each instrument in the stable patient group using the ICC whereas responsiveness was assessed by using the paired t-test, effect size (ES) and standardized response mean (SRM). For all the measures, the ICC ranged from 0.50 (role emotional domain from the SF-36) to 0.92 (physical functioning domain from the SF-36). The preference-based measures had moderate reliability (ICCs of: EQ-5D 0.66, SF-6D 0.72, SG 0.73). However, the sample from which these results were derived was very small (n=24) and thus, these estimates may not be robust. In terms of responsiveness, for Group 2; all the overall scores and domain scores for the SF-36 detected significant changes from baseline to the second measurement. Standardized response means (SRM) and ES were the largest for the pain VAS, the EQ-5D VAS, the SF-36 physical component scores, and the SF-36 vitality domain. In terms of the preference-based measures, the SRM and ES values were 0.67 and 0.64 for the EQ-5D, 1.40 and 0.87 for the SF-6D, and 0.49 and 0.43 for the SG. Despite the fact that the change described by the EQ-5D system was twice that described by the SF-6D, the responsiveness statistics were much smaller mainly due to the larger SD of the baseline and change scores of the EQ-5D. The authors concluded that the SF-6D might be a preferable to the EQ-5D in measuring clinically-relevant improvement in RA.
Conner-Spady et al. assessed the interchangeability of preference-based, indirect utility assessment instruments (the EQ-5D, HUI3, and the SF-6D) in a longitudinal study. One hundred and sixty one patients (of the 252 initially approached) with at least one of several rheumatological conditions (51% had RA, 19% had low back pain, 14% had knee osteoarthritis, 12% had fibromyalgia and 3% had psoriatic arthritis) participated in the baseline questionnaire and 98 patients had data both at baseline and 12 months later. Of the 98 patients. The mean scores (SD) of the instruments at baseline were 0.49 (0.31), 0.50 (0.54) and 0.62 (0.14) for the EQ-5D, HUI3 and SF-6D respectively. Distributions of the three measures were very different with the EQ-5D having a bimodal distribution with two gaps between 0.28 and 0.50 and another between 0.88 and 1.00. The HUI3 score distribution was more continuous with a wide range from -0.21 to 1.00. The SF-6D had a normal appearing distribution. Of the three instruments, it appeared that EQ-5D had some ceiling effects when compared to the other two instruments. For specific domains that were analogous across the instruments, \( \geq 97\% \) reported decrements on the pain domains across all instruments, 42\% (HUI3) to 77\% (SF-6D) reported impairment for mental health, and 52\% (HUI3) to 98\% (SF-6D) for impairments on mobility, ambulation, or physical functioning.

Responsiveness of the three instruments from baseline to 12 months was assessed using a self-reported change question (dividing the group into "better", "same" and "worse" subgroups) and the ES. After 12 months, 41\% reported their health to be better, 31\% the same and 28\% as worse compared with baseline. Using a repeated measures ANOVA, a significant tool effect with significantly higher SF-6D scores and a significant tool by time by group interaction with the EQ-5D scores showing a significantly greater mean improvement than the other two instruments (changes of 0.15, 0.07 and 0.05 for the EQ-5D,
HUI3, and the SF-6D respectively). For the group reporting their health to be “worse”, the EQ-5D showed a significantly greater mean decrease (0.19) than either the HUI3 (0.05) or the SF-6D (0.03). For the “better” and “worse” groups, there were no significant differences in the ES between the instruments. However, for the “worse” group, the EQ-5D had a significantly larger ES than the HUI3 and SF-6D. The authors concluded that the instruments, although measuring a similar underlying construct, were not interchangeable and could result in substantially different estimates if used in a cost-utility analysis. The main limitation of this study was the inclusion of several disease states which may have influenced different domains covered by the various instruments. Thus, it was difficult to separate out the performances of the instruments in any particular disease state.

Finally, the EQ-5D was compared to the HUI3 in sample of patients with rheumatic diseases in Singapore. Specifically, the authors compared overall utility scores, test-retest reliability, and construct validity of these instruments in 114 patients with rheumatic diseases (49 had RA, 31 had lupus, and 24 had osteoarthritis).27 Test-retest reliability was assessed using ICC values. Construct validity of the instruments was assessed by, based upon median values, dichotomizing SF-36 scores, pain VAS scores, tender points, and the number of other acute/chronic conditions and conducting t-tests and Mann Whitney U tests on the HUI3 or EQ-5D scores in each of these groups (i.e. mean HUI3 scores in those above and below the median of the SF-36 would be compared). Agreement between responses on analogous domains between the instruments was examined.

The test-retest reliability of the EQ-5D was 0.64 compared to 0.75 for the HUI3. The means (SD) of the preference-based scores were 0.75 (0.21) for the EQ-5D and 0.76 (0.17) for the HUI3. Correlation between the two instruments’ baseline scores was 0.45. The EQ-
5D system classified patients into 16 unique health states whereas the HUI3 system classified patients into 72 states. For the pain dimensions on the two instruments, 78% reported deficits on the EQ-5D while 90% reported decrements in this domain on the HUI3. As expected for both instruments, patients classified as having worse health status had lower scores than those classified as having better health status (by all of the criteria). Correlations between the two preference based measures and the SF-36 domain scores ranged from 0.23 to 0.55 for the EQ-5D and from 0.29 to 0.49 for the HUI3 (with the highest correlation for both instruments being with the “bodily pain” domain). The authors concluded that the EQ-5D and the HUI3 performed equally well in assessing HRQL although they measured different dimensions. As with the previous study, the inclusion of multiple disease states makes interpretation of the results difficult. In addition, despite collecting longitudinal data, responsiveness was not assessed.

2.4 QUALITY WEIGHTINGS IN THE ESTIMATION OF QALYS IN COST-UTILITY ANALYSES IN RA: WHAT ARE INVESTIGATORS USING?

With the availability of new, effective and costly pharmacotherapeutic interventions for RA, economic evaluations of these therapies are becoming more common. Increasingly, the methodology utilized to conduct these analyses falls under the cost-utility framework and an incremental cost per QALY is often calculated. This approach is supported by the publication of a recent consensus-based reference case for economic evaluations of programs or interventions in the management of RA. Recommendations outlined in the consensus document advocate the use of QALYs as outcome measures but also stated that disease-specific measures could also be considered.
The consensus document also attempts to address which sources should be utilized for QALY weightings and states that both direct and indirect (specifically naming the EQ-5D and the HUI3) methods are acceptable to utilize. 82

As outlined above in previous sections of this chapter, the use of different quality weighting sources in the estimation of QALYs across economic evaluations (even within the same disease area) could lead to very different estimations in the incremental number of QALYs and, therefore, the incremental cost-effectiveness (or cost-utility) ratio. 26,30,43,52,60 Although the magnitude of this potential problem has not been directly explored in an actual economic evaluation of a therapy or intervention for RA, Suarez-Almazor and Conner-Spady utilized a hypothetical intervention and results from small surveys conducted with the EQ-5D, RS, TTO and SG techniques in the general public (n=104), patients with RA (n=51) and health professionals (n=43). 81 Significant differences were found between the scores achieved on the different preference-based methods by technique and by the sample that was surveyed. As such, the incremental cost per QALYs calculated using these weights for a hypothetical intervention with RA ranged from $40,000 to $220,000.

Therefore, the question arises as to what investigators have been using as preference weights for the calculation of QALYs in economic evaluations of RA. For example, if standardization has already occurred through the mutual yet independent selection of an instrument or technique that appears to be the best suited to measure elements that are germane to RA, then there may be little or no problem in this regard. However, if there are several different weightings applied to economic evaluations in studies in the literature without an attempt to standardize the outcomes, the results would be very difficult to compare. As such, Table 2.1 was compiled to examine the different sources for QALY
weights that appear in economic evaluations for RA interventions. As can be seen from Table 1, sources for weighting come from both directly assessed and indirectly assessed preference measures. The SG is the most commonly applied utility weighting being used four times in economic evaluations. The RS and EQ-5D were the next most commonly used instruments followed by the TTO. The SF-6D and the HUI systems have not yet been applied in economic evaluations for interventions or treatments for RA.

2.5 SUMMARY

QALY is the preferred measure used to integrate both years of life and health–related quality of life into the effectiveness measure in economic evaluations. The quality weightings for QALYs are based upon HRQL measured anchored at 0 (dead) and 1.00 (full health). Preference-based weightings for the estimation of QALYs are generally recommended and can be either directly elicited or indirectly elicited by using a health classification and scoring system. Techniques to directly elicit preferences include RS, SG, and TTO methods. Indirect measurement techniques that are widely used include the HUI2, HUI3, EQ-5D, and the SF-6D. The application of the preference-based, indirect measurement techniques is expanding due to their ease and low cost of administration when compared to the directly elicited techniques.

In the literature, there have been a number of recent studies comparing the indirect preference measurement systems in a variety of disease states. Results from these studies suggest that although the instruments perform reasonably well on their own in terms of feasibility, reliability, validity and responsiveness, there are important differences among them. These differences could result in significant variation in the calculation of incremental
cost-effectiveness ratios when different instruments are applied as the quality weights in the estimation of QALYs.

In RA, there has been little work comparing the properties of the various indirect, preference-based, utility assessment instruments. The published economic evaluations that have included the QALY as an outcome measure make use of several different weighting sources (RS, TTO, SG, and the EQ-5D), making comparisons of outcomes between studies difficult. In addition, there have been no studies within RA that determine what the potential impact would be of using different indirect sources of QALY weightings on the outcomes of economic evaluations. As such, additional work is required in these areas.
2.6 REFERENCES


41. Essink-Bot M-L, Krabbe PFM, Bonsel GJ, Aaronson NK. An empirical comparison of four generic health status measures: The Nottingham Health Profile, the Medical Outcomes Study 36-item Short Form Health Survey, the COOP/WONCA charts and the EuroQol instrument. Med Care 1997;35:522-537.


49. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health and Quality of Life Outcomes 2003;1:4 (available at http://www.hqlo.com/content/1/1/4).

the bridge: agreement between the SF-6D utility algorithm and the Health Utilities

comparison of HUI2 and HUI3 utility scores in Alzheimer's disease. Med Decis
Making. 2000; 20: 413-422.

of the RAND-12 and Health Utilities Index Mark 2 and 3 in type 2 diabetes. Qual

54. Maddigan SL, Feeny DH, Johnson JA for the DOVE Investigators. A comparison of
the Health Utilities Indices Mark 2 and Mark 3 in type 2 diabetes. Med Decis Making

S. Is the Health Utilities Index responsive in total hip arthroplasty patients? J Clin
Epidemiol 2003;56:1046-1054.

S. Comparing community-preference-based and direct standard gamble utility scores:
evidence from elective total hip arthroplasty. Inter J Technol Assess 2003;19:362-
372.

57. Feeny D, Furlong W, Saigal S, Sun J. Comparing directly measured standard gamble
scores to HUI2 and HUI3 utility scores: group- and individual level comparisons. Soc


TABLE 2.1: SOURCE OF PREFERENCES USED FOR QALY WEIGHTS IN ECONOMIC EVALUATIONS OF RA

<table>
<thead>
<tr>
<th>Preference-Technique</th>
<th>DMARD</th>
<th>NSAID</th>
<th>Decision-</th>
<th>Clinical or</th>
<th>Reference #</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>analysis</td>
<td>Observational</td>
<td></td>
</tr>
<tr>
<td>Direct</td>
<td></td>
<td></td>
<td></td>
<td>Trial</td>
<td></td>
</tr>
<tr>
<td>RS</td>
<td>✓✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>67,68,71,79</td>
</tr>
<tr>
<td>TTO</td>
<td>✓✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>77,80</td>
</tr>
<tr>
<td>SG</td>
<td>✓✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>67,68,79,80</td>
</tr>
<tr>
<td>Indirect</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EQ-5D</td>
<td>✓✓✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>72,73,74</td>
</tr>
<tr>
<td>SF-6D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUI2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUI3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“DMARD” refers to a study examining the cost-effectiveness of a disease-modifying antirheumatic drug; “NSAID” refers to a study examining the cost-effectiveness of a traditional NSAID or COX-2 specific inhibitor; “Decision analysis” refers to the methodology used to perform the economic evaluation; “Clinical or Observational Trial” refers to whether the economic evaluation was conducted alongside a clinical or observation trial; “RS” = rating scale; “TTO” = time tradeoff; “SG” = standard gamble; “EQ-5D” = EuroQol index score; SF-6D = Short-Form 6D index score; HUI2 = Health Utilities Index Mark 2 index score; HUI3 = Health Utilities Index Mark 3 index score.
CHAPTER 3

A COMPARISON OF FOUR INDIRECT METHODS OF ASSESSING UTILITY VALUES IN RHEUMATOID ARTHRITIS

3.1 FOREWORD

This chapter is a cross-sectional comparison of four indirect utility instruments (HUI2, HUI3, SF-6D, EQ-5D) in a sample of patients with rheumatoid arthritis. The content of this chapter has been accepted for publication in Medical Care.

The candidate is first author on this manuscript, developed the hypotheses, entered and manipulated the data, performed the statistical analyses, and wrote the final manuscript. Co-authors of the study included Daphne Guh, a statistician, Drs. Andy Chalmers and Barry Koehler, rheumatologists who participated in the recruitment of patients, Dr. John Brazier, the developer of SF-6D and Drs. Anis, Esdaile, and Kopec, members of the supervisory committee.

3.2 INTRODUCTION

To integrate quality of life into economic analyses, the effectiveness of a health intervention is measured using a metric known as “utility” where values of zero and 1.0 equal death and perfect health, respectively (some measures permit values less than zero for health states ranked worse than death). Utilities are used to calculate quality-adjusted life years (QALY) gained by adjusting survival by the average utility weight derived from the outcome of that health intervention. Cost per QALY gained is a unique and preferred
measure of the economic value of different interventions, because it permits comparison both within and across disease groups, thereby facilitating funding allocation decisions.\(^1\)

A variety of methods exist for measuring health-related quality of life.\(^2\) However, in order to integrate such measures into an economic evaluation, the common approach is to use QALYs as the outcome and preference-based assessments (often referred to as "utilities") as the source of weightings to assign quality to life-years.\(^1\) The use of a pre-scaled index is often the most convenient and least expensive means of achieving this approach. While no validated index is available for economic evaluations specifically in musculoskeletal disease, several generic preference measures appear suitable for adaptation to economic evaluations in RA.\(^3,4\) Examples of these instruments include the Health Utilities Index 2 and 3 (HUI2 and HUI3), the Short Form 6D (SF-6D), and the EuroQol (EQ-5D). The major characteristics of these instruments have been summarized in Table 3.1 and comprehensive reviews of these instruments are available elsewhere.\(^1,5\) It is important to point out that there is no "gold standard" among these instruments and each likely has its own advantages and disadvantages.

Although a few studies have examined the appropriateness of individual indirect utility instruments specifically in RA,\(^6-9\) no study has directly compared these measures in the same RA population. However, Conner-Spady et al. recently reported on the interchangeability of preference-based instruments (the EQ-5D, the SF-6D, and the HUI3) in providing weights for QALYs.\(^10\) Specifically, they compared the global utility scores in a sample of 161 patients with five different musculoskeletal conditions.

With the increased popularity of economic evaluations of new therapies and programmes, the impact of the choice of utility measure to use in the weightings of QALYs
is uncertain. It is important to evaluate these instruments in terms of their agreement and also to identify specific deficits of preference-based measures in RA.

Thus, our primary objectives of this study were 1) to compare the global utility scores from the HUI2, HUI3, EQ-5D and the SF-6D at both a sample level and within individuals in a clinically heterogeneous sample of RA patients; and 2) to determine the extent to which global utility scores from the indirect utility assessment instruments were representative of dimensions of health status measured in a sample of RA patients.

3.3 METHODS

Three hundred and thirteen individuals participated in the study. In order to be included in the study, subjects had to have a rheumatologist-confirmed diagnosis of RA (as defined by the American College of Rheumatology diagnostic criteria), receive rheumatology care within the province of British Columbia in one of the urban study areas (Vancouver, Richmond) or one of the rural study areas (Vernon and Penticton), consent to answering the questionnaires and be sufficiently proficient in English to answer the questionnaires. Recruitment of RA patients began in October 2001 and ended in September 2002. Ethical approval for this study was obtained through the University of British Columbia’s Behavioural Ethics Committee and informed consent was obtained from each of the participants.

Eight private practice rheumatologists’ offices from the study areas referred subjects as part of their interactions in routine clinical practice. In addition, two of these rheumatologists’ practices sent letters to all of their patients with RA inviting them to participate in the survey. All patient questionnaires were self-administered, self-completed
and submitted via mail. The study physicians' offices supplied additional information from the patients' health record.

### 3.3.1. Measures

#### 3.3.1.1 Clinical

Participants were asked questions regarding their RA and medication history (including recent adverse events). Other self-reported clinical variables included swollen and tender joint count (using the mannequin-based 42 joint count methodology), a 10 cm pain visual analogue scale (VAS) and patient global assessment of disease activity (10 cm VAS). Erythrocyte sedimentation rate (ESR) values closest to the date of completion of the questionnaire (within 1 month) were extracted from the patient's chart for those patients whose rheumatologist used this measure for patient monitoring. In addition, the attending rheumatologists were asked to complete a physician global assessment of disease activity (10 cm VAS) for each patient.

#### 3.3.1.2 Questionnaires

Respondents self-completed three questionnaires allowing for the scoring of four indirect utility assessment instruments (the HUI2, the HUI3, the EQ-5D, and the SF-6D).

#### 3.3.1.3 Hypotheses

Since all the instruments purport to measure the same construct (namely, a global utility value), then, in theory, the values obtained with the different instruments should agree within individual subjects. However, since each instrument has been constructed (in terms of
domains assessed) and valued in different ways, we hypothesized that there would be significant differences between the instruments. In addition, we hypothesized that the global utility scores achieved with the different MAUT instruments would be represented by different dimensions of health status.

3.3.2. Data Analysis

Descriptive statistics were used to characterize the study sample. Repeated measures ANOVA was used to compare global utility values across instruments with Bonferroni’s correction to adjust for multiple comparisons between instruments. Due to the skewed nature of the distribution of some of the instrument’s global utility scores, nonparametric tests were also applied. Since the results using both approaches agreed, only the results of the parametric tests are reported here. Agreement among the utility scores obtained from the four instruments was assessed using the Intraclass Correlation Coefficient (ICC) with a two-way mixed effect model such that the subject effect is random and the measure effect is fixed.

Bland-Altman plots were used to examine patterns of inter-instrument agreement between every possible combination of instruments. These plots are useful to reveal a relationship between the differences and the averages to look for any systematic bias and to identify possible outliers. Since, in theory, these instruments should have perfect agreement as they are attempting to measure the same global utility, the difference scores should be randomly distributed closely around the zero line. By convention, if 95% of the differences fall within zero ± 1.96 times the standard deviation of the mean difference and are not interpreted to be clinically important, the two methods may be used interchangeably.
Minimally important differences (MID) in the utility values obtained by the MAUT instruments are thought to be between 0.03 and 0.07.\textsuperscript{9,15,16}

Exploratory factor analysis was utilized to identify the dimensions of health measured by the questions and to determine if their similar domains loaded into identified dimensions. Raw answers for all of the questions from the SF-6D, the EQ-5D and the HUI2/3 questionnaire were utilized for the factor analysis. Due to the data's skewed, categorical nature, techniques based on polychoric correlations were used. Based on the nature of the instruments, it was expected that the factor analysis would identify the following dimensions: functional ability, pain, cognition, hearing/vision/sensation, and mental/emotional health. Unweighted least squares was utilized as the estimation method for factor extraction. Promax rotation was used as it allows factors to be reasonably correlated. The criteria used to determine the appropriate number of factors consisted of the scree test, an eigenvalue greater than one, and the presence of residuals greater than 0.05 between the observed and the reproduced correlation matrices and the overall interpretation of the solution.\textsuperscript{17} The rotated factor pattern matrix, which represents a matrix that is uncontaminated by overlap among factors, was selected for interpretation of the factor solution.\textsuperscript{17}

By adapting the methods of Richardson and Zumbo\textsuperscript{18} to determine the extent to which the relative proportion of variation in the global utility scores were explained by the factor scores produced by the exploratory factor analysis, each global utility score was regressed onto the saved factor scores. To determine the relative contribution of each explanatory variable (i.e., each factor) to the regression equation, a relative Pratt index score was generated. This index quantifies the contribution each independent variable makes to the overall regression equation by partitioning the model $R^2$ into the proportion attributable to
each independent variable.\textsuperscript{19} One should note that the index is based on a geometric layout of regression and does not make any assumptions regarding the distribution of the variables. In using the modified Pratt Index, we were able to determine which aspects of health were relevant to RA patients in the different overall utility scores.

3.4 RESULTS

Three hundred and thirteen (245 female) respondents with confirmed RA completed the baseline questionnaire. One hundred and ninety seven (63\%) patients were recruited directly by the study rheumatologists whereas 116 were recruited via the mail survey. The completion rates differed for direct recruitment (91\%) compared to mail recruitment (38\%) after accounting for invalid mailings (address problems (n=69), had died (n=6), reported that they did not have RA (n=3), and already recruited (n=1)). The final sample represented a clinically heterogeneous cohort of patients with RA (Table 3.2). A more detailed description of our cohort is available elsewhere.\textsuperscript{20}

3.4.1 Comparison of Utility Scores

Summaries of the instrument utility scores are presented in Table 3.3. There were few missing values in our sample for any of the four instruments. The distributions of the utility values obtained from the four instruments were markedly different (Figure 3.1). The HUI3 global utility score was the lowest of the four instruments for 151 (50\%) of the participants followed by the SF-6D utility score in 96 participants (32\%). The HUI2 global utility score was the highest score in 141 participants (47\%) followed by the EQ-5D score in 92 (30\%) of participants. Sixteen participants (5\%) scored negative values on at least one measure with
the HUI3 global utility score being negative in 5% (n=15) of participants and the EQ-5D utility index in 1% (n=2). Both were negative in only one participant.

For the SF-6D, a total of 223 health states were defined (the most out of any of the instruments). The most common SF-6D health vector was '121212' (n=9, 2.9%), followed by '523434' (n=6, 1.9%) and '323323' (n=5, 1.6%). No participant indicated no problems ('111111') or the worst health state ('645655'). A total of 35 different health states were described by the EQ-5D health profile (vectors). The most frequent health vector in the sample for the EQ-5D was '21221' with 51 (16.2%) indicating this response (some problems in mobility, usual activities and moderate pain/discomfort with no problems in self-care or anxiety/depression). The other most common vectors were '11121' (n=39, 12.4%), '21222' (n=27, 8.6%), '22222' and '22221' (both n=26, 8.3%). Twenty-three (7.3%) indicated no problems ('11111') whereas the worst health state in the sample was '33331' (n=1, 0.3%).

Using the HUI2, a total of 136 health states were defined. The most common health states were '211112' (n=31, 10.2%) where respondents indicated that they required equipment to hear or see or speak and occasional pain without disruption of normal activities followed by '111112' (n=12, 3.9%). For the HUI3, a total of 217 health states were defined. Four (1%) respondents indicated no problems ('1111111') and 1% (n=5) indicated no problems except for mild or moderate pain that prevented no activities ('1111112'). The most common health state vectors were '21111112' (n=14, 4.6%) indicating normal vision with glasses and mild or moderate pain preventing no activities and '21112112' (n=13, 4.2%) indicating normal vision with glasses, mild limitations in the use of hands and fingers and mild or moderate pain that prevented no activities. The worst health state for the HUI3 system was '51144215' (n=1, 0.3%).
Using repeated measures ANOVA, there were significant differences among the utility scores obtained by the four instruments within individuals. In examining differences utility scores obtained from the different methods, all comparisons were significant (p<0.005).

3.4.2 Analysis of Agreement

The ICC for all of the measures (HUI2, HUI3, EQ-5D, SF-6D) was 0.67 (95% C.I. 0.62 to 0.71). The pairwise ICC values are summarized in Table 3.4. The Bland-Altman plots are presented in Figures 3.2-3.7. In general, for all of the plots, there appeared to be more agreement in the higher utility values compared to the lower utility values across instruments. All plots reveal that a substantial proportion of the observations in the lower utility values fall outside the area of zero ± 1.96 times the standard deviation of the differences and most of these differences between any two instruments exceed the MID.

3.4.3 Exploratory Factor Analysis

Bartlett’s Test of Sphericity was used to determine whether the correlation matrix differed from the identity matrix. The results of this test supported the use of factor analysis (p < 0.0001). There were five factors with eigenvalues greater than one; thus, a factor analysis extracting these factors was performed. This analysis accounted for 74% of the variation in the original variables.

Five factors were extracted and interpreted as follows (with strongest loading indicators in brackets): 1) Factor 1 = Physical Functioning (pain mobility/ambulation, self-
care, dexterity, usual activity, physical health, role limitations, social limitations); 2) Factor 2 = Emotional/Mental Health – (anxiety, happiness, and mental health); 3) Factor 3 = Speech (speaking); 4) Factor 4 = Cognition – (cognition, hearing); and 5) Factor 5 = Vision (vision). The rotated factor pattern matrix showing the individual loadings from the raw questions is shown in Table 3.5. The factor correlations are shown in Table 3.6. With the exception of physical functioning/pain and emotional/mental health, there were no moderate or strong correlations among the factors.

For multiple linear regressions of the global utility scores (dependent variable), the results revealed that the functional ability/pain factor score contributed the most in explaining the variance in all of the global utility scores. However, for the HUI2/3, cognition factor scores explained a large proportion of the variance in the global utility scores whereas the emotional/mental factor scores explained a large component of the variance in the SF-6D and EQ-5D utility scores (Table 3.7).

3.5 DISCUSSION

This is the first study to report on the results of administering and comparing the four instruments in a relatively large sample of participants with RA. The low number of missing values for each of the instruments attests to the fact that they are suitable to be self-administered in this type of cohort. There were significant differences in utility scores between instruments. In addition, the level of agreement was much lower than theoretically postulated (i.e. an ICC approaching 1.0) with ICC values ranging from 0.56 to 0.79. From the Bland-Altman plots, agreement was much poorer at lower utility values than higher utility values. Also, it would appear that the global utility scores from the various systems are
mostly measuring physical functioning/pain; however, the HUI2/3 are also measuring cognition while the SF-6D and EQ-5D are also measuring emotional/mental health. This finding is not surprising considering that the HUI systems were created to be “within the skin” measures; that is, they are primarily concerned with impairment rather than disability or handicap. On the other hand, the SF-6D is based on the psychometric instrument, the SF-36. As such, it is a measure of handicap and includes social and role functioning which are not assessed by the other instruments.

The SF-6D, due to its high lower boundary of +0.30 does not provide a wide range of utility values. This phenomenon is illustrated in the Bland-Altman plot comparing the global utilities from the SF-6D to the EQ-5D. However, this observation is not unexpected as Brazier stated that this instrument might be most appropriate in “groups experiencing mild to moderate health problems and in those expected to experience comparatively small changes”. Brazier also states that one of the potential advantages of the SF-6D is the much larger size of its predictive system as compared to the EQ-5D. This appears to be the case as there were 188 more health states defined using the SF-6D compared to the EQ-5D system within our cohort. This finding may impart a greater degree of sensitivity to change that this instrument can detect in longitudinal studies especially in those in mildly or moderately impaired.

The EQ-5D appears to have significant ceiling effects with 21% of the study participants reporting either no problems on all the domains or some problems on one domain only (with the remainder having no problems). More individuals reported having no problems under the EQ-5D classification system than reporting the lowest level of deficit on the SF-6D system. In addition, due to the limited number of health states that were reported
by the sample, there was a lack of variability in the responses to the five questions. This lack of variation in the responses and low number of possible descriptive health states may impede the sensitivity of the EQ-5D in longitudinal studies when compared to the other instruments.

The HUI2 and HUI3 displayed superior agreement to the other instruments as they had the highest pair-wise ICC values. However, as shown by the Bland-Altman plot comparing these two instruments, the global utility values defined by them tend to be quite different especially at the lower end of the utility scale. Both of these systems appear to define sufficient health states to enable them to discriminate small changes between patients or within patients over time, which is in agreement with previous research.

There have been comparisons between instruments in other disease states. For example, there have been comparisons between the HUI2 and HUI3 in Alzheimer’s disease, the HUI3 and the SF-6D in patients at an elevated risk of sudden cardiac death, and the SF-6D, HUI3, EQ-5D, the Assessment of Quality of Life Questionnaire (AQoL), and the Finnish 15D in a sample of Australian Residents. Generally, the results of these studies agreed with ours in that there were significantly different values achieved both between instruments and within individuals using the different instruments.

There are several limitations to this study. First of all, the scoring functions for both the EQ-5D and SF-6D global utility values were derived from samples of the UK population which may differ from preferences given by those in Canada. Secondly, although the SF-6D and the HUI systems utilized the standard gamble as a valuation technique for health states described by their attributes, there are several differences with how the health states
are valued. For example, the SF-6D health states were valued directly, whereas the HUI health states were directly valued by a rating scale that was transformed (using a power function) into a standard gamble utility. The EQ-5D health states were valued using the time trade-off technique. Also, the HUI systems utilize multiattribute utility theory and a multiplicative model of scoring whereas the SF-6D and EQ-5D use empiric methods and additive scoring models. Thus, differences between the health utilities obtained may be confounded by the utility evaluation and valuation methodology. In addition, we have studied a single cohort with a specific disease state thus limiting the generalizability to other diseases.

As we only analyzed cross-sectional data, it is not clear if changes in the utility values obtained from the MAUT instruments would be similar. If these changes are similar across instruments, O'Brien et al. postulate that this would increase the comfort level for inter-study comparability of QALYs. If these changes are similar across instruments, O'Brien et al. postulate that this would increase the comfort level for inter-study comparability of QALYs. Further research examining the longitudinal validity (responsiveness and sensitivity to change) of these MAUT instruments is presently ongoing. The results from this research will further guide which of these instruments is the most appropriate to utilize in the assessment of RA-related outcomes. A recent study comparing the MAUT instruments in a sample from the Australian general population recommended that researchers should select a MAUT instrument that is sensitive to the health states that are being investigated.

Although it is good practice in quality of life research to randomize the order of administration of questionnaires to enable the evaluation of an “order effect”, we did not randomize the sequence of questionnaires. In piloting of the survey in the first 25 patients, we found that patients were not completing the questionnaires in the order that they were
presented or another identifiable systematic pattern. Thus, a formal assessment of an “ordering effect” was not feasible.

In conclusion, the utility values obtained from the four MAUT instruments were statistically and clinically different. The SF-6D is bounded by a relatively high floor (+0.30) and the EQ-5D displayed important ceiling effects that could limit their ability to detect change in patients at the higher or lower limits of the utility scale. The HUI2 and HUI3 did not appear to suffer from these limitations. It is unlikely that the utility values from the MAUT instruments tested, if used as the weightings for QALYs in studies examining RA, would result in comparable estimates.
3.6 REFERENCES


<table>
<thead>
<tr>
<th>Dimensions/Domains/Attributes</th>
<th># of Possible Health States</th>
<th>Valuation Technique</th>
<th>Boundaries</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HUI2</strong> Sensation (vision, hearing, speech), Mobility, Emotion Cognition, Self-care, Pain</td>
<td>24,000</td>
<td>Standard Gamble</td>
<td>-0.03 – 1.00</td>
</tr>
<tr>
<td><strong>HUI3</strong> Vision, Hearing, Speech, Ambulation, Dexterity, Emotion, Cognition, Pain</td>
<td>972,000</td>
<td>Standard Gamble</td>
<td>-0.36 – 1.00</td>
</tr>
<tr>
<td><strong>SF-6D</strong> Physical Function, Role Limitation, Social Function, Pain, Mental Health, Vitality</td>
<td>18,000</td>
<td>Standard Gamble</td>
<td>0.30 – 1.00</td>
</tr>
<tr>
<td><strong>EQ-5D</strong> Mobility, Usual Activities, Self-Care, Pain, Anxiety</td>
<td>243</td>
<td>Time Trade Off</td>
<td>-0.59 – 1.00</td>
</tr>
</tbody>
</table>
TABLE 3.2: CLINICAL CHARACTERISTICS OF THE STUDY PARTICIPANTS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>SD*</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (yrs)</td>
<td>61.5 (25.9)</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESR</td>
<td>24.52 (21.02)</td>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAQoL (range 0 - 30)</td>
<td>12.82 (8.28)</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RA Duration (yrs)</td>
<td>13.87 (11.41)</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAQ</td>
<td>1.10 (0.77)</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pt. Global Assessment</td>
<td>59.82 (25.86)</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD Global Assessment</td>
<td>20.88 (23.39)</td>
<td>23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain VAS</td>
<td>43.12 (27.02)</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EQ-5D Health Thermometer</td>
<td>65.02 (19.27)</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tender Joint Count</td>
<td>15.09 (11.99)</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swollen Joint Count</td>
<td>9.14 (9.66)</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-Reported RA Severity, n %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Mild</td>
<td>9</td>
<td>2.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mild</td>
<td>34</td>
<td>10.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>120</td>
<td>38.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe</td>
<td>110</td>
<td>35.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Severe</td>
<td>27</td>
<td>8.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>13</td>
<td>4.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-Reported RA Control, n %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Well Controlled</td>
<td>33</td>
<td>10.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Well Controlled</td>
<td>76</td>
<td>24.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adequately Controlled</td>
<td>123</td>
<td>39.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Well Controlled</td>
<td>61</td>
<td>19.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Controlled At All</td>
<td>7</td>
<td>2.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>13</td>
<td>4.2%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* SD = Standard Deviation
### TABLE 3.3: OVERALL MEAN AND MEDIAN UTILITY SCORES FROM THE INSTRUMENTS IN THE SAMPLE OF RA PATIENTS

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Mean*</th>
<th>Lower</th>
<th>Upper</th>
<th>Median</th>
<th>IQR*</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF-6D</td>
<td>0.63</td>
<td>0.61</td>
<td>0.64</td>
<td>0.60</td>
<td>0.12</td>
<td>302</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>0.66</td>
<td>0.63</td>
<td>0.69</td>
<td>0.74</td>
<td>0.19</td>
<td>308</td>
</tr>
<tr>
<td>HUI2</td>
<td>0.71</td>
<td>0.69</td>
<td>0.73</td>
<td>0.75</td>
<td>0.28</td>
<td>304</td>
</tr>
<tr>
<td>HUI3</td>
<td>0.53</td>
<td>0.50</td>
<td>0.57</td>
<td>0.56</td>
<td>0.44</td>
<td>303</td>
</tr>
</tbody>
</table>

* IQR = interquartile range

* P-value <0.0001 for comparison of mean utility scores using repeated measures ANOVA (based on n= 302)

* P<0.005 for all pairwise comparisons with Bonferroni correction

* IQR = interquartile range

* P-value <0.0001 for comparison of mean utility scores using repeated measures ANOVA (based on n= 302)

* P<0.005 for all pairwise comparisons with Bonferroni correction
## Table 3.4: Intraclass Correlations and 95% Confidence Intervals Between Instruments

<table>
<thead>
<tr>
<th>Instrument</th>
<th>HUI2</th>
<th>HUI3</th>
<th>SF-6D</th>
<th>EQ-5D</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUI2</td>
<td>1.00</td>
<td>0.79</td>
<td>0.66</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.74-0.83)</td>
<td>(0.54-0.72)</td>
<td>(0.61-0.74)</td>
</tr>
<tr>
<td>HUI3</td>
<td></td>
<td>1.00</td>
<td>0.56</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.48-0.64)</td>
<td>(0.59-0.72)</td>
</tr>
<tr>
<td>SF-6D</td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.51-0.66)</td>
</tr>
<tr>
<td>EQ-5D</td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>
TABLE 3.5: ROTATED FACTOR PATTERN MATRIX

<table>
<thead>
<tr>
<th>Indicator (Questionnaire)*</th>
<th>Factor</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain (SF-6D)</td>
<td></td>
<td>0.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobility (EQ-5D)</td>
<td></td>
<td>0.86</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-care (EQ-5D)</td>
<td></td>
<td>0.86</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain (HUI2/3)</td>
<td></td>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain (EQ-5D)</td>
<td></td>
<td>0.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical Functioning (SF-6D)</td>
<td></td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain with medications (HUI2/3)</td>
<td></td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambulation (HUI2/3)</td>
<td></td>
<td>0.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-care (HUI2/3)</td>
<td></td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Limitations (SF-6D)</td>
<td></td>
<td>0.73</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dexterity (HUI2/3)</td>
<td></td>
<td>0.68</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vitality (SF-6D)</td>
<td></td>
<td>0.62</td>
<td>0.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Role Limitations</td>
<td></td>
<td>0.60</td>
<td>0.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotion (HUI2/3)</td>
<td></td>
<td></td>
<td></td>
<td>0.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety (EQ-5D)</td>
<td></td>
<td></td>
<td></td>
<td>0.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mental Health (SF-6D)</td>
<td></td>
<td></td>
<td></td>
<td>0.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happiness (HUI2/3)</td>
<td></td>
<td></td>
<td></td>
<td>0.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaking to those you know (HUI2/3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.84</td>
</tr>
<tr>
<td>Speaking to those you don’t know (HUI2/3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.93</td>
</tr>
<tr>
<td>Memory (HUI2/3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.37</td>
<td>0.72</td>
</tr>
<tr>
<td>Think/Solve Problems (HUI2/3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.51</td>
<td>0.70</td>
</tr>
<tr>
<td>Hearing (HUI2/3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.63</td>
</tr>
<tr>
<td>Vision – Reading (HUI2/3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vision – Recognizing a friend (HUI2/3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Indicators with loadings less than 0.30 have been omitted to improve interpretability
### TABLE 3.6: FACTOR CORRELATION MATRIX

<table>
<thead>
<tr>
<th>Factor</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Functioning/Pain</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotional/Mental Health</td>
<td>-0.32</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech</td>
<td>-0.08</td>
<td>-0.16</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognition</td>
<td>-0.15</td>
<td>-0.10</td>
<td>-0.15</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Vision</td>
<td>-0.15</td>
<td>-0.05</td>
<td>-0.07</td>
<td>-0.04</td>
<td>1.00</td>
</tr>
</tbody>
</table>
TABLE 3.7: RELATIVE PRATT INDEX SCORES ASSESSING RELATIVE CONTRIBUTION OF EACH FACTOR TO THE MODEL'S ADJUSTED $R^2$

<table>
<thead>
<tr>
<th>Instrument, Factor</th>
<th>Pearson Correlation</th>
<th>Standardized Beta Weight</th>
<th>Relative Pratt Index Score*</th>
<th>Model Adjusted R-square</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUI2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical functioning/Pain</td>
<td>-0.82</td>
<td>-0.75</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td>Cognition</td>
<td>-0.50</td>
<td>-0.33</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>HUI3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical Functioning/Pain</td>
<td>-0.86</td>
<td>-0.77</td>
<td>0.77</td>
<td>0.86</td>
</tr>
<tr>
<td>Cognition</td>
<td>-0.53</td>
<td>-0.36</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>SF-6D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical Functioning/Pain</td>
<td>-0.90</td>
<td>-0.77</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>Emotional/Mental Health</td>
<td>-0.68</td>
<td>-0.21</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>EQ-5D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical Functioning/Pain</td>
<td>-0.63</td>
<td>-0.71</td>
<td>0.87</td>
<td>0.51</td>
</tr>
<tr>
<td>Emotional/Mental Health</td>
<td>-0.55</td>
<td>-0.13</td>
<td>0.13</td>
<td></td>
</tr>
</tbody>
</table>

* The Relative Pratt Index represents the proportion of the model R-square that is explained by the variable.
FIGURE 3.1: DISTRIBUTIONS OF GLOBAL UTILITY VALUES ACROSS THE MAUT INSTRUMENTS
EQ-5D

Global Utility Values

SF-6D

Global Utility Values
FIGURE 3.2: BLAND-ALTMAN PLOT OF DIFFERENCE BETWEEN THE HUI2 AND HUI3 VS. THE AVERAGE SCORE WITHIN PATIENTS

The dotted lines represent ±1.96 times the standard deviation around the difference. The clustering of the data on the right around zero compared to the on the left, indicates that the two utility scores have better agreement in those with higher values and that, at lower values, the HUI2 yields much higher scores than the HUI3.
The dotted lines represent ±1.96 times the standard deviation around the difference. The pattern of observations reveals that for higher utility values, the SF-6D tends to have lower scores than the HUI3 whereas the converse is true at lower utility values (which is expected since the SF-6D is bounded at +0.30).
FIGURE 3.4: BLAND-ALTMAN PLOT OF DIFFERENCE BETWEEN THE HUI3 AND EQ-5D VS. THE AVERAGE SCORE OF THESE TWO INSTRUMENTS WITHIN PATIENTS

The dotted lines represent ±1.96 times the standard deviation around the difference. From the pattern of the points, it appears that the EQ-5D is lower than the HUI3 at higher values but this relationship reverses at lower values. Also, the odd linear patterns of the points are due to the gaps in the EQ-5D scoring system.
FIGURE 3.5: BLAND-ALTMAN PLOT OF THE DIFFERENCE BETWEEN THE EQ-5D AND SF-6D VS. THE AVERAGE SCORE WITHIN PATIENTS

The dotted lines represent ±1.96 times the standard deviation around the difference. There would appear to be higher agreement on the right compared to the left indicating that the two utility scores have better agreement in those with higher values and that, at lower values, the SF-6D yields much higher scores than the EQ-5D.
FIGURE 3.6: BLAND-ALTMAN PLOT OF DIFFERENCE BETWEEN THE HUI2 AND THE SF-6D VS. THE AVERAGE SCORE WITHIN PATIENTS

The dotted lines represent ±1.96 times the standard deviation around the difference. The pattern of observations reveals that for higher utility values, the SF-6D tends to have lower scores than the HUI2 whereas the converse is true at lower utility values (which is expected since the SF-6D is bounded at +0.30).
FIGURE 3.7: BLAND-ALTMAN PLOT OF DIFFERENCE BETWEEN THE HUI2 AND EQ-5D VS. THE AVERAGE SCORE OF THESE TWO INSTRUMENTS WITHIN PATIENTS

The dotted lines represent ±1.96 times the standard deviation around the difference. Also, the odd linear patterns of the points are due to the gaps in the EQ-5D scoring system.
CHAPTER 4

A COMPARISON OF GENERIC, INDIRECT UTILITY MEASURES (THE HUI2, HUI3, SF-6D, AND THE EQ-5D) AND DISEASE-SPECIFIC INSTRUMENTS (THE RAQOL AND THE HAQ) IN RHEUMATOID ARTHRITIS

4.1 FOREWORD

This chapter is currently under second review, under the same title, in the journal Social Science and Medicine. The candidate is first author of this manuscript which is co-authored by John Woolcott, a health economist, Drs. Kam Shojania and Robert Offer, clinical rheumatologists who assisted with patient recruitment, Dr. John Brazier, the developer of the SF-6D and by Drs. Aslam Anis, John Esdaile and Jacek Kopec, members of the candidate’s committee. The candidate’s role in this manuscript was the development of the primary hypothesis and methods, data entry and manipulation, statistical analysis, and writing of the final manuscript.

4.2 INTRODUCTION

Rheumatoid arthritis (RA) is a chronic, progressive disease that places a substantial burden on those afflicted and their families. Specifically, the disease itself, its treatments, and complications arising from both, result in detrimental effects on many areas of life including physical, psychological, and social functioning.1 Yet many clinical measures do not adequately capture the overall impact of the disease on individuals. Furthermore, because
the ultimate goal in any therapeutic intervention in RA is to improve health-related quality of life (HRQL), it is important to measure this outcome accurately.

The two basic categories of instruments used to measure HRQL are generic instruments and disease-specific instruments. Utility, or preference-based, measures are an example of generic instruments that are derived from decision and utility theories. As such, preference-based approaches integrate different aspects of health into a single index anchored by a value of ‘1.00’ for full health and ‘0’ for death (health states considered worse than death can be represented by negative values). In turn, these measures are used in economic evaluations to integrate survival and HRQL into a single metric, the quality adjusted life year (QALY).

Preferences for health outcomes can be considered to be utility values when they are choice-based response methods that are framed under uncertainty. Health utilities can be either measured directly (using trade-off techniques such as the standard gamble or time trade-off) or indirectly (using multidimensional HRQL questionnaires developed using multi-attribute utility theory [MAUT]) such as the Health Utilities Index 2 and 3 (HUI2 and HUI3), the Short Form 6D (SF-6D), and the EuroQol (EQ-5D). Due to their ease of administration, these indirect measures are commonly used as the source of utility weightings in economic evaluations. A brief overview of these instruments has been provided in Table 4.1 and comprehensive reviews are available.

Disease-specific measures are commonly utilized to assess HRQL in RA. Specifically, the Health Assessment Questionnaire (HAQ) Disability Index is a commonly used disease-specific measure whereas the Rheumatoid Arthritis Quality of Life (RAQoL) questionnaire is a newly developed instrument. The HAQ was originally developed as one
of the first self-report, functional status (disability) measures and has become one of the dominant instruments in musculoskeletal diseases including rheumatoid arthritis. The HAQ has been utilized to assess disability for approximately two decades and is a mandated outcome for clinical trials in RA. The RAQoL is the first patient-completed instrument specifically designed for use with RA patients. It was derived directly from qualitative interviews with relevant patients and considers aspects of many areas of life that have been affected detrimentally by RA. The goal of the RAQoL is to be a comprehensive, disease-specific scale that will be more responsive to change than previous scales used in RA.

When comparing disease-specific to generic measures of HRQL, disease-specific measures focus on the particular problems that are often unique to the disease that they are developed to assess. As such, these measures may have greater ability to measure functional impairments resulting from the disease and detect smaller changes in health relative to generic measures. However, generic measures permit comparisons across disease states which may provide useful data for health policy and resource allocation decision-making.

However, there is agreement in the literature that any instrument utilized to assess HRQL needs to be valid and reliable and the most rigorous approach to establishing validity is construct validity. Generally, validity can be defined as the extent to which an instrument measures the property that it is intended to measure. Construct validity is an assessment of the extent to which the scores of an instrument correlate with other hypothesized measures or indicators of the health concept or concepts of interest.

In a recent paper by Brazier et al., an argument is made against traditional means to assess construct validity for preference-based measures. One method of assessing construct validity is to test a measure’s ability to discriminate between groups hypothesized to differ in
terms of health. However, Brazier et al. stated concern that the aspect of health used to subdivide groups might not reflect preferences (for example, age). To rectify these problems, Brazier et al.\textsuperscript{12} make a case for empirical validity (a form of construct validity) to be the "acid test" for preference-based measures. Specifically, they propose that a preference-based instrument should generate values that reflect people's preferences and supply a hierarchy of evidence of how to determine this: revealed preference data, stated preference data and hypothesized preferences. The authors state that when revealed preference data and stated preference data are lacking (which is common in health care), hypothesized preference data can be used.

Hypothesized preferences are very similar to construct validity in that the researcher must hypothesize or construct the expected difference. Provided that care is taken in examining health states where clear preference differences would exist, the hypothesized preference approach should yield valid results. A method of assessing this is to examine whether the global utility scores of an instrument reproduce the expected differences between groups of patients. Thus, as long as care is taken in selecting the groups such that preferences for health states would be expected to differ, this would be an appropriate method to assess the construct validity of preference-based instruments.

There remains a gap in the literature regarding the assessment of construct validity for both global and single-attribute indirect measures of health utility. Although construct validity has been investigated for the HUI2 and HUI3 in Type 2 diabetes,\textsuperscript{13} the HUI3 in self-reported stroke and arthritis,\textsuperscript{14} the HUI2 and HUI3 for Alzheimer's disease,\textsuperscript{15} the HUI3 and the EQ-5D in intermittent claudication,\textsuperscript{16} and the EQ-5D in rheumatoid arthritis,\textsuperscript{17} there are little comparative data across all four instruments in the same patient sample with a well,
delineated, chronic disease. In addition, there are no data that compare the construct validity of disease-specific instruments like the Health Assessment Questionnaire (HAQ) and the Rheumatoid Arthritis Quality of Life Questionnaire (RAQoL). Thus, the objectives of this study were to examine the cross-sectional construct validity of the global and single-attribute scores from the indirect utility instruments in terms of their ability to distinguish between subgroups of individuals with different levels of RA severity, compare amongst them and compare them to disease-specific instruments. In addition, for each of the instruments, the minimally important difference was determined and was compared to previously defined values where available.

4.3 METHODS

4.3.1 Sample

Three hundred and thirteen individuals participated in the study. To be included, subjects had to have a rheumatologist-confirmed diagnosis of RA (as defined by the American College of Rheumatology diagnostic criteria), receive rheumatology care within the province of British Columbia in one of the study areas (Vancouver, Richmond, Vernon and Penticton), consent to answer the questionnaires, be sufficiently proficient in English to answer the questionnaires, and be willing to participate in follow-up surveys. Recruitment of RA patients began in October 2001 and ended in September 2002. Ethical approval for this study was obtained through the University of British Columbia's Behavioural Ethics Committee and informed consent was obtained from each of the participants.

Eight private rheumatologists' offices from the study areas referred subjects into the cohort during their interactions in routine clinical practice. In addition, two of these
rheumatologists' practices sent letters to all of their patients with RA inviting them to participate in the survey. All patient questionnaires were self-administered, self-completed and submitted via mail. The study physicians' offices supplied additional information from the patients' health record.

**4.3.2 Measures**

**4.3.2.1 Clinical**

Participants were asked questions regarding their RA and medication history including adverse reactions over the past three months. Other self-reported clinical variables included swollen joint count (SJC) and tender joint count (TJC) (using the mannequin-based 42 joint count methodology),
19 a 10 cm pain visual analogue scale (VAS), a patient global assessment of disease activity (10 cm VAS), and RA severity and RA control (both using a 5 point Likert scale). Erythrocyte sedimentation rate (ESR) values closest to the date of completion of the questionnaire (within 1 month) were extracted from the patient's chart for those patients whose rheumatologist used this measure for patient monitoring. In addition, the attending rheumatologists were asked to complete a physician global assessment of disease activity (10 cm VAS) for each patient.

**4.3.2.2 Questionnaires**

**4.3.2.2.1 Health Assessment Questionnaire (HAQ) Disability Index**

The HAQ is a measure of physical disability that assesses a respondent's ability to complete everyday tasks in areas such as dressing and grooming, rising, eating, walking, personal hygiene, reach, grip and other activities (such as getting into and out of a car). Each
of these areas is assigned a section score that is further adjusted to account for the use of any aids, devices or help from another person. These are then summed and averaged to give an overall score between 0.0 (best possible function) to 3.0 (worst function). A HAQ score difference of 0.25 is said to represent the minimally important difference (MID).\(^{20,21}\)

### 4.3.2.2 Rheumatoid Arthritis Quality of Life Questionnaire (RAQoL)

The RAQoL consists of 30 questions (answered by yes/no) that assess such aspects of RA as moods and emotions, social life, hobbies, everyday tasks, personal and social relationships, and physical contact. The RAQoL is scored by assigning a point for each affirmative response and no points for negative responses. Thus, scores range from 0 (least severity) to 30 (highest severity). To date, the MID for the RAQoL has not been determined.

### 4.3.2.3 Preference Based Measures – MAUT Instruments

The multi-attribute utility theory (MAUT) based instruments used in the questionnaire were the HUI2, HUI3, SF-6D, and the EQ-5D. The major differences between these instruments are outlined in Table 4.1. The MID for the HUI2 and the HUI3 overall utility scores is considered to be 0.03.\(^4\) For the EQ-5D, 0.03 has been postulated to be the MCID, as it is the smallest of the coefficients in the York weights (i.e., the smallest difference in moving from one level to another on any of the 5 dimensions).\(^{22}\) Finally for the SF-6D global utility scores, the MID has been estimated to be 0.033 (95% CI: 0.029 to 0.037) from an analysis examining seven longitudinal studies involving the SF-36.\(^{23}\)

In addition to the overall utility scores, single attribute utility scores can be determined for each of the dimensions that are assessed by HUI2 and HUI3. Although
single-attribute scores are not generally computed for the EQ-5D and SF-6D systems, they can be calculated within the scoring functions by holding all other domains constant at no impairment and determining the score for the domain of interest by utilizing the reported categorical response. While we do not advocate the widespread adoption of this methodology, we utilized this approach in an exploratory fashion to facilitate comparisons between these instruments and the HUI systems. As such, these single attribute scores range from 0.0 to 1.0 and represent functional capacity on each of the dimensions independent of the other attributes in the instrument.

4.3.3 Data Analysis

As an assessment of construct validity of the MAUT instruments global utility scores, the HAQ disability score and the RAQoL, tests of statistical significance were used to determine the ability of each summary score to discriminate between groups of differing disease severity. More severe or advanced RA was defined in terms of self-reported severity and control, recent adverse events to RA drug therapy, hospitalizations due to RA in the past year, other chronic diseases besides RA, absenteeism from work or school in the past year due to RA, and use of allied health/home services and the use of special equipment for RA. For all scales, it was hypothesized that groups with more advanced or severe RA would have lower scores.

The effect size, the standardized mean difference between two groups on a measured outcome, was calculated for each of the dichotomous clinical variables for each instrument. An effect size of 1 indicates a change in magnitude equivalent to one standard deviation. According to Cohen, the absolute value of effect sizes (d) can be categorized as small (d =
Comparing the effect sizes across the different indirect utility and disease-specific instruments allowed for a comparison in the instruments' abilities to discriminate between groups of different disease severity, with a larger effect size indicating better discriminative ability.

To further assess and compare the construct validity among the instruments, relationships between continuous clinical variables and the MAUT global utilities, the single attribute utilities, the HAQ and the RAQoL were assessed with Spearman’s correlations. It was postulated that strong correlations/relationships would exist between the overall scores from all the instruments and measures of RA severity. For the single attribute utility scores, it was postulated that strong correlations/relationships would exist between mobility, self-care and pain (from the HUI2), ambulation, dexterity and pain (from the HUI3), physical functioning, role limitations, pain and vitality (from the SF-6D), mobility, self-care, usual activities, and pain/discomfort (from the EQ-5D) and the continuous measures of RA severity. Due to the skewed nature of the data, non-parametric correlations (Spearman’s rho) were calculated. A Spearman’s rho of > 0.50 or < -0.50 were considered be strong, while values between -0.49 to -0.30 or 0.30 to 0.49 were considered moderate and values between -0.30 and 0.30 were considered to be weak.

According to the methods outlined by Samsa et al., MID values were calculated for each of the MAUT instruments using the calculated effect sizes. In brief, the methodology is as follows: 1) using Cohen’s criteria, the absolute value of the effects under consideration were considered to be small (d=0.20); 2) the standard deviation of the global utility from the instruments were determined (Table 4.2); 3) a preliminary estimate of the MID was determined by multiplying the effect size (0.2) by the standard deviation for each of the
summary scores. Differences in scores for each dichotomous clinical parameter were compared to the MID estimate to determine if they were clinically important. Thus, our hypothesis was that, using the methods of estimating effect size-based clinically important differences (CID) from cross-sectional data, the differences in overall instrument scores between groups of different RA severity, would exceed the minimally important difference (MID).

The HAQ was also utilized as a means to estimate the MID for the overall utility scores. Using simple, ordinary least squares (OLS), linear regression, the MAUT overall utility scores and the RAQoL score (independent variables) were regressed on the HAQ score (dependent variable). Since it is well accepted that the MID for the HAQ is 0.25, the MID for each instrument was estimated from the beta coefficient from the regression model to produce a 0.25 change in HAQ score.

Descriptive statistics were used to characterize the study sample. Parametric tests (t-tests and ANOVA) were used to test for important differences between those with missing data from the MAUT instruments. Both parametric (t-tests and ANOVA) and non-parametric (Mann Whitney tests and Kruskal-Wallis tests) were used due to the skewed nature of the data. However, since the parametric and non-parametric approaches agreed, only the results of the parametric tests are reported.

4.4 RESULTS

4.4.1 Sample

Three hundred and thirteen (245 female) respondents with confirmed RA completed the baseline questionnaire. One hundred and ninety seven (63%) patients were recruited
directly by the study rheumatologists whereas 116 were recruited via mail. The completion rates of the surveys differed according to the method of recruitment. For direct recruitment by a study rheumatologist, 91% completed the baseline questionnaire, whereas for recruitment by mail, there was a 38% completion rate after accounting for invalid mailings (returned due to address problems (n=69), patient had died (n=6), patient did not have RA (n=3), and patient already recruited by a different rheumatologist (n=1)). Those recruited by mail tended to be older (62 vs. 58 years, p=0.01), had RA for a longer period of time (15 vs. 10 years, p=0.0002), had better perceived control of their RA (16% vs. 26% rated as “not well controlled” or “not controlled” at all, p=0.03) and included more females (84% vs. 74%, p=0.03) than those recruited directly by a rheumatologist. The final sample represented a clinically heterogeneous cohort of patients with RA (Table 4.2).

There were few missing values in the HRQL or clinical variables. For the HRQL questionnaires, the lowest completion rate was for the SF-6D with 11 (<4%) missing values. There were no significant differences in demographic or RA characteristics identified between those with complete and missing values.

4.4.2 Description of Global and Single-Attribute Utilities

In Table 4.3, a summary of the results of the multiattribute and single attribute utility values for the MAUT instruments is displayed. In Table 4.4, the specific domain responses for the MAUT instruments are given. A comparison of the distributions of the utility scores is illustrated in Figure 4.1.
4.4.3 Construct Validity

As hypothesized, all of the MAUT instrument global utility scores were lower in groups thought to have higher RA severity and most of these relationships were statistically significant (Tables 4.5 and 4.6). For self-reported disease-severity and control (each with five categories of responses), there was a gradient across all the instruments’ global scores with the highest level of severity/control having the lowest utility and vice versa (Table 4.5). The Spearman correlation coefficients were very similar across both the disease-specific and generic instruments.

Other relationships between dichotomous, disease severity indicators and the summary scores for each of the instruments are shown in Table 4.6. For all of the severity variables, the hypothesized relationship of a better score (a higher global utility for the MAUT instruments and lower scores for the disease-specific instruments) was found to be valid. Of these, 29 of the 36 were significant at p<0.05. For the effect size analysis, 32 out of the 36 calculated effect sizes exceeded Cohen’s low limits of 0.2 (Table 4.6). The HAQ and RAQoL were generally better able to discriminate among the groups of lower and higher severity as indicated by the larger effect sizes. However, all of the MAUT instruments appeared to have discriminative ability with the HUI3 having the largest magnitude in overall differences across the dichotomous severity measures.

For the correlation analysis between the multi-attribute and single-attribute utility scores and the disease severity measures, all the expected correlations were in the hypothesized direction and most were highly significant. Strong correlations were observed consistently with the RAQoL score, the patient global VAS, the HAQ disability score, and
the pain VAS with the MAUT global utility scores (Table 4.7). For the single attribute utility scores that were postulated to be highly correlated, consistent strong correlations existed only with the RAQoL scores and the HAQ disability scores. Generally, with the exception of the pain/discomfort single attribute scores from the EQ-5D, the pain VAS was also strongly correlated with the pain single attribute scores, the mobility single attribute score from the HUI2, and the physical functioning and the role limitations single attribute scores from the SF-6D. The patient global VAS was strongly correlated with the pain single attribute scores from the HUI2, HUI3, and the SF-6D (along with the physical functioning and role limitations single attribute scores from this measure) and the usual activities single attribute score from the EQ-5D. For the disease specific measures, the RAQoL score was strongly correlated with all of the disease severity measures with the exception of RA duration, whereas the HAQ displayed a similar correlation pattern as the global utility scores with strong correlations with the RAQoL, the patient global VAS, and the pain VAS.

Using the effect size methodology to estimate the MID, these values for each of the MAUT instruments were 0.04 for the HUI2, 0.06 for the HUI3, 0.03 for the SF-6D, and 0.05 for the EQ-5D. For the disease-specific measures, the estimated MID was 1.70 for the RAQoL and 0.15 for the HAQ disability index. As it can be seen in Table 4.6, the differences in global scores between naturally occurring groups based on clinical characteristics generally exceeds these MID estimates.

Finally, the results of the simple, linear regression revealed strong associations between the MAUT instruments’ global utility values and the RAQoL (dependent variables) and the HAQ disability index (Table 4.8). In estimating the MID of the MAUT instruments and the RAQoL using the accepted MID of the HAQ and the beta coefficients from the linear
regression, it was found that the MID estimates were in general agreement to those
determined by the effect size methodology (0.04 vs. 0.03 for the HUI2, 0.07 vs. 0.06 for the
HUI3, 0.03 and 0.05 for the SF-6D and the EQ-5D, respectively). For the RAQoL, both
methodologies yielded similar results (1.70 and 2.0).

4.5 DISCUSSION

This is the first study to examine the construct validity of these four generic, MAUT
instruments simultaneously in a relatively large cohort of participants with a single, well-
deﬁned, chronic disease. In addition, it is the first to compare the generic MAUT instruments
to two disease speciﬁc measures (the HAQ and the RAQoL) in their relative abilities to
discriminate across RA severity. Finally, the estimates of the MID values from each of the
instruments both serves as a comparison to those with prior MID values estimated in the
literature (HUI2 and HUI3, and the HAQ),\textsuperscript{4,20,21} and provides new information for those
instruments without prior MID estimates (SF-6D, EQ-5D, and the RAQoL). The low
number of missing values (all < 4%) for each of the instruments attests that they are suitable
for self-administration. Overall, all the instruments tend to discriminate across disease
severity based on multiple criteria with worse scores being associated with measures
indicating a higher severity of RA.

The results of the differences in the overall scores across known, naturally occurring
groups (Table 4.6) general supports construct validity for each of the instruments. However,
there were some important differences among the generic, MAUT instruments. For example,
only the EQ-5D and the SF-6D overall scores were signiﬁcantly different between those
experiencing adverse events to RA drug therapy over the previous three months compared to those who did not. Similarly, of the MAUT instruments, only the HUI2 and HUI3 overall scores were significantly different between groups with and without RA hospitalization and other chronic diseases. It is not clear why these differences among the instrument scores exist but may be due to the different aspects of health that are represented by each of the systems. Of note, differences across groups based on RA severity were consistently significant for the RAQoL and the HAQ severity index (with the exception of adverse events for the latter).

Contrary to the hypothesis that strong correlations would exist between overall instrument scores or selected single attribute utility scores and RA duration, this finding was not observed. Similarly, few strong correlations were observed for SJC or TJC and the scale results. All other hypothesized relationships were found to be significant. The length of RA is a somewhat imprecise measure of severity as some patients have severe, aggressive disease from the onset and others have a slow, insidious disease process.

As shown in Table 4.4, more individuals reported having no problems under the EQ-5D and the HUI2 classification systems than reporting the lowest level of deficit on the other systems. This lack of variation in the responses and low number of possible descriptive health states may impede the sensitivity of the EQ-5D and HUI2 in longitudinal studies when compared to the other instruments. Conversely, both the SF-6D and the HUI-3 tended to have responses across the full range of severity for most of the domains assessed. Therefore, the SF-6D and the HUI3 may have a higher degree of sensitivity to the disease burden of RA.

The estimates of MID that we obtained using the effect-size and regression methodologies closely agreed for the HUI2 and HUI3 (within 0.01) and were exactly the same for the SF-6D and the EQ-5D. However, although the results of the HUI2 MID
estimates agree with what has been postulated in the literature (Drummond stated that a difference of 0.03 in utility values should be used for the basis of sample size calculations for the HUI2 and HUI3), the estimates we obtained for the HUI3 MID were higher at 0.06 and 0.07. Moreover, the MID for the SF-6D estimated by Brazier et al. was identical to those that we obtained using different methodologies. However, using any of the criteria, it would appear that, for the global utility values, the differences between groups in Table 4.6 exceed the MID.

The results of this study lend support to the construct validity of all of the generic MAUT instruments and the disease-specific instruments in RA and provide some detail regarding their limitations and strengths. As mentioned, Brazier et al.'s concern is that a score of a preference-based measure may fail to detect a hypothesized difference simply because the difference is not valued by patients or the result of an insensitive scoring system. We believe that we have chosen health states in RA that would clearly result in hypothesized differences in preferences. Besides, we did see differences among these different health states in the anticipated directions, which would appear to address Brazier's concern that these instruments might not identify changes appropriately.

Another limitation was that most of the data obtained was self-reported without subsequent verification from clinical records. Thus, it is possible that study participants did not accurately or objectively describe the severity of their RA. However, we believe that the risk of this is low as previous research has shown that patient self-reporting of symptoms is valid and reliable in RA.

In conclusion, the overall and certain single-attribute scores of both the generic MAUT instruments and the disease-specific instruments are all able to distinguish between
groups that were defined by measures of RA severity. As expected, the disease-specific instruments appeared to be slightly superior to the generic measures; however, the MAUT instruments appeared to have construct validity for RA. Effect-size and regression-based estimates of the MID for the instruments agreed providing comparison with the MID values previously postulated in the literature or new information for those measures without prior estimates. Most of the differences across RA severity in our cohort exceeded the MID for overall scores using any criteria.
4.6 REFERENCES


23. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health Qual Life Outcomes. 2003;11:4-12.


illustration to the Health Utilities Index Mark II. Pharmacoeconomics 1999;15:141-155.

## TABLE 4.1: OVERVIEW OF MAUT INSTRUMENT PROPERTIES

<table>
<thead>
<tr>
<th>Dimensions/Domains/Attributes</th>
<th># of Possible Health States</th>
<th>Valuation Technique</th>
<th>Boundaries</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HUI2</strong> Sensation (vision, hearing, speech), Mobility, Emotion Cognition, Self-care, Pain</td>
<td>24,000</td>
<td>Standard Gamble</td>
<td>-0.03 – 1.00</td>
</tr>
<tr>
<td><strong>HUI3</strong> Vision, Hearing, Speech, Ambulation, Dexterity, Emotion, Cognition, Pain</td>
<td>972,000</td>
<td>Standard Gamble</td>
<td>-0.36 – 1.00</td>
</tr>
<tr>
<td><strong>SF-6D</strong> Physical Function, Role Limitation, Social Function, Pain, Mental Health, Vitality</td>
<td>18,000</td>
<td>Standard Gamble</td>
<td>0.30 – 1.00</td>
</tr>
<tr>
<td><strong>EQ-5D</strong> Mobility, Usual Activities, Self-Care, Pain, Anxiety</td>
<td>243</td>
<td>Time Trade Off</td>
<td>-0.59 – 1.00</td>
</tr>
</tbody>
</table>
## TABLE 4.2: CHARACTERISTICS OF THE STUDY PARTICIPANTS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean (SD)*</th>
<th>Median (IQR)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (yrs)</td>
<td>61.5 (25.9)</td>
<td>63.0 (19.0)</td>
</tr>
<tr>
<td>ESR</td>
<td>24.52 (21.02)</td>
<td>18.0 (24.0)</td>
</tr>
<tr>
<td>RAQoL score (range 0 - 30)</td>
<td>12.82 (8.28)</td>
<td>12.5 (13.0)</td>
</tr>
<tr>
<td>RA Duration (yrs)</td>
<td>13.87 (11.41)</td>
<td>12.00 (15.67)</td>
</tr>
<tr>
<td>HAQ Disability Index (range 0 - 3.00)</td>
<td>1.10 (0.77)</td>
<td>1.125 (1.375)</td>
</tr>
<tr>
<td>Patient Global Assessment (100 mm VAS)</td>
<td>59.82 (25.86)</td>
<td>65.00 (37.00)</td>
</tr>
<tr>
<td>MD Global Assessment (100mm VAS)</td>
<td>20.88 (23.39)</td>
<td>10.5 (29.0)</td>
</tr>
<tr>
<td>Pain VAS (100mm VAS)</td>
<td>43.12 (27.02)</td>
<td>42.00 (44.00)</td>
</tr>
<tr>
<td>EQ-5D Health Thermometer</td>
<td>65.02 (19.27)</td>
<td>70 (30)</td>
</tr>
<tr>
<td>Tender Joint Count (range 0 – 28)</td>
<td>15.09 (11.99)</td>
<td>12.00 (15.00)</td>
</tr>
<tr>
<td>Swollen Joint Count (range 0 –28)</td>
<td>9.14 (9.66)</td>
<td>6.00 (11.00)</td>
</tr>
<tr>
<td>Self-Reported RA Severity, n %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Mild</td>
<td>9</td>
<td>3%</td>
</tr>
<tr>
<td>Mild</td>
<td>34</td>
<td>11%</td>
</tr>
<tr>
<td>Moderate</td>
<td>120</td>
<td>38%</td>
</tr>
<tr>
<td>Severe</td>
<td>110</td>
<td>35%</td>
</tr>
<tr>
<td>Very Severe</td>
<td>27</td>
<td>9%</td>
</tr>
<tr>
<td>Missing</td>
<td>13</td>
<td>4%</td>
</tr>
<tr>
<td>Self-Reported RA Control, n %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Well Controlled</td>
<td>33</td>
<td>11%</td>
</tr>
<tr>
<td>Well Controlled</td>
<td>76</td>
<td>24%</td>
</tr>
<tr>
<td>Adequately Controlled</td>
<td>123</td>
<td>39%</td>
</tr>
<tr>
<td>Not Well Controlled</td>
<td>61</td>
<td>19%</td>
</tr>
<tr>
<td>Not Controlled At All</td>
<td>7</td>
<td>2%</td>
</tr>
<tr>
<td>Missing</td>
<td>13</td>
<td>4%</td>
</tr>
<tr>
<td>Adverse Drug Reaction to RA Medication in Last 3 Months, n %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>108</td>
<td>35%</td>
</tr>
<tr>
<td>No</td>
<td>202</td>
<td>65%</td>
</tr>
<tr>
<td>Hospitalized For RA in Last 12 Months, n%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>45</td>
<td>15%</td>
</tr>
<tr>
<td>No</td>
<td>253</td>
<td>85%</td>
</tr>
<tr>
<td>Missed Work or School Due to RA in Last 12 Months, n%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>59</td>
<td>37%</td>
</tr>
<tr>
<td>No</td>
<td>100</td>
<td>63%</td>
</tr>
<tr>
<td>Purchased or Rented Equipment for RA in Last 12 Months, n%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>72</td>
<td>28%</td>
</tr>
<tr>
<td>No</td>
<td>183</td>
<td>72%</td>
</tr>
<tr>
<td>Used Allied Health Professional/Home Care Services for RA in Last 12 months, n %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>129</td>
<td>42%</td>
</tr>
<tr>
<td>No</td>
<td>174</td>
<td>58%</td>
</tr>
<tr>
<td>Concomitant Chronic Illness Other Than RA, n%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>192</td>
<td>62%</td>
</tr>
<tr>
<td>No</td>
<td>118</td>
<td>38%</td>
</tr>
</tbody>
</table>

* SD = Standard deviation; IQR = interquartile range
TABLE 4.3: MULTI-ATTRIBUTE AND SINGLE ATTRIBUTE UTILITY SCORES
FROM THE MAUT INSTRUMENTS

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>STD*</th>
<th>Median</th>
<th>IQR*</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUI2 Global Utility</td>
<td>304</td>
<td>0.71</td>
<td>0.20</td>
<td>0.75</td>
<td>0.28</td>
<td>0.11</td>
<td>1.00</td>
</tr>
<tr>
<td>Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUI2 Single Attribute</td>
<td>304</td>
<td>0.95</td>
<td>0.04</td>
<td>0.95</td>
<td>-</td>
<td>0.61</td>
<td>1.00</td>
</tr>
<tr>
<td>Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensation</td>
<td>304</td>
<td>0.96</td>
<td>0.06</td>
<td>0.97</td>
<td>0.03</td>
<td>0.73</td>
<td>1.00</td>
</tr>
<tr>
<td>Mobility</td>
<td>304</td>
<td>0.96</td>
<td>0.05</td>
<td>0.93</td>
<td>0.07</td>
<td>0.70</td>
<td>1.00</td>
</tr>
<tr>
<td>Emotion</td>
<td>304</td>
<td>0.98</td>
<td>0.03</td>
<td>1.00</td>
<td>0.05</td>
<td>0.65</td>
<td>1.00</td>
</tr>
<tr>
<td>Cognition</td>
<td>304</td>
<td>0.99</td>
<td>0.03</td>
<td>1.00</td>
<td>0.03</td>
<td>0.80</td>
<td>1.00</td>
</tr>
<tr>
<td>Self care</td>
<td>304</td>
<td>0.86</td>
<td>0.15</td>
<td>0.85</td>
<td>0.12</td>
<td>0.38</td>
<td>1.00</td>
</tr>
<tr>
<td>Pain</td>
<td>304</td>
<td>0.53</td>
<td>0.29</td>
<td>0.56</td>
<td>0.44</td>
<td>-0.16</td>
<td>1.00</td>
</tr>
<tr>
<td>HUI3 Global Utility</td>
<td>303</td>
<td>0.53</td>
<td>0.29</td>
<td>0.56</td>
<td>0.44</td>
<td>-0.16</td>
<td>1.00</td>
</tr>
<tr>
<td>Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUI3 Single Attribute</td>
<td>303</td>
<td>0.98</td>
<td>0.04</td>
<td>0.98</td>
<td>-</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td>Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vision</td>
<td>303</td>
<td>0.98</td>
<td>0.05</td>
<td>1.00</td>
<td>-</td>
<td>0.61</td>
<td>1.00</td>
</tr>
<tr>
<td>Hearing</td>
<td>303</td>
<td>0.99</td>
<td>0.02</td>
<td>1.00</td>
<td>-</td>
<td>0.81</td>
<td>1.00</td>
</tr>
<tr>
<td>Speech</td>
<td>303</td>
<td>0.94</td>
<td>0.08</td>
<td>0.93</td>
<td>0.07</td>
<td>0.58</td>
<td>1.00</td>
</tr>
<tr>
<td>Ambulation</td>
<td>303</td>
<td>0.89</td>
<td>0.11</td>
<td>0.95</td>
<td>0.19</td>
<td>0.56</td>
<td>1.00</td>
</tr>
<tr>
<td>Dexterity</td>
<td>303</td>
<td>0.95</td>
<td>0.08</td>
<td>1.00</td>
<td>0.05</td>
<td>0.46</td>
<td>1.00</td>
</tr>
<tr>
<td>Emotion</td>
<td>303</td>
<td>0.95</td>
<td>0.08</td>
<td>1.00</td>
<td>0.05</td>
<td>0.42</td>
<td>1.00</td>
</tr>
<tr>
<td>Cognition</td>
<td>303</td>
<td>0.88</td>
<td>0.11</td>
<td>0.90</td>
<td>0.19</td>
<td>0.55</td>
<td>1.00</td>
</tr>
<tr>
<td>Pain</td>
<td>303</td>
<td>0.63</td>
<td>0.13</td>
<td>0.60</td>
<td>0.12</td>
<td>0.31</td>
<td>1.00</td>
</tr>
<tr>
<td>SF-6D Global Utility</td>
<td>302</td>
<td>0.63</td>
<td>0.13</td>
<td>0.60</td>
<td>0.12</td>
<td>0.31</td>
<td>1.00</td>
</tr>
<tr>
<td>Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF-6D Single Attribute</td>
<td>302</td>
<td>0.91</td>
<td>0.04</td>
<td>0.90</td>
<td>0.06</td>
<td>0.83</td>
<td>1.00</td>
</tr>
<tr>
<td>Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical Functioning</td>
<td>302</td>
<td>0.93</td>
<td>0.03</td>
<td>0.93</td>
<td>-</td>
<td>0.90</td>
<td>1.00</td>
</tr>
<tr>
<td>Role Limitations</td>
<td>302</td>
<td>0.92</td>
<td>0.03</td>
<td>0.92</td>
<td>-</td>
<td>0.88</td>
<td>1.00</td>
</tr>
<tr>
<td>Social Functioning</td>
<td>302</td>
<td>0.92</td>
<td>0.05</td>
<td>0.92</td>
<td>0.05</td>
<td>0.79</td>
<td>1.00</td>
</tr>
<tr>
<td>Pain</td>
<td>302</td>
<td>0.95</td>
<td>0.03</td>
<td>0.94</td>
<td>-</td>
<td>0.83</td>
<td>1.00</td>
</tr>
<tr>
<td>Mental Health</td>
<td>302</td>
<td>0.98</td>
<td>0.02</td>
<td>0.98</td>
<td>-</td>
<td>0.93</td>
<td>1.00</td>
</tr>
<tr>
<td>Vitality</td>
<td>302</td>
<td>0.66</td>
<td>0.24</td>
<td>0.74</td>
<td>0.19</td>
<td>-0.21</td>
<td>1.00</td>
</tr>
<tr>
<td>EQ-5D Global Utility</td>
<td>308</td>
<td>0.66</td>
<td>0.24</td>
<td>0.74</td>
<td>0.19</td>
<td>-0.21</td>
<td>1.00</td>
</tr>
<tr>
<td>Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EQ-5D Single Attribute</td>
<td>308</td>
<td>0.91</td>
<td>0.08</td>
<td>0.85</td>
<td>0.15</td>
<td>0.34</td>
<td>1.00</td>
</tr>
<tr>
<td>Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobility</td>
<td>308</td>
<td>0.94</td>
<td>0.09</td>
<td>1.00</td>
<td>0.18</td>
<td>0.44</td>
<td>1.00</td>
</tr>
<tr>
<td>Self-care</td>
<td>308</td>
<td>0.90</td>
<td>0.11</td>
<td>0.88</td>
<td>0.12</td>
<td>0.56</td>
<td>1.00</td>
</tr>
<tr>
<td>Usual Activities</td>
<td>308</td>
<td>0.70</td>
<td>0.16</td>
<td>0.80</td>
<td>0.27</td>
<td>0.26</td>
<td>1.00</td>
</tr>
<tr>
<td>Pain</td>
<td>308</td>
<td>0.94</td>
<td>0.09</td>
<td>1.00</td>
<td>0.15</td>
<td>0.41</td>
<td>1.00</td>
</tr>
<tr>
<td>Anxiety/Depression</td>
<td>308</td>
<td>0.94</td>
<td>0.09</td>
<td>1.00</td>
<td>0.15</td>
<td>0.41</td>
<td>1.00</td>
</tr>
</tbody>
</table>

* SD = Standard deviation; IQR = interquartile range
- = < 0.0001
TABLE 4.4: DOMAIN RESPONSES FOR THE MAUT INSTRUMENTS

<table>
<thead>
<tr>
<th>HUI2 Levels*</th>
<th>Cognition (n=308)</th>
<th>Self Care (n=308)</th>
<th>Emotion (n=308)</th>
<th>Pain (n=307)</th>
<th>Sensation (n=306)</th>
<th>Mobility (n=307)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>%</td>
<td>Count</td>
<td>%</td>
<td>Count</td>
<td>%</td>
<td>Count</td>
</tr>
<tr>
<td>1</td>
<td>185</td>
<td>60</td>
<td>212</td>
<td>67</td>
<td>152</td>
<td>49</td>
</tr>
<tr>
<td>2</td>
<td>115</td>
<td>37</td>
<td>81</td>
<td>27</td>
<td>137</td>
<td>45</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>2</td>
<td>11</td>
<td>5</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HUI3 Levels*</th>
<th>Pain (n=306)</th>
<th>Emotion (n=306)</th>
<th>Vision (n=309)</th>
<th>Hearing (n=308)</th>
<th>Speech (n=310)</th>
<th>Cognition (n=308)</th>
<th>Ambulation (n=306)</th>
<th>Dexterity (n=307)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>%</td>
<td>Count</td>
<td>%</td>
<td>Count</td>
<td>%</td>
<td>Count</td>
<td>%</td>
<td>Count</td>
</tr>
<tr>
<td>1</td>
<td>22</td>
<td>7</td>
<td>157</td>
<td>50</td>
<td>45</td>
<td>15</td>
<td>284</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>93</td>
<td>31</td>
<td>105</td>
<td>34</td>
<td>249</td>
<td>80</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>111</td>
<td>36</td>
<td>33</td>
<td>11</td>
<td>4</td>
<td>1</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>64</td>
<td>21</td>
<td>10</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

130
Physical Functioning (n=307) | Role Limitations (n=302) | Social Functioning (n=302) | Pain (n=303) | Mental Health (n=303) | Vitality (n=308)
---|---|---|---|---|---
SF-6D Levels* | Count | % | Count | % | Count | % | Count | % | Count | % | Count | %
1 | 7 | 3 | 50 | 17 | 44 | 15 | 12 | 4 | 58 | 19 | 2 | 1
2 | 54 | 18 | 192 | 63 | 75 | 24 | 55 | 18 | 123 | 41 | 71 | 23
3 | 75 | 24 | 11 | 4 | 129 | 43 | 72 | 24 | 104 | 34 | 108 | 35
4 | 50 | 16 | 49 | 16 | 43 | 14 | 81 | 27 | 15 | 5 | 84 | 27
5 | 95 | 31 | - | - | 11 | 4 | 58 | 19 | 3 | 1 | 38 | 12
6 | 26 | 8 | - | - | - | - | 25 | 8 | - | - | 5 | 2

For the HUI2, HUI3, SF-6D and EQ-5D, higher levels represent higher degrees of limitation or increased symptoms. For the SF-6D, Physical Functioning and Pain have six levels, Social Functioning, Mental Health and Vitality have five levels and Role Limitations has four levels. For the HUI2, all the domains have four levels except for Mobility and Pain which have five. For the HUI3, all the domains have six levels with the exception of Pain, Vision and Speech which have five.
TABLE 4.5: RELATIONSHIP BETWEEN RA SEVERITY AND CONTROL AND THE GLOBAL UTILITY SCORES FOR EACH OF THE MAUT INSTRUMENTS

<table>
<thead>
<tr>
<th>Self-reported RA severity</th>
<th>Mean Score (SD)</th>
<th>Global Utility Score, Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HAQ ¹</td>
<td>RAQoL ²</td>
</tr>
<tr>
<td>Very mild</td>
<td>0.28 (0.43)</td>
<td>3.22 (3.27)</td>
</tr>
<tr>
<td>Mild</td>
<td>0.43 (0.51)</td>
<td>5.15 (4.40)</td>
</tr>
<tr>
<td>Moderate</td>
<td>1.01 (0.64)</td>
<td>11.09 (6.65)</td>
</tr>
<tr>
<td>Severe</td>
<td>1.37 (0.72)</td>
<td>16.76 (7.86)</td>
</tr>
<tr>
<td>Very Severe</td>
<td>1.68 (0.83)</td>
<td>18.96 (8.54)</td>
</tr>
<tr>
<td><strong>Spearman’s Correlation Coefficient</strong></td>
<td><strong>0.46</strong>*</td>
<td><strong>0.52</strong>*</td>
</tr>
</tbody>
</table>

| Self-reported RA control        |                 |                                |        |        |        |        |
|                                 |                 |                                |        |        |        |        |
| Very well controlled            | 0.65 (0.78)     | 6.15 (7.02)                    | 0.82 (0.20) | 0.74 (0.26) | 0.73 (0.14) | 0.81(0.18) |
| Well controlled                 | 0.79 (0.68)     | 8.0 (6.11)                     | 0.79 (0.13) | 0.66 (0.24) | 0.69 (0.11) | 0.77 (0.13) |
| Adequately controlled           | 1.09 (0.70)     | 13.10 (6.97)                   | 0.73 (0.17) | 0.54 (0.25) | 0.62 (0.11) | 0.67 (0.11) |
| Not well controlled             | 1.63 (0.59)     | 20.37 (6.50)                   | 0.55 (0.19) | 0.31 (0.24) | 0.52 (0.09) | 0.48 (0.28) |
| Not controlled at all           | 2.04 (0.38)     | 24.0 (3.51)                    | 0.42 (0.08) | 0.00 (0.11) | 0.46 (0.08) | 0.25 (0.23) |
| **Spearman’s Correlation Coefficient** | **0.45***      | **0.59***                     | **-0.49*** | **-0.52*** | **-0.58*** | **-0.50*** |

¹Comparison (using ANOVA) of mean values stratified by severity/control – all p<0.0001

* p<0.0001
TABLE 4.6: DICHOTOMOUS MEASURES OF RA SEVERITY

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean Score (SD)</th>
<th>Global Utility Score, Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HAQ</td>
<td>RAQoL</td>
</tr>
<tr>
<td>Adverse Events to RA Drug Therapy in Last 3 months</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1.20 (0.72)</td>
<td>14.7 (7.9)</td>
</tr>
<tr>
<td>No</td>
<td>1.06 (0.77)</td>
<td>11.7 (8.2)*</td>
</tr>
<tr>
<td>Effect Size</td>
<td>0.19</td>
<td>0.37</td>
</tr>
<tr>
<td>Hospitalized in Past Year for RA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1.38 (0.73)</td>
<td>14.9 (7.9)</td>
</tr>
<tr>
<td>No</td>
<td>1.05 (0.77)</td>
<td>12.3 (8.2)*</td>
</tr>
<tr>
<td>Effect Size</td>
<td>0.44</td>
<td>0.32</td>
</tr>
<tr>
<td>Other Chronic Diseases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 or more</td>
<td>1.19 (0.76)</td>
<td>13.6 (8.3)</td>
</tr>
<tr>
<td>None</td>
<td>0.97 (0.75)*</td>
<td>11.6 (8.1)*</td>
</tr>
<tr>
<td>Effect Size</td>
<td>0.29</td>
<td>0.24</td>
</tr>
<tr>
<td>Days Off Work/School Due to RA in Past Year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1.27 (0.71)</td>
<td>16.4 (7.7)</td>
</tr>
<tr>
<td>No</td>
<td>0.83 (0.75) ‡</td>
<td>10.1 (7.8) †</td>
</tr>
<tr>
<td>Effect Size</td>
<td>0.60</td>
<td>0.81</td>
</tr>
<tr>
<td>Use of Allied Health/Home Services** for RA in Past Year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1.34 (0.72)</td>
<td>15.1 (7.9)</td>
</tr>
<tr>
<td>No</td>
<td>0.80 (0.74) †</td>
<td>11.0 (8.1) †</td>
</tr>
<tr>
<td>Effect Size</td>
<td>0.74</td>
<td>0.51</td>
</tr>
<tr>
<td>Rent or Purchase Equipment for RA in Past Year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1.34 (0.69)</td>
<td>16.6 (8.1)</td>
</tr>
<tr>
<td>No</td>
<td>0.90 (0.76) †</td>
<td>11.8 (8.0) †</td>
</tr>
<tr>
<td>Effect Size</td>
<td>0.61</td>
<td>0.60</td>
</tr>
</tbody>
</table>

**Physiotherapy, Occupational Therapy, Massage Therapy, Home Care,
†p-value < 0.00001 from a t-test between the two groups , ‡p-value < 0.001,  p-value < 0.01, * p-value < 0.05
Large effect sizes (>0.5) are highlighted in bold text
TABLE 4.7: CORRELATIONS FOR MULTIATTRIBUTE AND SELECT SINGLE ATTRIBUTE UTILITY SCORES WITH RA SEVERITY

<table>
<thead>
<tr>
<th>MEASURE</th>
<th>RA Duration (years)</th>
<th>RAQoL Score (0–30)</th>
<th>Swollen Joint Count (0–28)</th>
<th>Tender Joint Count (0–28)</th>
<th>Patient Global VAS Score</th>
<th>HAQ Disability Score</th>
<th>Pain VAS (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUI2 Global Utility</td>
<td>-0.11</td>
<td>-0.70†</td>
<td>-0.38†</td>
<td>-0.44†</td>
<td>0.55†</td>
<td>-0.66†</td>
<td>-0.59†</td>
</tr>
<tr>
<td>Mobility (HUI2)</td>
<td>-0.25†</td>
<td>-0.52†</td>
<td>-0.37†</td>
<td>-0.36†</td>
<td>0.49†</td>
<td>-0.64†</td>
<td>-0.44†</td>
</tr>
<tr>
<td>Self-Care (HUI2)</td>
<td>-0.10</td>
<td>-0.55†</td>
<td>-0.36†</td>
<td>-0.34†</td>
<td>0.42†</td>
<td>-0.60†</td>
<td>-0.40†</td>
</tr>
<tr>
<td>Pain (HUI2)</td>
<td>-0.04</td>
<td>-0.62†</td>
<td>-0.45†</td>
<td>-0.42†</td>
<td>0.51†</td>
<td>-0.54†</td>
<td>-0.59†</td>
</tr>
<tr>
<td>HUI3 Global Utility</td>
<td>-0.20‡</td>
<td>-0.75†</td>
<td>-0.42†</td>
<td>-0.48†</td>
<td>0.58†</td>
<td>-0.76†</td>
<td>-0.60†</td>
</tr>
<tr>
<td>Ambulation (HUI3)</td>
<td>-0.24†</td>
<td>-0.53†</td>
<td>-0.37†</td>
<td>-0.36†</td>
<td>0.49†</td>
<td>-0.65†</td>
<td>-0.44†</td>
</tr>
<tr>
<td>Dexterity (HUI3)</td>
<td>-0.24†</td>
<td>-0.62†</td>
<td>-0.47†</td>
<td>-0.43†</td>
<td>0.44†</td>
<td>-0.68†</td>
<td>-0.39†</td>
</tr>
<tr>
<td>Pain (HUI3)</td>
<td>-0.09</td>
<td>-0.70†</td>
<td>-0.51†</td>
<td>-0.45†</td>
<td>0.57†</td>
<td>-0.61†</td>
<td>-0.70†</td>
</tr>
<tr>
<td>SF-6D Global Utility</td>
<td>-0.17‡</td>
<td>-0.80†</td>
<td>-0.47†</td>
<td>-0.53†</td>
<td>0.63†</td>
<td>-0.73†</td>
<td>-0.62†</td>
</tr>
<tr>
<td>Physical Functioning</td>
<td>-0.21‡</td>
<td>-0.64†</td>
<td>-0.42†</td>
<td>-0.41†</td>
<td>0.50†</td>
<td>-0.69†</td>
<td>-0.50†</td>
</tr>
<tr>
<td>Role Limitations</td>
<td>-0.14*</td>
<td>-0.57†</td>
<td>-0.36†</td>
<td>-0.36†</td>
<td>0.42†</td>
<td>-0.50†</td>
<td>-0.42†</td>
</tr>
<tr>
<td>Social Functioning</td>
<td>-0.16‡</td>
<td>-0.70†</td>
<td>-0.42†</td>
<td>-0.44†</td>
<td>0.56†</td>
<td>-0.63†</td>
<td>-0.52†</td>
</tr>
<tr>
<td>Pain</td>
<td>-0.18‡</td>
<td>-0.74†</td>
<td>-0.50†</td>
<td>-0.48†</td>
<td>0.62†</td>
<td>-0.67†</td>
<td>-0.63†</td>
</tr>
<tr>
<td>EQ-5D Global Utility</td>
<td>-0.11</td>
<td>-0.70†</td>
<td>-0.47†</td>
<td>-0.42†</td>
<td>0.58†</td>
<td>-0.61†</td>
<td>-0.60†</td>
</tr>
<tr>
<td>Mobility (EQ-5D)</td>
<td>-0.17‡</td>
<td>-0.53†</td>
<td>-0.41†</td>
<td>-0.37†</td>
<td>0.46†</td>
<td>-0.53†</td>
<td>-0.44†</td>
</tr>
<tr>
<td>Usual Activities (EQ-5D)</td>
<td>-0.10</td>
<td>-0.64†</td>
<td>-0.44†</td>
<td>-0.45†</td>
<td>0.51†</td>
<td>-0.57†</td>
<td>-0.51†</td>
</tr>
<tr>
<td>Pain/Discomfort (EQ-5D)</td>
<td>0.00</td>
<td>-0.34†</td>
<td>-0.26†</td>
<td>-0.32†</td>
<td>0.30†</td>
<td>-0.31†</td>
<td>-0.32†</td>
</tr>
<tr>
<td>Self-Care (EQ-5D)</td>
<td>-0.14*</td>
<td>-0.59†</td>
<td>-0.34†</td>
<td>-0.37†</td>
<td>0.43†</td>
<td>-0.60†</td>
<td>-0.43†</td>
</tr>
<tr>
<td>RAQoL Score</td>
<td>0.20‡</td>
<td>-0.53†</td>
<td>0.54†</td>
<td>-0.62†</td>
<td>0.76†</td>
<td>0.62†</td>
<td></td>
</tr>
<tr>
<td>HAQ Disability Score</td>
<td>0.28‡</td>
<td>0.76†</td>
<td>0.48†</td>
<td>0.46†</td>
<td>-0.53†</td>
<td>0.54†</td>
<td></td>
</tr>
</tbody>
</table>

* p < 0.05; ‡ p < 0.01; † p < 0.001; Variables for which correlations were hypothesized to be strong are in bolded type in the measures column; Correlations considered to be strong are in bold type in the results column.
<table>
<thead>
<tr>
<th>Variable</th>
<th>HAQ Disability Index (0 – 3.0)</th>
<th>Minimally Important Difference (MID) Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUI2 Global Utility</td>
<td></td>
<td>For each 1.0 change in HAQ, the HUI2 changes 0.16. Therefore, a 0.25 change in HAQ (MID) results in a 0.04 change in the HUI2 (estimated MID)</td>
</tr>
<tr>
<td>Beta-coefficient (SE)*</td>
<td>-0.16 (0.01)</td>
<td>Model p-value: 0.0001</td>
</tr>
<tr>
<td>Model p-value</td>
<td></td>
<td>Model p-value: 0.43</td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td>R²: 0.43</td>
</tr>
</tbody>
</table>

* SE = Standard Error*

| HUI3 Global Utility      |                                | For each 1.0 change in HAQ, the HUI3 changes 0.29. Therefore, a 0.25 change in HAQ (MID) results in a 0.07 change in the HUI3 (estimated MID) |
| Beta-coefficient (SE)*    | -0.29 (0.01)                   | Model p-value: 0.0001                                                                                              |
| Model p-value             |                                | Model p-value: 0.58                                                                                               |
| R²                        |                                | R²: 0.58                                                                                                          |

| SF-6D Global Utility     |                                | For each 1.0 change in HAQ, the SF-6D changes 0.12. Therefore, a 0.25 change in HAQ (MID) results in a 0.03 change in the SF-6D (estimated MID) |
| Beta-coefficient (SE)*    | -0.12 (0.007)                  | Model p-value: 0.0001                                                                                              |
| Model p-value             |                                | Model p-value: 0.53                                                                                               |
| R²                        |                                | R²: 0.53                                                                                                          |

| EQ-5D Global Utility     |                                | For each 1.0 change in HAQ, the EQ-5D changes 0.19. Therefore, a 0.25 change in HAQ (MID) results in a 0.05 change in the EQ-5D (estimated MID) |
| Beta-coefficient (SE)*    | -0.19 (0.01)                   | Model p-value: 0.0001                                                                                              |
| Model p-value             |                                | Model p-value: 0.37                                                                                               |
| R²                        |                                | R²: 0.37                                                                                                          |

| RAQoL Score              |                                | For each 1.0 change in HAQ, the RAQoL changes 8.13. Therefore, a 0.25 change in HAQ (MID) results in a 2.0 change in the RAQoL (estimated MID) |
| Beta-coefficient (SE)*    | 8.13 (0.003)                   | Model p-value: 0.0001                                                                                              |
| Model p-value             |                                | Model p-value: 0.57                                                                                               |
| R²                        |                                | R²: 0.57                                                                                                          |
FIGURE 4.1: BOX PLOT OF MAUT INSTRUMENT GLOBAL UTILITY SCORES

MAUT Instrument

- Outliers are marked by an open circle.
CHAPTER 5

NOT ALL QALYS ARE EQUAL: THE IMPACT OF USING DIFFERENT INDIRECT UTILITY MEASURES ON ESTIMATING THE COST-UTILITY OF INFliximab IN RHEUMATOID ARTHRITIS

5.1 FOREWORD

This manuscript is currently under review under the same title for publication in the Medical Decision Making. The candidate is first author of this manuscript which is co-authored by Drs. Stephen Marion, Fred Wolfe, John Esdaile, Monique Gignac, Ann Clarke and Aslam Anis. In addition, a statistician, Ms. Daphne Guh, was a co-author. Dr. Stephen Marion completed all of the complicated mathematical modeling such as the construction of the transition probability matrix and instructed the candidate in these methods. Drs. Wolfe, Gignac and Clarke provided access to databases that facilitated estimation of the costs and effectiveness outcomes. Ms. Guh provided assistance with the statistical analysis. Drs. Anis and Esdaile are co-supervisors of the candidate.

The candidate's role in this manuscript involved the development of the primary hypothesis, study design, model design (with Dr. Marion), statistical analyses and the writing of the final manuscript.

5.2 INTRODUCTION

With the introduction of new, expensive biological agents for the treatment of rheumatoid arthritis (RA), the costs to manage this debilitating chronic disease have increased. This observation has drawn a great deal of attention as governments and third-
party payers struggle to make decisions about how to incorporate these therapeutic agents into their funding envelopes. Because of the limited funds available for health care, RA drugs often compete for funding with those used to treat other disease areas. However, unlike HIV and cancer, arthritis cripples and does not immediately result in death. Drugs used for RA tend to improve quality of life and do not immediately impact on mortality. Because of this fact, treatments for RA can be undervalued when compared to treatments for other chronic diseases. Economic evaluations have become increasingly important in the allocation of funding to treatments and programmes. Thus, if RA patients are not to be short-changed by policies and decisions, then measures that incorporate quality of life changes into economic studies are key.

In order to integrate quality of life into economic analyses, the effectiveness of the health interventions in question is measured using a metric known as "utility," which ranges in value from 0 (dead) to 1 (perfect health). Results are reported as cost per quality-adjusted life years (QALY) gained, which are derived by incorporating the utilities as weights in the life expectancy calculation. Cost per QALY gained is a unique and preferred measure of the economic value of different interventions, because it permits comparison across disease groups, thereby facilitating funding allocation decisions. The ultimate goal is a comparison of the cost per quality adjusted life year (QALY) gained so as to prioritize funding according to cost/QALY. This determination of a cost/QALY permits comparison across therapies within disease states (i.e. Drug A vs. Drug B, or Drug vs. physiotherapy vs. surgery) as well as permitting comparisons across diseases (RA vs. lupus or congestive heart disease).

To determine the utility weightings for QALYs, the use of a pre-scaled multi-attribute utility index is often the most convenient and least expensive means of achieving this approach. While no validated multi-attribute utility index is available for economic
evaluations specifically in musculoskeletal disease, several generic measures of HRQOL (health related quality of life) appear suitable for adaptation to economic evaluations in RA.\textsuperscript{3,4} Such generic utility-based instruments for use in economic evaluations are the Health Utilities Index Mark 2 and 3 (HUI2 and HUI3), the EuroQol 5D (EQ-5D) and the Short Form 6-D (SF-6D).\textsuperscript{5} These measures each capture different attributes and domains and thus assess different aspects of quality of life and utilize different methodologies to calculate the utility score. However, all of these instruments purport to integrate the health states obtained from the population under study with predetermined societal preference ratings for said states to produce an overall index score. There is no “gold standard” among these instruments and each likely has its own advantages and disadvantages. However, little is known how the choice of these instruments influences the outcome in economic evaluations of RA drug therapy.

Thus, the primary objective of this study was to determine the impact on incremental cost effectiveness ratios (ICERs) from using different indirect utility instruments in an economic evaluation of a new biological agent (infliximab) used to treat RA. A secondary objective of the study was to determine the incremental costs, QALYs, and cost-utility of infliximab over standard therapy for the treatment of active, refractory RA.

5.3 METHODS

5.3.1 Clinical Trial Data Source

The largest randomized controlled trial to evaluate the efficacy of infliximab in patients with RA refractory to other disease modifying antirheumatic drugs (DMARDs) was
the ATTRACT trial. Patients were eligible for this trial if they had active RA (6 or more tender and swollen joints and symptoms or signs (at least two of morning stiffness for at least 45 minutes, an erythrocyte sedimentation rate of at least 28 mm/hr, and serum C-reactive protein of at least 2.0 mg/dL.) despite methotrexate (MTX) doses of more than 12.5 mg per week. 428 patients from 34 centers in the United States, Canada, and Europe were randomly allocated to one of five treatment groups: MTX alone; MTX plus 3 mg/kg of infliximab every 4 or 8 weeks; or MTX plus 10 mg/kg of infliximab every 4 or 8 weeks. At baseline, most patients were also receiving non-steroidal anti-inflammatory drugs (74%) and corticosteroids (61%). DMARDs other than MTX were not allowed and, if necessary, withdrawn before beginning study treatment. The 5 treatment groups were comparable with respect to age (median, 54 years), gender distribution (78% were women), disease duration (8-9 years), functional class, number of swollen and tender joints, and levels of C-reactive protein. The primary outcome of the study was a comparison of ACR20 (American College of Rheumatology function class: a 20% or greater improvement in swollen and tender joint counts and a 20% or greater improvement in 3 of patient's global assessment, physician's global assessment, physical disability score (as measured by the Health Assessment Questionnaire [HAQ] disability index), erythrocyte sedimentation rate, and patient's assessment of pain) at 54 weeks between groups. At the end of the 54 weeks, 52% of the infliximab-treated patients had achieved an ACR 20% response compared with 17% of MTX-only controls (P<0.001). Improvement with infliximab was also evident when comparing patient outcomes in terms of ACR 50% (33% vs. 18%) and ACR 70% (18% vs. 3%).
5.3.2 Overview of Model

The framework of our Markov decision-analytic model is presented as a schematic diagram in Figure 5.1 and is outlined in Appendix I. Consistent with the inclusion/exclusion criteria of the ATTRACT trial, the decision model was used to compare the costs and effects of two different drug therapy strategies for adult patients with RA refractory to standard therapy including MTX: 1) intravenous infliximab 3mg/kg every 8 weeks and intravenous MTX of at least 12.5mg/week; and 2) continued usual RA management with MTX as described above. The time horizon for our model was ten years since we assumed that this would likely be the minimal time frame that infliximab would be used in clinical practice and the perspective of both the costs and outcomes were from society. We utilized a 3% discount rate for both costs and QALYs as generally recommended in recent guidelines.

The infliximab plus MTX strategy was based on the pooled results across the three infliximab treatment arms in the ATTRACT trial. Similar to Wong et al., because of the similarity of the observed outcomes in the different doses and dosing intervals for the infliximab arms in the ATTRACT trial, these were pooled to estimate the health state transitions associated with infliximab treatment. Treatment with infliximab was assumed to be continuous during the ten year model unless, during individual patient simulations, three consecutive months residing in HAQ states > 2.0 were observed. At this point, the costs of infliximab were doubled (indicating a shortening in the dosage interval from 8 weeks to 4 weeks which has been shown to be the course most rheumatologists would take considering the lack of response. Finally, if no improvement in HAQ was made after a further three consecutive months, it was assumed that infliximab would be discontinued and the MTX
alone strategy would be adopted. The MTX alone strategy was based on the results achieved from the MTX plus placebo arm in the ATTRACT trial.\textsuperscript{6}

In defining health states for our Markov model, we utilized the HAQ in increments of 0.125 (scores range from 0, which is perfect function, to 3.0, which is maximum functional impairment in 0.125 increments) as discrete health states (Figure 5.1). The HAQ is the best predictor of mortality in RA,\textsuperscript{10} work disability,\textsuperscript{11} and health care resource utilization.\textsuperscript{12} Therefore, within the model, transitions were made between different HAQ-defined health states while costs and outcomes (QALYs) were accrued according to the treatment strategy. All cause mortality functioned as the absorbing state in the model.

5.3.3 Transition Probability Matrices and Statistical Modeling

Wong et al. published transition probabilities based upon an on-treatment analysis from the ATTRACT trial for the period of baseline to 30 weeks and 30 weeks to 54 weeks.\textsuperscript{8} These investigators grouped the HAQ into four discrete health states (HAQ 0 = no disability; HAQ 0.1 to 1.0 = mild impairment; HAQ 1.1 to 2.0 = moderate impairment; and HAQ 2.1 to 3.0 = advanced impairment) and reported the transitions to and from these health states. Wong et al. also reported their transition probabilities as single state transitions across a 30 week (baseline to 30 weeks) or 24 week (30 weeks to 54 weeks) time frame. Since there was substantial improvement in the first 30 weeks of the ATTRACT trial in both the placebo and the infliximab arms,\textsuperscript{6} we classified this time frame as being an adjustment phase where considerable regression to the mean was occurring in both treatment arms. Thus, we did not utilize this time frame to estimate the long-term transition probabilities. Instead, we utilized Wong et al.'s 30 to 54 week transition probabilities for the MTX (Table 5.1) and infliximab
plus MTX (Table 5.2) strategies as the basis of our long-term transition probability matrix. However, we used the baseline HAQ distribution for all arms in the ATTRACT trial as the baseline HAQ distribution for entry into the model.

Therefore, the underlying mathematical model that we adopted was a continuous time Markov process. As such, transitions can occur at any time, not just at discrete weekly or monthly time points and the state of the patient with respect to RA is fully defined by knowing which of 25 HAQ states she/he is in at any time. There may be some error, however, in the measurement of the HAQ state. The true HAQ is the measured HAQ +/- an error. The transitions in HAQ state at any time-point are always to one of the neighboring HAQ states (i.e. from HAQ 0.250 to HAQ 0.375 or HAQ 0.125). The interval between transitions is assumed to be exponentially distributed with a rate parameter that depends on the current HAQ; i.e. the distribution of between event times is:

\[ r \exp(-r t) \]

(\text{where} \ "r" \ \text{is the transition rate and is a function of the HAQ level from which the transition will occur}).

The transition rate has three multiplicative components: i) a purely random fluctuation component which is equal in both directions, but which is larger for HAQ scores in the middle of the range than at the extremes; ii) a systematic excess tendency for drift in either an upper or a downward direction; and iii) a factor which allows the systematic drift to increase or decrease (and even reverse) across the range of possible HAQ states. This assumption reduces the 600 (24*25) independent transition probabilities in an unstructured
model, to 5 parameters in this structured transition rate model. The observed 4 x 4 transition matrix (from Wong et al.) arises by running the underlying model for 24 weeks and then collapsing the resultant 25 x 25 transition matrix to 4 x 4. The 5 parameters were estimated by standard maximum likelihood methods for non-linear models (S-Plus® 6.1 for Windows procedure NLMIN, but with a more stringent convergence criterion than the default). The estimates for the two treatment strategies were carried out independently. Mortality over six months was ignored in these calculations. The 25 x 25 weekly transition probability matrices for MTX and infliximab plus MTX strategies are shown in Table 5.3 and Table 5.4, respectively. Mortality was then estimated from other data (see below) and superimposed on the HAQ transition rate model (see the Appendix I for details of the model and Appendix II for the C-code for the model).

5.3.4 Mortality Rate

We modeled the all-cause mortality rate from a large, dataset spanning from 1974 and continuing through 1999, consisting of 1,922 consecutive RA patients seen at the Wichita (Kansas) Arthritis Center, an outpatient rheumatology clinic. Demographic, clinical, laboratory, and self-report data (including HAQ) were obtained at each follow-up clinic visit. The details of this data set in regard to mortality have been reported previously.\textsuperscript{10,13} The death rate was calculated using Poisson regression using time at risk as an offset variable. The covariates in the model were age, age-squared, the HAQ and HAQ-squared.\textsuperscript{6} From this data, we determined the probability of death by HAQ state as people aged during the simulation (Table 5.5). All patients were assumed to be 52.6 ± 9.2 years of age at the time of initiating the treatments as this was the mean age of enrollment in the ATTRACT study.\textsuperscript{6} The
regression model fit was assessed by examining the model deviance divided by its degrees of freedom.

5.3.5 Utilities and QALYS

The primary outcome measure used in the analysis was QALYs over a ten year period associated with the use of either treatment strategy. Utility weights for the determination of QALYs were calculated using multiple linear regression models assessing the relationship between the HAQ score, age and utility values. The models were estimated using baseline data from a longitudinal study of 317 RA patients comparing different utility instruments (the HUI2, the HUI3, the EQ-5D, and the SF-6D). Details of this study can be found elsewhere.\textsuperscript{14,15} The model fit was assessed using $R^2$ and residual plots were used to assess the fit of each model.

5.3.6 Cost Estimation

Our cost analysis was performed from the Canadian societal perspective. The cost components include both the direct medical costs and indirect costs incurred by loss of work due to RA. The methodology used to calculate each cost category is outlined below. All costs were deflated by the Consumer Price Index for healthcare products and are in 2002 Canadian dollars.\textsuperscript{16} Unit costs and other equations in our model are summarized in Table 5.5.

5.3.6.1 Direct Drug Costs

According to Schering Canada, infliximab is marketed at a cost of $\text{SCAN}2002\, 909.51 / 100 mg vial). Since the infliximab strategy used in our model (and in clinical practice) is
dosed based upon body weight, we applied a weight of 66kg as reported in another Canadian clinical trial of RA drug therapy\textsuperscript{17} resulting in the use of two vials every eight weeks. However, a report out of the United States suggested that the average weight of infliximab users in clinical practice was 77kg.\textsuperscript{9} Furthermore, Malone et al. revealed in their report that 78\% of clinicians surveyed gave patients the entire vial when the calculated dose based on weight was less than the entire vial.\textsuperscript{9} Thus, we used a blended cost assuming that 67\% received 2 vials and 33\% received 3 vials of infliximab. This assumption was tested in univariate sensitivity analysis. In addition, the costs of pharmacy (for preparation), nursing (for preparation and monitoring), and baseline TB screening tests consisting of a chest X-ray, a PPD skin test, and a rheumatologist follow-up visit (for the infliximab strategy) were obtained from the provincial medical fee guide and included in the model. The cost of MTX is $1.00 per 2.5mg tablet and $9.75 per 50mg vial. An analysis of MTX prescriptions in 2000 for all RA patients in the province of British Columbia based on a population-based RA cohort showed that 90\% of MTX prescriptions were oral tablets and 5\% were injectable solution with preservative and 5\% were injectable solution without preservative.\textsuperscript{18} The monitoring costs of anti-nuclear antibodies and anti-DNA antibodies done twice a year were included in the infliximab strategy whereas other monitoring costs were assumed to be the same across the two strategies.

5.3.6.2 Other Direct Costs

Other direct costs beside drug cost included in this study were derived from a longitudinal study of 1063 Canadian patients who reported semi-annually on their health services utilization over the preceding 6 months during 1983 and 1994. A detailed description of the determination of these costs is available from the literature.\textsuperscript{19} The direct
costs of RA care were comprised of long-term care, rehabilitation, nursing homes, health professional visits, medications, diagnostic tests, acute hospitalization, emergency department visits, ambulance services, dialysis and outpatient surgeries. Using a subset of the database, a mixed-effect regression model to estimate direct cost (log transformed) over the next 6-month period was generated where the predictors were gender (fixed effect), disease duration (fixed effect) and HAQ (random effect) at 0 month and within-patient correlation of observations over time were adjusted. These costs were divided by 24 to give average weekly costs by HAQ health state for input into the Markov model with weekly state transitions.

5.3.6.3 Indirect Costs

Indirect cost caused by work disability due to RA was estimated from a prospective longitudinal cohort of 120 employed RA patients recruited in Ontario, Canada from September 1999 to December 2001. In the self-report questionnaire, participants were asked the number of days missed due to RA in the past 6 months and their regular weekly working hours. Participants' disability level was assessed by items drawn from the HAQ and the Multidimensional Functional Assessment Questionnaire and supplemented with additional items to assess discretionary activities such as hobbies, leisure pursuits. A multiple linear regression model was constructed between work capacity and the disability score. Using the disability score as a proxy of the HAQ score, we estimated the work capacity for our study groups based on the baseline HAQ and improvement in HAQ. A gender-weighted average income of Canadian population aged from 45-64 was multiplied by work missed to estimate the cost of lost work capacity. For the model, once the age of the cohort was
equivalent to 65, these indirect costs were no longer accrued as it was assumed that patients would have retired.

5.3.7 Survival Analysis

From the 100,000 simulations described below (50,000 for each treatment strategy), we conducted Kaplan Meier survival analysis and Cox regression. Since time at risk and the occurrence of death as a binary variable were tracked, we estimated the probability of survival over the 10 years of the model. The log-rank test was used to test the null hypotheses that the survival time between the treatment strategies was the same. Cox regression was used to determine the hazard ratio associated with the use of infliximab plus MTX as compared to MTX alone. Right censoring was used for those who had not died after the 10 year time horizon of the Markov model.

5.3.8 Cost-Utility and Probabilistic Analysis

The analysis of the state-transition model provides expected costs and expected QALYs over a 10 year follow-up period. If the infliximab strategy was both more effective and more costly, we calculated the incremental cost-utility ratio of the additional cost per QALY. To quantify the precision of our cost-utility estimates, we conducted probabilistic analyses. We utilized both 1st order (random walk) and 2nd order (random draws from specified distributions) Monte Carlo simulation methods. For the 2nd order simulations, we conducted 1000 iterations. For each of the 2nd order iterations with sampling from the specified distributions, 50 random walks were conducted for each strategy. Therefore, a total
of 100,000 simulations were conducted (50,000 per treatment strategy). For this model, probability distributions were defined for three sets of key model parameters: 1) gender was assumed to follow a Bernoulli; 2) baseline age was assumed to follow a normal distribution; and 3) baseline HAQ scores were assumed to follow the distribution at randomization in the MTX arm of the ATTRACT trial.6,8 The variables were chosen as they were key variables in the cost calculation (Table 5.5) and the baseline HAQ distribution sets the starting point for the transition probability matrices.

The 95% confidence region surrounding the incremental cost-utility ratio was estimated using Fieller’s theorem.22 This analysis was also used to generate plots on the cost-effectiveness plane using each of the indirect utility measurements and to generate cost-acceptability curves with the often quoted threshold of society’s willingness to pay (WTP) of $50,000 per QALY as the ceiling ratio.23

5.3.9 Univariate Sensitivity Analysis

For parameter estimates that were uncertain but where evidence of prior probability distributions was uncertain or inapplicable, we conducted deterministic, univariate sensitivity analysis. Specifically, we calculated the cost per QALY by varying the discount rate from 0%, 3% (base case), 5%, and 7%. In addition, to account for the potential of higher doses of infliximab being used to achieve the same benefit (Malone et al. reported that approximately 1/3 of infliximab patients receive a higher dose than 3 mg/kg every eight weeks9), we varied the weekly cost of infliximab from $200 to $500 per week.
5.4 RESULTS

5.4.1 Simulation Results

Under the assumptions of our model, the mean final HAQ states for those still alive after 10 years or those who expired during the simulations were $2.40 \pm 0.41$ for the MTX alone strategy and $1.38 \pm 0.92$ for the infliximab plus MTX strategy ($p<0.0001$ by student’s t-test). The Kaplan-Meier survival curve from the analysis of the 100,000 simulated patients is presented in Figure 5.2. The result of the log-rank test revealed that there was a significant benefit in survival by using infliximab plus MTX over MTX alone ($p<0.0001$). The hazard ratio associated with infliximab plus MTX was 0.63 (95% confidence interval 0.62 to 0.65, $p<0.0001$) when compared to MTX alone.

Similar to Wong et al., to compare the benefit of infliximab as predicted by our model compared to that achieved in the ATTRACT trial, we determined the predicted mean HAQ score for the infliximab plus MTX and the MTX alone arms after 54 weeks. Our model predicted a mean difference in improvement in HAQ score of 0.4 for infliximab plus MTX versus MTX alone after 54 weeks which was identical to that observed in the ATTRACT trial.

5.4.2 Utility and QALY Values

The results of the multiple linear regression analyses of the indirect utility measures and HAQ are presented in Table 5.6. The discounted QALYs produced by using these equations in the Markov model by treatment strategy are provided in Table 5.7. The SF-6D produced the highest estimations of QALYs secondary to its high lower bound (0.30) as
compared to the other three instruments which permit utility values less than zero (presumably, those health states valued less than death). The HUI3 produced the largest incremental difference between the infliximab plus MTX and the MTX alone strategies.

5.4.3 Cost-Utility and Probabilistic Analysis

The results of the expected costs, expected QALYs and the incremental cost-utility ratios (with 95% confidence limits generated by the 1st and 2nd order Monte Carlo simulations) of using the infliximab plus MTX over the MTX alone strategy are presented in Table 5.8. The mean incremental cost per QALY was the highest for utility weightings provided by the SF-6D compared to the lowest for utility weighting provided by the HUI3 with the HUI2 and EQ-5D utility weightings providing estimates in the middle of these results. The results of the probabilistic analysis are shown graphically on the cost-effectiveness plane in Figure 5.3. Finally, in Figure 5.4, the cost-acceptability curves for each of the indirect utility measures are shown. These results suggest that for ceiling ratios below $50,000, the HUI3 and EQ-5D would most likely yield results below this figure (100% and 99% probability, respectively) as compared to the HUI2 and SF-6D (8% and 0% probability, respectively). For a ceiling ratio of $100,000 per QALY, the results indicate that estimates obtained with any of the indirect utility estimates would be judged to be cost-effective (100% probability).
5.4.4 Traditional Sensitivity Analysis

Results of the univariate sensitivity analysis of varying discount rates and the cost of infliximab are presented in Table 5.9. The incremental cost-utility ratios are relatively robust to the different discount rates. However, increasing the cost of infliximab causes a large increase in the incremental cost per QALY for all of the indirect utility measures.

5.5 DISCUSSION

Our analysis reveals that there is considerable variation in the incremental cost per QALY of using different indirect utility instruments as weightings for QALY estimation in economic evaluation of new therapies for RA. It appears that the SF-6D yields the least optimistic while the HUI3 yields the most optimistic incremental cost per QALY gained. These findings were further supplemented by the results from the cost-acceptability curves which showed that, under a ceiling ratio of society’s WTP of $50,000 per QALY, the HUI3 and EQ-5D based results had a 100% probability of being under this limit. Also, under the assumptions of our model, we demonstrated that the addition of infliximab to MTX in patients with refractory RA results in an improvement in both quality of life (regardless of the indirect utility measurement technique employed) and survival. However, this benefit comes at an increased cost which is due mostly to the acquisition cost of the drug.

Recently, a significant amount of attention has been paid to the fact that the available indirect utility instruments could result in drastically different results when applied to the calculation of QALYs in rheumatology. Conner-Spady et al. administered the EQ-5D, the SF-6D, and the HUI3 in a consecutive sample of rheumatology patients (98 patients were included in the analysis; of these, 51% had RA with the remainder having other
rheumatological conditions) at baseline, 3, 6 and 12 months. They calculated a theoretical QALY by summing the average QALY by instrument for each time interval (for example, 0 to 3 months). They found that the EQ-5D derived QALYs were larger in those reporting better health than HUI3 or SF-6D derived QALYs. The authors concluded the three indirect tools they tested were not interchangeable which could have important ramifications for economic evaluations. The analysis by Luo et al. also conducted in a relatively small sample (n=114) of patients with a variety of rheumatic diseases concluded that the HUI3 and the EQ-5D performed “equally well” in measuring utilities in rheumatic diseases based upon assessment of construct validity. While we agree that both of these instruments have construct validity in RA, we agree with Conner-Spady et al. that they are clearly not interchangeable for use in economic analysis as QALY weightings.

Our analysis, which is based on indirect utility assessment using the HUI2, HUI3, SF-6D, and EQ-5D from a sample of 317 patients with rheumatologist-confirmed RA, is the first attempt at quantifying the differences in indirect utility weightings for QALY measurement in an actual economic evaluation in rheumatic arthritis. From the incremental cost-utility results, it can be seen that there was a relative difference of over 100% in incremental cost per QALY between the lowest (from the HUI3 derived QALYs) and highest (from the SF-6D derived QALYs) estimates. In addition, the range in incremental cost per QALYs generated spans the often-quoted threshold of $50,000 per QALY for programs to be funded, making decision-making difficult. For example, using HUI3 or EQ-5D generated QALYs and the $50,000 per QALY threshold as shown in the cost-acceptability curves (Figure 5.4), decision-makers might determine that infliximab is an economically attractive strategy. However, the same model yields less optimistic findings with the HUI2, and the SF-6D derived QALYs potentially resulting in a conflicting decision. Thus, given the wide
range of incremental cost per QALYs generated by using the various instruments, policies regarding new medications should not be based on a single measure wherever possible. At the very least, economic evaluations should either attempt to explore this issue in sensitivity analysis or, the choice of outcome measures should be standardized across economic evaluations of rheumatoid arthritis.

These finding substantiate the conjecture by Conner-Spady et al. that there might be important implications in economic evaluation by employing different indirect utility instruments. Our analysis builds on those previously described as the utility estimates are derived from a much larger, homogeneous (all patients had RA) sample. In addition, we utilized all four of the most commonly utilized indirect utility instruments thus permitting a complete comparison amongst them. Our findings are similar to those who have compared the outcomes in economic analyses using different indirect utility measures. For example, a study by Neumann et al., showed that incremental cost-effectiveness estimates for a new drug for Alzheimer’s disease were more economically attractive when using the HUI3 as compared to the HUI2.

With respect to the secondary objective of our study, namely to determine the incremental cost per QALY in adding infliximab to MTX in the treatment of refractory RA, we found that the cost per additional discounted QALY by adopting the infliximab strategy under the assumptions of our model ranged from $38,161 to $58,991 depending on the utility weighting method utilized. Our results generally agree with other evaluations of the cost-effectiveness of infliximab that have been recently published in that, under certain conditions, the infliximab strategy could be construed to be economically attractive, although there are important methodological differences. All of these analyses (including ours) make use of a Markov model based upon states derived from the Health Assessment
Questionnaire (HAQ) and utilize the ATTRACT trial as the primary source of clinical data. Wong et al. extended the results from the 54 week follow-up from the ATTRACT trial over the lifetime of RA patients. These authors found that, from the perspective of society, the infliximab strategy had an incremental cost-utility ratio of $9,100 (US) per QALY gained versus MTX monotherapy. Kobelt et al. conducted a similar analysis from the Swedish healthcare perspective over a 1 year and a 2 year period. These investigators found, under the assumptions of their model, the incremental cost-utility ratio for the infliximab over the MTX strategy was 3440 EURO per QALY (from the Swedish societal perspective) and 34,800 EURO per QALY (from the UK societal perspective).

Much of the differences in the results across these studies can be explained by basic differences in the construction of Markov models. Wong et al. utilized a 4x4 transition probability matrix, Kobelt et al. utilized a 6x6 matrix, while ours was 25x25 (with a continuous, underlying process) allowing for increased sensitivity in the relationship between HAQ, cost and QALYs. In addition, the utility weightings employed by each of the studies were different. Wong et al. utilized a visual analogue scale (VAS) with death equal to zero and one equal to perfect health whereas Kobelt et al. utilized the EQ-5D. In both cases, the sample size from which these values were derived or the methodology to integrate them into the Markov model was not clear. In addition, while the VAS is a preference-based measure, it is not a choice-base method and thus not a utility. The comparison of these two studies is a perfect illustration how weightings for QALYs vary considerably across economic analysis even for the same drug therapy.

Our analysis assumes that infliximab is continued for ten years for those that respond to therapy. This assumption is not unreasonable as it is conceivable that these refractory patients will continue to receive a drug therapy that is as costly as infliximab for the duration
of their disease if it is successful. However, there are several limitations to our study. Since
the available results for the ATTRACT trial only account for 54 weeks of follow-up, the
transition probabilities could only be estimated from within this time window. We have
ignored the occurrence, costs and outcomes associated with adverse events in the infliximab
arm which could include infusion reactions, superficial upper respiratory tract infections,
demyelination, and serious opportunistic infections. However, in the ATTRACT trial, there
was no difference in the number of serious adverse events (requiring hospitalization or
judged to be life-threatening) between the two treatment groups.\(^6\) Many adverse events for
the tumor necrosis factor alpha inhibitors were identified due to post-marketing surveillance.
As such, due to either the lack of significant medical intervention required for the
management of these adverse events, their low probability of occurrence, their elimination
through monitoring that we have built into the model (ie. for example, activation of latent TB
is a concern with anti-TNF alpha therapy but we have included screening mechanisms prior
to treatment for those in the infliximab strategy) and/or their negligible impact on utility, we
determined that the costs incurred to manage this complication and changes in utility would
not affect the overall model.

A recent study has shown that patients encountered in clinical practice are quite
different than those enrolled in the ATTRACT trial.\(^{28}\) In fact, in those patients with “long-
term” RA in a clinical practice, only 5% would have fit the inclusion/exclusion criteria of the
ATTRACT trial. This finding may seriously limit the applicability of all of the infliximab
cost-utility analyses to the majority of patients encountered in clinical practice. A recent
editorial by Wolfe et al., outlines the potential pitfalls of extrapolating randomized controlled
trial data in the conduct of long-term cost-effectiveness studies.\(^{29}\) Further research in this
regard is warranted.
Other limitations with the model involve the use of HAQ health states (in 0.125 increments) as the underlying predictor of utility, mortality and work disability. Wolfe recently outlined the non-linear nature of the HAQ (changes in the HAQ score in the range of 1 to 2 represent much less change in function than changes in the HAQ score in range from 0 to 1). While we did not assume linearity in the HAQ in terms of the probability calculation, we made this assumption in the regression models. However, as stated by Wolfe, the HAQ is "a good, sensitive questionnaire, the best we have to date and one that has stood the test of time." Thus, despite its limitations, it is still likely the best method currently available in defining RA health states for transition models as we describe.

The use of different indirect utility measurement methods as weightings for QALYs yields quite different incremental cost-utility ratios in the economic evaluation of new therapies for RA. Thus, these differences should be explored in sensitivity analysis or, the choice of outcome measures should be standardized across economic evaluations of rheumatoid arthritis.

Infliximab as add on therapy to patients with RA who are refractory to MTX results in additional years of life even when an adjustment for quality is made. Depending on the method of measurement of utility adopted and the ceiling ratio of society’s WTP for a QALY, infliximab may represent good value in certain health care environments.
5.6 REFERENCES


Arthritis Rheum 2002;46(suppl.):s95.


TABLE 5.1: OBSERVED TRANSITION PROBABILITY MATRICES FOR METHOTREXATE FROM THE ATTRACT TRIAL (FROM WEEK 30 TO WEEK 54)

<table>
<thead>
<tr>
<th>HAQ SCORE GROUPS*</th>
<th>0</th>
<th>0.1 – 1.0</th>
<th>1.1 – 2.0</th>
<th>2.1 – 3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5</td>
<td>0.143</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.1 – 1.0</td>
<td>0.5</td>
<td>0.524</td>
<td>0.063</td>
<td>0</td>
</tr>
<tr>
<td>1.1 – 2.0</td>
<td>0</td>
<td>0.333</td>
<td>0.781</td>
<td>0.2</td>
</tr>
<tr>
<td>2.1 – 3.0</td>
<td>0</td>
<td>0</td>
<td>0.156</td>
<td>0.8</td>
</tr>
</tbody>
</table>

*HAQ transition states defined as 0 = no impairment; 0.1 to 1 = mild impairment; 1.1 to 2 = moderate impairment; 2.1 to 3 = advanced impairment
TABLE 5.2: OBSERVED TRANSITION PROBABILITY MATRICES FOR INFLIXIMAB FROM THE ATTRACT TRIAL (FROM WEEK 30 TO WEEK 54)

<table>
<thead>
<tr>
<th>HAQ SCORE GROUPS*</th>
<th>0</th>
<th>0.1 - 1.0</th>
<th>1.1 - 2.0</th>
<th>2.1 - 3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.679</td>
<td>0.079</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.1 - 1.0</td>
<td>0.286</td>
<td>0.822</td>
<td>0.158</td>
<td>0</td>
</tr>
<tr>
<td>1.1 - 2.0</td>
<td>0.035</td>
<td>0.089</td>
<td>0.806</td>
<td>0.342</td>
</tr>
<tr>
<td>2.1 - 3.0</td>
<td>0</td>
<td>0.01</td>
<td>0.036</td>
<td>0.658</td>
</tr>
</tbody>
</table>

*HAQ transition states defined as 0 = no impairment; 0.1 to 1 = mild impairment; 1.1 to 2 = moderate impairment; 2.1 to 3 = advanced impairment.
TABLE 5.3: CALCULATED WEEKLY TRANSITION PROBABILITY MATRIX FOR METHOTREXATE

The upper left hand corner represents the probability of transition from HAQ 0 to HAQ 0 in a one week time frame \((p_{11})\). The rows represent the probability of transition from HAQ states in increments of 0.125 (for example, the second row, first column represents the probability of a transition from HAQ 0.125 to HAQ 0 \([p_{21}]\)). Similarly, the columns are in 0.125 increments in HAQ (for example, the first row represent probabilities of transition to a HAQ 0, the second row represents transitions to HAQ 0.125. The bottom right hand corner represents the probability of transitioning from HAQ 3.0 \([p_{25.25}]\).
The upper left hand corner represents the probability of transition from HAQ 0 to HAQ 0 in a one week time frame \((p_{1,1})\). The rows represent the probability of transition from HAQ states in increments of 0.125 (for example, the second row, first column represents the probability of a transition from HAQ 0.125 to HAQ 0 \([p_{2,1}]\)). Similarly, the columns are in 0.125 increments in HAQ (for example, the first row represent probabilities of transition to a HAQ 0, the second row represents transitions to HAQ 0.125. The bottom right hand corner represents the probability of transitioning from HAQ 3.0 \([p_{25,25}]\).
TABLE 5.5: UNIT COSTS (IN CANADIAN DOLLARS), OTHER PARAMETERS AND EQUATIONS IN THE MARKOV MODEL

<table>
<thead>
<tr>
<th>Model Parameter</th>
<th>Parameter</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infliximab*</td>
<td>$264.89/week</td>
<td>(exp (6.49 -1.18<em>age + 0.15</em>(female)+0.39<em>HAQ + 0.5</em>1.66))</td>
</tr>
<tr>
<td>Methotrexate 15mg/week</td>
<td>$6.05/week</td>
<td>(1-workcapacity) * ((47,085*(male) + 33,774*(female)))</td>
</tr>
<tr>
<td>Pharmacist preparation¥</td>
<td>$6.59/week</td>
<td>-</td>
</tr>
<tr>
<td>Nursing monitoring‡</td>
<td>$19.02/week</td>
<td>-</td>
</tr>
<tr>
<td>Laboratory monitoring‡</td>
<td>$5.66/week</td>
<td>-</td>
</tr>
<tr>
<td>Chest X-ray*</td>
<td>$84.62 once</td>
<td>-</td>
</tr>
<tr>
<td>PPD Skin Test*</td>
<td>$8.42 once</td>
<td>-</td>
</tr>
<tr>
<td>Rheumatology Clinic Visit*</td>
<td>$72.62 once</td>
<td>-</td>
</tr>
<tr>
<td>Proportion females</td>
<td>78%</td>
<td>-</td>
</tr>
<tr>
<td>Mean age, yrs/100 (SD)</td>
<td>0.52 (0.09)</td>
<td>-</td>
</tr>
<tr>
<td>6 Month Direct Medical Costs</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Annual Indirect Costs</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Work Capacity</td>
<td>-</td>
<td>1.09 - (0.18*HAQ)</td>
</tr>
<tr>
<td>Annual Death Rate</td>
<td>-</td>
<td>exp(-11.42+(17.05<em>age)-(0.26</em>HAQ)-(7.95<em>age²)+(0.29</em>HAQ²))</td>
</tr>
</tbody>
</table>

*Based on an assumption of 67% receive 2 * 100mg vials and 33% receive 3 * 100mg vials
¥ For the preparation, administration and monitoring of infliximab infusion
‡ For the monitoring of anti-nuclear antibodies and anti-DNA antibodies twice a year
† For the screening of latent tuberculosis prior to infliximab therapy

In the equations, “male” and “female” refer to proportion of that gender as reported in the ATTRACT trial; “HAQ” refers to the disability index from 0 to 3.0; and “age” refers to the age in years divided by 100.
### TABLE 5.6: MULTIPLE LINEAR REGRESSION MODELS OF THE INDIRECT UTILITY MEASURES

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Independent Variable, Beta Coefficient, P-Value</th>
<th>Model R²</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HUI2 Utility Score</strong></td>
<td>Intercept, 0.88, &lt;0.0001</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>HAQ, -0.17, &lt;0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age*, 0.03, 0.71</td>
<td></td>
</tr>
<tr>
<td><strong>HUI3 Utility Score</strong></td>
<td>Intercept 0.83, &lt;0.0001</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>HAQ, -0.29, &lt;0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age*, 0.01, 0.55</td>
<td></td>
</tr>
<tr>
<td><strong>SF-6D Utility Score</strong></td>
<td>Intercept, 0.69, &lt;0.0001</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>HAQ, -0.13, &lt;0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age*, 0.13, 0.002</td>
<td></td>
</tr>
<tr>
<td><strong>EQ-5D Utility Score</strong></td>
<td>Intercept, 0.72, &lt;0.0001</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>HAQ, -0.20, &lt;0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age*, 0.25, 0.008</td>
<td></td>
</tr>
</tbody>
</table>

All models based on a sample size of 317 respondents

*Age is age in years divided by 100
<table>
<thead>
<tr>
<th>Indirect Utility Method</th>
<th>Discounted* Mean QALY ± SD</th>
<th>Mean Difference ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Infliximab plus MTX Strategy</td>
<td>MTX Alone Strategy</td>
</tr>
<tr>
<td>HUI2</td>
<td>5.38 ± 0.89</td>
<td>4.12 ± 0.63</td>
</tr>
<tr>
<td>HUI3</td>
<td>3.74 ± 1.27</td>
<td>1.74 ± 0.72</td>
</tr>
<tr>
<td>SF-6D</td>
<td>4.74 ± 0.69</td>
<td>3.75 ± 0.50</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>4.72 ± 0.94</td>
<td>3.30 ± 0.58</td>
</tr>
</tbody>
</table>

*Discounted at 3% per annum
TABLE 5.8: EXPECTED COSTS AND INCREMENTAL COST-UTILITY RATIOS GENERATED BY THE INDIRECT UTILITY METHODS

<table>
<thead>
<tr>
<th>Strategies</th>
<th>Cost $ per Patient*</th>
<th>95% Confidence Limits‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) MTX alone</td>
<td>$133,737</td>
<td>$131,007 - $136,466</td>
</tr>
<tr>
<td>(2) Infliximab plus MTX</td>
<td>$199,729</td>
<td>$197,747 - $201,713</td>
</tr>
<tr>
<td>Difference (2) – (1)</td>
<td>$65,993</td>
<td>$64,412 - $67,573</td>
</tr>
</tbody>
</table>

Incremental Cost per QALY*

| HUI2-QALY                           | $52,078             | $48,850 - $55,528       |
| HUI3-QALY                           | $33,092             | $30,887 - $35,436       |
| SF-6D-QALY                          | $67,005             | $62,773 - $71,540       |
| EQ-5D-QALY                          | $46,159             | $43,086 - $49,438       |

* Discounted at 3% per annum

* Discounted at 3% per annum

† Generated by application of Fieller’s theorem
TABLE 5.9: UNIVARIATE SENSITIVITY ANALYSIS – INCREMENTAL COST-UTILITY RATIO (INCREMENTAL COST PER QALY) BY INDIRECT UTILITY METHOD

<table>
<thead>
<tr>
<th></th>
<th>Discount 0%</th>
<th>Discount 5%</th>
<th>Discount 7%</th>
<th>Infliximab Cost*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$200 to $500 per week</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUI2- QALY</td>
<td>$56,318</td>
<td>$51,302</td>
<td>$53,762</td>
<td>$40,918 to $147,454</td>
</tr>
<tr>
<td>HUI3- QALY</td>
<td>$39,190</td>
<td>$34,851</td>
<td>$35,951</td>
<td>$27,485 to $98,207</td>
</tr>
<tr>
<td>SF-6D- QALY</td>
<td>$75,142</td>
<td>$84,902</td>
<td>$77,483</td>
<td>$46,616 to $169,237</td>
</tr>
<tr>
<td>EQ-5D- QALY</td>
<td>$53,518</td>
<td>$47,830</td>
<td>$49,609</td>
<td>$37,084 to $138,098</td>
</tr>
</tbody>
</table>

* At the base case discount rate (3%) for both costs and QALYs
"Stage" refers to the Markov transition cycle. For simplicity, only transitions from HAQ 0 are shown.
FIGURE 5.2: KAPLAN-MEIER SURVIVAL CURVES FROM THE 100,000 MONTE CARLO SIMULATIONS

Survival curves for those on infliximab plus MTX (upper blue line) and MTX alone (lower red line). The diamonds (◇) represent those who were censored.
Each ellipse within each utility-defined QALY covers 5%, 50% and 95% of integrated joint density between cost and the QALY differences. The lines are the upper and lower confidence limits using Fieller's theorem. As noted by Briggs et al. (ref. 22), the "wedge" defined by Fieller's confidence limits falls inside the 95% ellipse.
FIGURE 5.4: COST-UTILITY (or COST-EFFECTIVENESS) ACCEPTABILITY CURVES FOR EACH INDIRECT UTILITY MEASURE

Cost-Effectiveness Acceptability Curve (Fieller's Theorem)
APPENDIX I: MARKOV MODEL

To conduct a cost-effectiveness analysis of infliximab plus methotrexate (MTX) compared to MTX alone in the treatment of severe rheumatoid arthritis using the ATTRACT trial as a source of clinical outcomes, we defined a Markov model with 26 health states: 25 states based upon increments of 0.125 in a Health Assessment Questionnaire (HAQ) score from 0 (no disability) to 3.0 (worst level of disability) and one absorbing state, death. The length of our Markov cycle was one week.

An important assumption made was that patients in a given initial health state have a constant probability per unit time of making a transition into any other given state independently of how much time has already passed in the initial state. This assumption, referred to as the Markov property, was essential in modeling the prognosis with a finite number of states. We also assumed that, given that death does not occur, the intermediate transition probabilities from one HAQ state to another are gender and age independent and therefore constant through time. The probability of all-cause mortality in a unit of time for people with rheumatoid arthritis however is a function of the HAQ state and age which we have denoted as \( p(h,a) \). As mentioned, the final transition probabilities were the combination of the intermediate transition probabilities and mortality probability. The transition matrix below describes the transition probabilities in the Markov model (also shown in Tables 5.3 and 5.4).
The utility accrued for cycle $t$, referred to as the cycle sum, was calculated by:

$$\text{Cycle Sum} = \sum_{i=1,2,\ldots,26} P_i(t) \times U_i$$

where $P_i(t)$ was the distribution of patients in the 26 health states at cycle $t$ and $U_i$ was the utility associated with state $i$. Simulations were run for 10 years (520 cycles). The cycle sum was then added to a running total - the cumulative utility - which is what was required for cost-utility analysis.
The distribution of patients at each cycle (i.e. one week) was calculated using the transition matrices $T(h,a)$ and the distribution of patients at age $t_0$, distributed among the 26 health states ($p_1(t_0), p_2(t_0) p_3(t_0) p_4(t_0) p_5(t_0) p_6(t_0) p_7(t_0) \ldots \ldots p_{26}(t_0)$), then the distribution of patients at age $t$ was given by:

\[
(p(t), p_2(t), p_3(t) \ldots p_{26}(t)) = (p_1(t_0), p_2(t_0), p_3(t_0) \ldots p_{26}(t_0)) \prod_{\text{age}=t_0+1}^{t} T(HAQ, age)
\]

where the right hand side of the equation was the matrix product of a row vector $(p_1(t_0), p_2(t_0) p_3(t_0) p_4(t_0) p_5(t_0) p_6(t_0) p_7(t_0) \ldots \ldots p_{26}(t_0))$, and $t-t_0-1$ transition matrices.
APPENDIX II: C-CODE USED TO RUN THE MARKOV MODEL

/* --- The following code was compiled under lcc-win32 --- */
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <TIME.H>
#include "uniform.c"
#include "ltqnorm.c"
define ranunif genrand_real3
define setseed init_genrand
define str(x) #x
define xstr(x) str(x)
define xstr(x) str(x)

//changeable:
define RESTART
#define MM 25
#define MM1 24
#define ENDSTAGE 520 (520 weeks = 10 years)
#define SEED 14357
#define RATE 0.03 (discount rate)
#define BLKN 1000 (for 2nd order Monte Carlo - # of simulations)
#define BLKZ 50 (for 1st order Monte Carlo - # of patients simulated per 2nd order Monte Carlo)
#define CFLX 500 (for univariate sensitivity analysis = cost of infliximab up to 500 dollars)
#define MCCOUT sim500.txt

int haqmcc(void);

int rchoose (double *xx, int mm, double ru)
// chooses the first element of a vector of length mm that equals or exceeds ru
{
    int jj;
    int ii;
    jj = mm-1;
    for (ii = 0; ii <= mm-1; ii++)
        if (ru < xx[ii])
            jj = ii;
    break;
    return jj;
}

// the function that does the simulations:
int haqmcc(void) {

//data input
    char *rnames="block outtime age0 gender dead haqi0 haqi1 finaldose rcost rqaly1 rqaly2 rqaly3 rqaly4 inflix";
    FILE *fpt, *fout;
    double xcol=0; //intermediate, now misnamed
double **ctmtx, **ctflx, **ctmat, *chaq0, *chaq;
    int nread=0;
    int ii, jj, stage, iter, block;

    int haqi, haqi0, trak, dose0, outtime,gender,dead;
double *vhaq=(double *)malloc(MM*sizeof(double));

    178
unsigned long s=SEED;
double age, age0, haq, lpdying, rate, dfact;
double directcost, indirectcost, wcage, cinflix, cmethotrexate=6.05;
double cost, rcost;
double log52=log(52);
double *workcapacity=(double *)malloc(MM*sizeof(double));
double *cinfliximab=(double *)malloc(MM*sizeof(double));
double *bchoices=(double *)malloc(2*sizeof(double));
double util[4] = {0,0,0,0}; double rqaly[4] = {0,0,0,0};

// read matrix ctmtx in
// ctmtx is matrix with columns that are cumulated versions of the columns of the mtx transition matrix
ctmtx=(double **)malloc(MM*sizeof(double *));
ctmtx[0]=(double *)malloc(MM*MM*sizeof(double));
for(ii=1;ii<=24;ii++) ctmtx[i]=ctmtx[ii-1]+MM;
if((fpt=fopen("ctmtx.txt","r"))==NULL)
printf("%s","Error opening file ctmtx for reading");
for(ii=0;ii<=24;ii++) {
for(jj=0; jj<=24; jj++) {
 nread=nread+fsscanf(fpt,"%le", &xcol);
 ctmtx[jj][ii]=xcol; //column major order
}}
fclose(fpt);

// read matrix ctflix in
// ctflix is formed from the transition matrix for the infliximab group
nread=0;
ctflix=(double **)malloc(MM*sizeof(double *));
ctflix[0]=(double *)malloc(MM*MM*sizeof(double));
for(ii=1;ii<=24;ii++) ctflix[i]=ctflix[ii-1]+MM;
if((fpt=fopen("ctflix.txt","r"))==NULL)
printf("%s","Error opening file ctflix for reading");
for(ii=0;ii<=24;ii++) {
for(jj=0; jj<=24; jj++) {
 nread=nread+fsscanf(fpt,"%le", &xcol);
 ctflix[jj][ii]=xcol; //column order
}}
fclose(fpt);

// read vector chaq0 in
// baseline cumulative HAQ distribution
nread=0;
chaq0=(double *)malloc(MM*sizeof(double));
if((fpt=fopen("newchaq0.txt","r"))==NULL)
printf("%s","Error opening file chaq0 for reading");
for(ii=0;ii<=24;ii++) {
 nread=nread+fsscanf(fpt,"%le", &xcol);
 chaq0[ii]=xcol;
}
fclose(fpt);

#ifdef RESTART
fout=fopen(xstr(MCCOUT),"w");
fprintf(fout,"%s",rnames);
#else

#endif

179
fout=fopen(xstr(MCCOUT),"a");
#endif /* RESTART */

setseed(s);
bchoices[1]=1.0;

for(ii=0;ii<=MMl;ii++) {
    vhaq[ii]=(double)ii/8; // possible haq values
    workcapacity[ii]=1.09132 - 0.18404*vhaq[ii];
    cinfliximab[ii]= (CFLX *0.67) + (2*CFLX*0.33) + 25.50; // double vial use
}

for(block=1;block<=BLKN;block++) {
    //constants that are not altered during the Markov process can go here.
    ageO = .52+0.092*ltqnorm(ranunif()));
    bchoices[0]=1-0.2243;
    gender=rchoose(bchoices,2,ranunif());
    haqiO=rchoose(chaqO,MM,ranunif());
    iter=1;

    for(dose0=0;dose0<=l; dose0++) {
        for(iter=1;iter<=BLKZ; iter++) {
            // initial values for quantities that do change, go here.
            haqi=haqiO;
            haq=vhaq[haqiO];
            dead=0;
            dose=dose0;
            age=ageO;
            ctmat=(dose==0) ? ctmtx : ctflx;
            rcost=0;
            for(ii=0;ii<4;ii++) rqaly[ii]=0;
            outtime=0;
            for(stage=0;stage < ENDSTAGE ;stage++) {
                lpdying=-11.4184+(17.0462*age)-(0.2582*haq)-(7.9548*age*age)+(0.2903*haq*haq)-log52;
                dead=(ranunif()<exp(lpdying));
                if(dead==1) break;
                dfact= 1 /pow(( 1 +RATE),(stage/52));

                // weekly costs
                directcost=(exp (6.4924-1.183*age+0.1459*(1-gender)+0.3918*haq+0.5*1.6663))/26;
                if(age<=0.65) wcage=workcapacity[haqi]; else wcage=1;
                indirectcost=((1-wcage)*(47085*(1-gender) + 33774*gender))/52;
                cost=(directcost+indirectcost+dose*cinfliximab[haqi]+cmethotrexate)*dfact;

                util[0]= 0.8286 - 0.2947*haq +0.0545*age; //ii+1: 1=HUI3, 2=SF-6D; 3=HUI2, 4=EQ5D
                util[1]= 0.6863 - 0.126*haq +0.1316*age;
                util[2]= 0.8759 - 0.1673*haq + 0.0263*age;
                util[3]= 0.7207 - 0.197*haq + 0.247*age;
                for(ii=0;ii<=3;ii++) {
                    util[ii]=util[ii]*dfact; // if dead, utility, cost are zero
                    rqaly[ii]=rqaly[ii]+util[ii]/52;
                }
                rcost=rcost+cost;

            }
        }
    }
}

180
if(dose>=1) { //if dose==0, infliximab is permanently discontinued, so no need to track
if(haq>2.0) trak=trak+1; else trak=0;
if(trak==12) {dose=(dose+1)%3;cinflux=dose*cinfluximab[haqi];trak=0;}
}
if(dose<1) cmum=cmumx; // "else" would give wrong result. Would fail to change cmum.
age=age + 1/5200;
chaq=ctmat[haqi];
haqi=rchoose(chaq,MM,ranunif());
haq=vhaq[haqi];
outtime=stage;
} fprintf(fout,"%d %d %f %d %d %d %d %i %f %f %f %d",
block, outtime, ageO, gender, dead, haqiO, haqi, dose,
(long)rcost, rqaly[0], rqaly[1], rqaly[2], rqaly[3], dose0);
//printf("%f",clock()/CLOCKS_PER_SEC);
}
if(block % 50 == 0) printf("%d",block);
fprintf(fout,"\n")
fclose(fout);
return 0;
}

int main(void){
    int rtv;
    time_t t1,t2,tt;
    t1=time(&tt);
    rtv=haqmcc();
    t2=time(&tt);
    printf("\n\n\n%f",difftime(t2,t1));
    return rtv;
}

*For calculation of ltnorm
* Lower tail quantile for standard normal distribution function.
* This function returns an approximation of the inverse cumulative
* standard normal distribution function. I.e., given P, it returns
* an approximation to the X satisfying P = Pr\{Z <= X\} where Z is a
* random variable from the standard normal distribution.
* The algorithm uses a minimax approximation by rational functions
* and the result has a relative error whose absolute value is less
* than 1.15e-9.
* Author: Peter J. Acklam
* Time-stamp: 2002-06-09 18:45:44 +0200
* E-mail: jacklam@math.uio.no
* WWW URL: http://www.math.uio.no/~jacklam
* C implementation adapted from Peter's Perl version
*/

#include <math.h>
#include <errno.h>
#include <stdio.h>
/* Coefficients in rational approximations. */
static const double a[ ] =
{
  -3.969683028665376e+01,
  2.209460984245205e+02,
  -2.759285104469687e+02,
  1.383577518672690e+02,
  -3.066479806614716e+01,
  2.506628277459239e+00
};
static const double b[ ] =
{
  -5.447609879822406e+01,
  1.615858368580409e+02,
  -1.556989798598866e+02,
  6.680131188771972e+01,
  -1.328068155288572e+01
};
static const double c[ ] =
{
  -7.784894002430293e-03,
  -3.223964580411365e-01,
  -2.400758277161838e+00,
  -2.549732539343734e+00,
  4.37466414146968e+00,
  2.938163982698783e+00
};
static const double d[ ] =
{
  7.784695709041462e-03,
  3.224671290700398e-01,
  2.445134137142996e+00,
  3.754408661907416e+00
};
#define LOW 0.02425
#define HIGH 0.97575

double
ltnorm(double p)
{
  double q, r;
  errno = 0;
  if (p < 0 || p > 1)
  {
    errno = EDOM;
    return 0.0;
  }
  else if (p == 0)
  {
    errno = ERANGE;
    return -HUGE_VAL /* minus "infinity" */;
  }
} else if (p == 1)
{
    errno = ERANGE;
    return HUGE_VAL /* "infinity" */;
}
else if (p < LOW)
{
    /* Rational approximation for lower region */
    q = sqrt(-2*$log(p)$);
    return ((((c[0]*q+c[1])*q+c[2])*q+c[3])*q+c[4])*q+c[5]) /
    ((((d[0]*q+d[1])*q+d[2])*q+d[3])*q+1);
}
else if (p > HIGH)
{
    /* Rational approximation for upper region */
    q = sqrt(-2*log(1-p));
    return -((((c[0]*q+c[1])*q+c[2])*q+c[3])*q+c[4])*q+c[5]) /
    ((((d[0]*q+d[1])*q+d[2])*q+d[3])*q+1);
}
else
{
    /* Rational approximation for central region */
    q = p - 0.5;
    r = q*q;
    return (((((a[0]*r+a[1])*r+a[2])*r+a[3])*r+a[4])*r+a[5])*q /
    ((((b[0]*r+b[1])*r+b[2])*r+b[3])*r+b[4])*r+1);
}
/* calculates uniform
A C-program for MT19937, with initialization improved 2002/1/26.
Coded by Takuji Nishimura and Makoto Matsumoto.

Before using, initialize the state by using init_genrand(seed)
or init_by_array(init_key, key_length).
Copyright (C) 1997 - 2002, Makoto Matsumoto and Takuji Nishimura, All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that
the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the
   following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the
   following disclaimer in the documentation and/or other materials provided with the distribution.
3. The names of its contributors may not be used to endorse or promote products derived from this software
   without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS"
AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE
IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR
PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR
CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL,
EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO,
PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR
PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF
LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING
NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS
SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.)

Any feedback is very welcome.
http://www.math.keio.ac.jp/matumoto/emt.html
email: matumoto@math.keio.ac.jp
*/

#include <stdio.h>

/* Period parameters */
#define N 624
#define M 397
#define MATRIX_A 0x9908b0dfUL /* constant vector a */
#define UPPER_MASK 0x80000000UL /* most significant w-r bits */
#define LOWER_MASK 0x7fffffffUL /* least significant r bits */

static unsigned long mt[N]; /* the array for the state vector */
static int mti=N+1; /* mti==N+1 means mt[N] is not initialized */

/* initializes mt[N] with a seed */
void init_genrand(unsigned long s)
{
    mt[0]= s & 0xffffffffUL;
    for (mti=1; mti<N; mti++) {
        mt[mti] =
            (1812433253UL * (mt[mti-1] ^ (mt[mti-1] >> 30)) + mti);
        /* See Knuth TAOCP Vol2. 3rd Ed. P.106 for multiplier. */
        /* In the previous versions, MSBs of the seed affect */
        /* only MSBs of the array mt[]. */
        /* 2002/01/09 modified by Makoto Matsumoto */
        mt[mti] &= 0xffffffffUL;
        /* for >32 bit machines */
    }
}

/* initialize by an array with array-length */
/* init_key is the array for initializing keys */
/* key_length is its length */
void init_by_array(init_key, key_length)
unsigned long init_key[], key_length;
{
    int i, j, k;
    init_genrand(19650218UL);
    i=1; j=0;
    k = (N>key_length ? N : key_length);
    for (; k--; ) {
        m[i] = (mt[i] ^ (mt[i] >> 30)) * 1664525UL);
        /* non linear */
        mt[i] &= 0xffffffffUL; /* for WORDSIZE > 32 machines */
        i++; j++;
        if (i>=N) { mt[0] = mt[N-1]; i=1; }
        if (j>=key_length) j=0;
}
for (k=N-1; k--; ) {
    mt[i] = (mt[i] ^ ((mt[i-1] ^ (mt[i-1] >> 30)) * 1566083941UL))
            - i; /* non linear */
    mt[i] &= 0xffffffffUL; /* for WORDSIZE > 32 machines */
    i++;
    if (i>=N) { mt[0] = mt[N-1]; i=1; }
}

mt[0] = 0x80000000UL; /* MSB is 1; assuring non-zero initial array */

/* generates a random number on [0,0xffffffff]-interval */
unsigned long genrand_int32(void)
{
    unsigned long y;
    static unsigned long mag01[2]={0x0UL, MATRIX_A};
    /* mag01[x] = x * MATRIX_A for x=0,1 */

    if (mti >= N) {/* generate N words at one time */
        int kk;

        if (mti == N+1) /* if init_genrand() has not been called, */
            init_genrand(5489UL); /* a default initial seed is used */

        for (kk=0;kk<N-M;kk++) {
            y = (mt[kk]&UPPER_MASK)|(mt[kk+1]&LOWER_MASK);
            mt[kk] = mt[kk+M] ^ (y >> 1) ^ mag01[y & 0x1UL];
        }
        for (kk<N-1;kk++) {
            y = (mt[kk]&UPPER_MASK)|(mt[kk+1]&LOWER_MASK);
            mt[kk] = mt[kk+(M-N)] ^ (y >> 1) ^ mag01[y & 0x1UL];
        }
        y = (mt[N-1]&UPPER_MASK)|(mt[0]&LOWER_MASK);
        mt[N-1] = mt[M-1] ^ (y >> 1) ^ mag01[y & 0x1UL];

        mti = 0;
    }

    y = mt[mti++];

    /* Tempering */
    y ^= (y >> 11);
    y ^= (y << 7) & 0x9d2c5680UL;
    y ^= (y << 15) & 0xefc60000UL;
    y ^= (y >> 18);

    return y;
}

/* generates a random number on [0,0x7fffffff]-interval */
long genrand_int31(void)
{
    return (long)(genrand_int32()>>1);
}
/* generates a random number on [0,1]-real-interval */
double genrand_real1(void)
{
    return genrand_int32()/(1.0/4294967295.0);
    /* divided by 2^32-1 */
}

/* generates a random number on [0,1)-real-interval */
double genrand_real2(void)
{
    return genrand_int32()/(1.0/4294967296.0);
    /* divided by 2^32 */
}

/* generates a random number on (0,1]-real-interval */
double genrand_real3(void)
{
    return (((double)genrand_int32()) + 0.5)/(1.0/4294967296.0);
    /* divided by 2^32 */
}

/* generates a random number on [0,1) with 53-bit resolution*/
double genrand_res53(void)
{
    unsigned long a=genrandint32()>>5, b=genrand_int32()>>6;
    return(a*67108864.0+b)/(1.0/9007199254740992.0);
}

/* These real versions are due to Isaku Wada, 2002/01/09 added */

/*
int main(void)
{
    int i;
    unsigned long init[4]={0x123, 0x234, 0x345, 0x456}, length=4;
    init_by_array(init, length);
    printf("1000 outputs of genrand_int32()\n");
    for (i=0; i<1000; i++)
        printf("%10lu", genrand_int32());
    printf("\n");
    printf("\n1000 outputs of genrand_real2()\n");
    for (i=0; i<1000; i++)
        printf("%10.8f", genrand_real2());
    return 0;
} */
CHAPTER 6

THE IMPACT OF LOW FAMILY INCOME ON SELF-REPORTED HEALTH OUTCOMES IN PATIENTS WITH RHEUMATOID ARTHRITIS WITHIN A PUBLICLY-FUNDED HEALTH CARE ENVIRONMENT

6.1 FOREWORD

This chapter has been accepted for publication in Rheumatology. The candidate is first author of this manuscript which was co-authored by Dr. Larry Lynd, currently a post-doctoral fellow who had trained under Dr. Anis and whose thesis was on the relationship between socioeconomic status and beta-agonist use. Drs. Aslam Anis and John Esdaile, co-supervisors of the candidate are also included as co-authors of the submitted manuscript.

The candidate’s role in this manuscript was the conception of the problem, data entry and manipulation, co-ordination of study recruitment, development and performance of all statistical analysis, and the writing of the final manuscript.

6.2 INTRODUCTION

As comprehensively outlined by recent reviews, it has been well-established that self-rated health is an independent, strong predictor of morbidity and mortality. As such, self-reported health outcomes are increasingly being assessed in the evaluation of chronic disease states such as rheumatoid arthritis (RA). Socio-economic status (SES) has been shown to be strongly associated with self-reported health and therefore, more attention is being directed towards its determination and its role in the development and progression of disease.
Preference-based generic health related quality of life (HRQL) instruments are often used as self-reported measures of health and are increasingly used in economic evaluations as weighting factors for quality adjusted life years (QALYs). Similarly, disease-specific HRQL and functional status measures are often used as self-reported measures to fully assess the impact of chronic diseases and as monitoring tools in clinical practice. Despite the well-known association between self-reported health outcomes and SES, there has been little work evaluating the impact of SES on the results obtained using preference-based and disease specific HRQL measures in patients with RA.

In Canada, there is a universal health care system that is governed by the principles outlined in the Canada Health Act. However, despite the principles of public administration, universality, accessibility, portability, and comprehensiveness outlined in the Act, there remain large socio-economic inequalities in health within our system. In British Columbia, these disparities have been investigated in well-established, chronic diseases such as asthma, and HIV/AIDS. However, to our knowledge, the role of SES in self-reported health outcomes experienced by those with RA in a North American country with universal health care has not been investigated to date.

Therefore, the purpose of this study was to investigate the relationship between SES and self-reported health outcomes (both preference-based generic and disease-specific HRQL and functional status) in a sample of RA patients. Our hypothesis was that, despite adjustment for measures of disease severity, self-perceived generic and disease-specific HRQL and functional status would be worse in patients of lower SES.
6.3 METHODS

6.3.1 Study Sample and Design

Three hundred and thirteen English speaking adults between 19 and 90 years of age diagnosed with rheumatoid arthritis by a rheumatologist and who resided in the Greater Vancouver Regional District (GVRD) or the rural Okanagan region of British Columbia were recruited into this cross sectional study. One hundred and ninety seven (63%) patients were recruited directly by study rheumatologists whereas 116 were recruited via targeted mailouts to individuals with RA identified by their rheumatologists. The institutional and university ethics review boards approved the study protocol, and informed consent was obtained from each participant.

6.3.2 Generic Health-Related Quality of Life Measurement

The Short Form 6D (SF-6D), the Health Utilities Index Mark 2 and Mark 3 (HUI2 and HUI3, respectively), and the EuroQoL (EQ-5D) were used to measure generic HRQL, all of which have shown to have cross-sectional construct validity in patients with RA. Since these instruments measure different dimensions/attributes, all were included to assess a broader range of possible health outcomes.

Brazier et al. created the SF-6D to derive a preference-based measure of health from the Short Form-36. The SF-6D measures six dimensions, each with four to six levels and include physical functioning, role limitation, social functioning, pain, mental health, and vitality. A total of 18,000 health states can be defined by this classification system. Of these states, 249 were valued using the standard gamble (SG) in a sample of 611 UK participants.
from the general population. Modeling was used to generate the remainder of the values for the health states. Thus, in the final model, the boundaries of the SF-6D multi-attribute utility values were +0.30 and 1.00 (where zero is death and 1.00 is perfect health). The minimally important difference (MID) is thought to be about 0.03.\textsuperscript{10,13}

The HUI2 was created to measure HRQL in pediatric cancer patients and captures up to seven attributes of health: sensation (vision, hearing, speech), mobility, emotion cognition, self-care, pain, and fertility (optional).\textsuperscript{14,15} The HUI2 uses 4 or 5 levels on each attribute for a total of 24,000 possible unique health states. The boundaries on the overall multiattribute utility score are -0.03 to 1.00 and the MID is thought to be about 0.03 to 0.05. The HUI3 was created initially to measure HRQL in the National Population Health Survey.\textsuperscript{14} The 8 attributes of the HUI3 are: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain. Preference scores can be calculated for each attribute. The HUI3 system uses 5 or 6 levels for each attribute for a total of 972,000 possible unique health states.\textsuperscript{15} The boundaries on the overall multiattribute utility score on the HUI3 are -0.36 to 1.00 and a (MID) is thought to be about 0.06.\textsuperscript{10}

The EQ-5D assesses five domains of health: mobility, self-care, usual activity, pain/discomfort, and anxiety/depression.\textsuperscript{15} Each domain has 3 levels corresponding to a total of 243 possible unique health states. The boundaries on the EQ-5D are -0.59 to 1.00. The MID for the EQ-5D is thought to be from 0.03 to 0.05.\textsuperscript{10,15}

6.3.3 Functional Status Measurement

The Health Assessment Questionnaire (HAQ)\textsuperscript{16} was one of the first self-reported, functional status (disability) measures developed and has become one of the dominant instruments in musculoskeletal diseases including RA.\textsuperscript{16} The HAQ has been utilized to assess
functional status for approximately two decades and is a mandated outcome for clinical trials in RA. The HAQ is a measure of physical disability that assesses a respondent's ability to complete everyday tasks in areas such as dressing and grooming, rising, eating, walking, personal hygiene, reach, grip and other activities (such as getting into and out of a car). Each of these areas is assigned a section score that is further adjusted to account for the use of any aids, devices or help from another person. These scores are then summed and averaged to give an overall score between 0.0 (best possible function) to 3.0 (worst function). A change in HAQ score of 0.25 is considered to represent the MID.\textsuperscript{17,18}

\subsection*{6.3.4 RA Specific Quality of Life Measure}

The Rheumatoid Arthritis Quality of Life (RAQoL) questionnaire\textsuperscript{19} is a newly developed instrument and is the first patient-completed instrument specifically designed for use with RA patients.\textsuperscript{19} It was derived directly from qualitative interviews with relevant patients and considers aspects of many areas of life that are detrimentally impacted by RA. The RAQoL is meant to be a comprehensive, disease-specific scale that will be more responsive to change than previous scales used in RA. The RAQoL consists of 30 questions with binary responses that assess such aspects of RA as moods and emotions, social life, hobbies, everyday tasks, personal and social relationships, and physical contact. The RAQoL is scored by assigning one point for each affirmative response and no points for negative responses. Thus, scores range from 0 (best RA-specific quality of life) to 30 (worst RA-specific quality of life). We have estimated that the MID for the RAQoL is approximately 1.7 to 2.0.\textsuperscript{10}
6.3.5 Clinical Measurements

In addition to demographic questions, participants were asked questions regarding their RA management and severity including DMARD and prednisone therapy over the past three months. These questions included swollen and tender joint count (using the mannequin-based 42 joint count methodology), 10 cm pain visual analogue scale (VAS), and five point Likert scales of self-perceived RA severity and control and duration of RA. The erythrocyte sedimentation rate (ESR) was obtained from the health record. Health utilization measurements over the previous year such as hospitalization, use of other professional services (physiotherapy, occupational therapy, home care, massage therapy, etc.), and the rental or purchase of physical aides (walker, wheelchair, cane, etc.) were also collected.

6.3.6 Socioeconomic Status

The association between SES and generic and disease-specific HRQL was tested at both the individual and population level of SES. Individual measures of SES were based on self reported annual household income and education. Annual household income was classified as less than $20,000, from $20,000 to $50,000, and greater than $50,000 as previously defined. Since number of people living in each household can have an impact on annual household income, this variable was included in all analyses. Education was classified based on the number of years of post-secondary education and highest level of education completed, ordinally categorized as less than high school, high school or trade diploma, or at least a university bachelor’s degree.

Using a Postal Code Conversion file, we determined the census tract where each participant’s current residence was located. From this we determined the neighbourhood
characteristics related to SES for each participant’s current residence that could be derived from the province’s (British Columbia) Census data. Census tract level variables deemed representative of SES included median neighbourhood income, the proportion of the population over 20 years of age completing at least a bachelor’s education, and the neighbourhood unemployment rate.

6.3.7 Statistical Analysis

Spearman’s rho was used to examine the correlations between the different measures of SES (a Spearman’s rho of $> 0.50$ or $<-0.50$ were considered be strong, while values between -0.49 to -0.30 or 0.30 to 0.49 were considered moderate and values between -0.30 and 0.30 were considered to be weak). The dependent variables for all other analyses were the global utility scores for the SF-6D, the HUI2, the HUI3, the EQ-5D and the overall RAQoL and HAQ disability score. Univariate associations between the dependent variables and demographic characteristics (age, and gender), disease severity measures (duration of RA, self-reported pain, swollen/tender joint count, self-reported severity and control), and measures of SES were assessed using simple linear regression. Comparisons of self-reported health and RA severity measures across categories of SES were conducted. Statistical comparisons were made using either ANOVA, Student’s t-test or $\chi^2$ test, where appropriate.

For the primary analysis, ordinary least squares (OLS) regression was used to adjust for RA clinical measures and then to assess the relationship between SES and the dependent variables. Each SES variable was modeled separately. All two-way interactions between SES and RA clinical measures were also tested in the multiple regression models. Model fit was
assessed using adjusted $R^2$ and standardized residuals were plotted against standardized predicted scores to assess each model for homoscedasticity.

To account for the possibility that there is an inter-relationship between the dependent variables (the generic or disease-specific HRQL or functional status score) and annual household income, two-stage least squares (TSLS) regression was used. For TSLS, the "problematic" predictor variable (in our case, income) must be continuous rather than categorical. Therefore, for the TSLS analysis, the self-reported annual income variable was converted to a measure in increments of $10,000.00 (as reported on the original questionnaire). Instrumental variables are not influenced by others in the model but have influence on the variable of interest. Thus, for the first stage of the regression, the instrumental variables used in our analysis to predict annual household income were marital status, number of people in the household, and educational status which were all highly correlated with income but lowly correlated with HRQL or functional status measures. In the second stage, the predicted values of income were regressed on the generic HRQL scores (HUI2, HUI3, SF-6D, or EQ-5D), the disease-specific HRQL scores (RAQoL) and the HAQ scores to yield unbiased parameter estimates. These parameters were compared to OLS regression coefficients (from models using the same annual household income variable) to determine how closely they matched.

6.4 RESULTS

Characteristics of the study participants are presented in Table 6.1. The average age of the sample was $61.5 \pm 25.9$ years, there were more women (79%) and males tended to be
older (66 vs. 60 years, p=0.004). The mean number of years since being diagnosed with RA was 13.9 ± 11.4. Fifty eight (19%) of the 313 participants did not report their annual household income and were therefore excluded from the income analysis at the individual level. To determine if this subgroup was different from the group that reported annual income levels, we compared the values for the variables presented in Table 6.1 between these two groups. There were no differences between these variables in the two subgroups adding confidence that no bias was introduced through these missing data.

The sample was well distributed across levels of SES (Table 6.2). Although 50 (16%) had an annual household income below $20,000, 47 (15%) reported incomes over $70,000, 15 of which exceeded $100,000. There was a significant relationship between self-reported annual income and number of household members with 56% of those reporting income of less <20K per year being the only household member compared with 17% and 9% of those reporting annual household incomes of 20 – 50K and >50K per year, respectively (p<0.0001). The mean number of household members was also significantly lower for those with an annual household income reported as <20K (mean 1.7, standard deviation 1.1), compared with those reporting annual household incomes of >50K per year (mean 2.6, standard deviation 1.4, p<0.0001).

With respect to self-reported education, 114 (51%) completed at least one year of post secondary education (median 1.0) and 52 (17%) received at least a bachelors degree; 32 (10%) completed at least five years of post-secondary education. The sample was also heterogeneous for the contextual measures (neighbourhood median income, prevalence of having received at least a bachelor’s degree, and percent neighbourhood unemployment), with participants residing in neighbourhoods with unemployment rates varying from 1% to
18% (median = 9%) and the prevalence of a bachelor's degree ranging from 2% to 52% (median=12%).

The self-reported measures of SES (annual income and education) were moderately correlated (Spearman's rho 0.33, p<0.0001). However, the contextual measures (tended to be more highly correlated amongst themselves with correlation coefficients ranging from 0.56 to 0.73 (all p<0.0001). Correlations between self-reported and contextual measures were mostly low with coefficients ranging from 0.11 to 0.31 (all p<0.05).

Unadjusted associations between the generic HRQL measures (the SF-6D, the HUI3, the HUI2, and the EQ-5D), or the RAQoL and the HAQ, and demographic, SES and RA severity variables are presented in Tables 6.3, 6.4 and 6.5, respectively. There were no associations between age and any of the generic health related quality of life measures; however, there was a significant positive association with the HAQ disability index (p=0.02). On average, men had significantly better generic and disease-specific quality of life and functional status as measured by most of the instruments. Most of the RA severity variables were significant across all of the HRQL instruments and the HAQ. No associations were found between contextual measures of SES (neighbourhood median income, prevalence of bachelor's degrees, and proportion of neighbourhood unemployed) and any of the generic or disease-specific HRQL measures or the HAQ (Tables 6.3, 6.4 and 6.5) with the exception of median neighbourhood income and the EQ-5D (p=0.02).

For both self-reported income and education levels, comparison of mean values of all measures (SF-6D, HUI3, HUI2, EQ-5D, RAQoL, HAQ) by ANOVA showed a significant gradient across SES categories (Table 6.6). For example, lower levels of income were associated with poorer generic and disease-specific HRQL and physical function. Results were confirmed with nonparametric tests. In general, all measures of health utilization
hospitalization, use of professional services, and use of physical aides/equipment), joint
damage, and health status showed a consistent gradient of worse functioning in those in
lower self-reported SES categories (self-reported annual family income and self-reported
education level) (Table 6.6). There were no differences across SES categories for type or
number of DMARDs or the use of prednisone over the past three months. There were no
gradients or associations between any of these variables and the contextual measures of SES.

For the proximate (self-reported) SES measures, the differences in these variables
were statistically significant for most of the subjective RA severity measures (self-reported
RA severity, patient global assessment of disease activity) and both the global scores of the
generic and disease-specific quality of life measures. Of note, most of the physically-based
clinical measures (such as joint counts) were not significantly different between the different
SES levels suggesting no physical differences in disease severity.

The results of the OLS regressions with adjustment for disease severity measures
show significant associations between the HUI3 and the SF-6D overall scores and the HAQ
for self-reported income (Figures 6.1, 6.2 and 6.3). There were no other significant
associations for other measures of SES (self-reported education of contextual measures) after
adjustment for disease severity. To account for differences in household size across self-
reported annual household income categories, we included number of people in the
household in the regression models. Other measures of disease management and severity that
were tested but did not improve the overall fit of the model included number and type of
DMARDs used within past three months, number of other chronic diseases, swollen joint
count (collinear with tender joint count) and erythrocyte sedimentation rate. Of note, the
RAQoL did not significantly differ across self-reported income categories. All differences in
the $\beta$-estimates between the lowest and highest income categories exceeded the MCID for the SF-6D, HUI3, and the HAQ. The results of the TSLS regression analyses also reveal that self-reported annual income was significantly associated with most of the generic HRQL measures ($p=0.03$, $p=0.03$ and 0.04 for the HUI2, HUI3 and SF-6D, respectively) and the HAQ ($p=0.002$) but not for the RAQoL ($p=0.07$), or the EQ-5D ($p=0.14$). A comparison of the beta-coefficients between OLS and TSLS regression revealed close agreement for the HUI2 (0.02 vs. 0.04), HUI3 (0.03 for both), the SF-6D (0.01 for both), the EQ-5D (0.03 and 0.04), the HAQ (-0.09 vs. -0.08) and the RAQoL (-0.50 vs. -0.70). Thus, there likely was little, if any, feedback between income and HRQL or functional status in the sample adding further credence to the utilization of OLS.

6.5 DISCUSSION

This study demonstrates a consistent and significant gradient in both generic and disease-specific HRQL and functional status and other self-reported health measures across income and education categories. This gradient is maintained even after adjustment for RA severity and number of people in the household for self-reported income (but not education) categories. Of note, there was no significant gradient across SES measures for physically defined RA severity measures (disease process and joint damage measures). Thus, these results highlight how SES impacts a well-defined chronic disease such as RA by influencing how patients perceive and report their health status.

These findings become particularly important when one considers that self-rated health predicts mortality even after controlling for a wide range of factors (demographic, psychosocial, prior illness, physician's assessments and physiological measures). Thus,
from our results, we have determined that low SES predicts poor self-reported health independently of RA severity and may thus be a strong contributing factor to the early mortality and substantial morbidity seen in RA patients with low SES.\textsuperscript{24,25}

Another important finding is that the magnitudes of utility values assessed by both the HUI3 and the SF-6D significantly vary by SES independently of RA severity measures. In addition, although not statistically significant for the HUI2 and the EQ-5D, there was a pattern for higher scores in higher income groups. This finding has potentially important ramifications for results of cost-utility analyses in therapies for RA as investigators need to ensure balance between treatment groups in not only clinical and disease specific factors but also SES. Therefore, in order to avoid potential confounding or bias due to SES status in economic evaluation, one would need to: 1) verify SES at baseline in randomized controlled trials (RCT) where utility measures will be used in a cost-utility analysis; 2) control for SES in observational studies where such measures may be used; and 3) ensure that the results obtained from the sample will be generalizable to the population of interest (i.e. to the extent that studies do not report SES and/or the SES is not similar to the general population).

Sculpher and O'Brien outline some additional concerns with using the results of indirectly assessed utility scores that are influenced by income in cost-utility analysis.\textsuperscript{26} They state that often, in cost-benefit analysis, where willingness to pay is often utilized to assign monetary value to benefits, ability to pay biases such data in favour of the more affluent. QALYs, as used as outcomes in cost-utility analyses, are thought to avoid this potential bias. However, the authors argue (and our results suggest) that this may not be the case. Specifically, the authors state that the effects of income could come into play when individuals are asked to value health states to generate utilities for the indirect utility instrument scoring functions or when the instrument is applied in the field. In our study, the
latter scenario is applicable. In this situation, the authors state that there is no reason why income effects should be excluded as these could be a relevant component of illness that may contribute to deficits in health status. However, these income effects could bias cost-utility analysis in at least two different ways: 1) when cross-national comparisons are being made and there are differences between countries in the levels of income maintenance available to the sick; and 2) the possibility of double-counting if income effects of reduced health have already been factored into the valuation stage of the instrument. For example, reduced quality of life that is mediated by loss of income should be counted in the denominator of the cost-effectiveness ratio. However, if one also includes loss of income as an indirect cost in the numerator, than there is a potential to count these effects twice – in the numerator and the denominator.

In RA, while it is well established that there are associations between low SES and morbidity and mortality, the mechanisms behind these associations are largely unknown. Callahan et al.\textsuperscript{27} reported that scores on a helplessness scale appeared to mediate a component of the association between formal education level and five year mortality. In a study attempting to identify a partial explanation for the association between low education and poor outcome in RA, Katz\textsuperscript{28} identified that self-care was strongly associated with education and thus concluded that low education was a proxy for a constellation of factors responsible for poor health outcomes. Therefore, the differences in self-reported health that we observed on both the generic and disease-specific HRQL and the HAQ scales might be indicative of helplessness or inability to complete self-care tasks in patients with lower SES.

Our results generally support the findings by Brekke et al.\textsuperscript{29} and McEntegart et al.\textsuperscript{30} who showed that self-reported health outcomes, but not objective indices of disease activity, differed across groups based upon SES. Specifically, McEntegart et al.\textsuperscript{30}, revealed how
patients living in more deprived areas in Scotland had poorer HAQ scores as compared to those living in more affluent areas. Similarly, Brekke et al., who conducted their study as a comparison of RA patients from affluent west Oslo to those from deprived east Oslo, extended these findings to disease-specific and generic quality of life measures. Both of these analyses used contextual measures of SES. The study by McEntegart et al. utilized the Carstairs index (a composite score using postal code that draws on measures of overcrowding, male unemployment, social class and car ownership) while Brekke et al. utilized neighbourhood factors (such as income, education, employment, mortality, housing standard and proportion of third world citizens) to define the two areas of Oslo as affluent or deprived.

Our findings build on those previously reported by including multiple measures of SES including those directly reported by the patient as opposed to only performing neighbourhood level analyses and the addition of two preference-based, generic HRQL instruments and the RAQoL. Since we collected patient-specific RA drug treatment data, we were able to determine that there were no treatment differences across SES categories that could have influenced self-reported outcomes. Similarly, since all of our subjects were under the care of rheumatologists, any differences in specialist versus non-specialist care that may have been due to SES and could potentially have influenced self-reported outcomes were avoided.

In addition, our study is the first to examine if this relationship holds true in a North American country with universal access to health care. Of note, we adjusted our model by the number of people living in the household. While we found that there was a significant difference in number of people per household across self-reported annual income with higher levels of income reported by those with larger families, this variable was not significantly
associated with the self-reported health variables, did not effect the magnitude or significance of the association of annual household income with the dependent variables, and did not significantly improve the multiple linear regression model fit.

Another point of interest that arises from the results of our study was the lack of a consistent gradient (except for the HUI3) across the income categories for the adjusted models of self-reported generic and disease-specific HRQL and functional status (Figures 6.1 and 6.2). For example, with the SF-6D, it appears that the biggest difference across income categories is between the middle and highest groups rather than between the lowest and highest groups. These results bring up the possibility that there may not be a perfect gradient across the three separate income categories and that it may be a dichotomous phenomenon (i.e. high versus low income) with an annual household income cut-off of approximately $50,000 defining the two groups. Another possible explanation is that there is another factor that is somehow influencing the self-reported health outcomes for the middle income category making it lower than both the high and low income categories.

Of interest, in our study, there was a low correlation between the proximate (self-reported) and contextual measures of SES. With both annual household income and education, there were strong univariate associations with self-reported health. However, once adjustments for RA severity were made, only self-reported annual income remained significant. A possible explanation for the lack of association between the self-reported HRQL measures and education and contextual SES measures is that they may not be indicative of SES in elderly populations such as those with RA. Our sample was mostly comprised of subjects who had worked in an era when there was less emphasis on education. Therefore, in contrast to a younger, employed sample of asthmatics from the same geographical area where education and income was highly correlated, results from our
sample revealed that these two variables were less correlated. Similarly, in the
aforementioned asthma sample, there were strong correlations between contextual and
proximate measures of SES that were not observed in our RA sample indicating that these
measures of SES may be more robust for younger participants who are more likely to be
currently employed. Another finding from our sample that supports this premise is that older
individuals (>50 years of age) who were still in the work force tended to have less education
but similar income to those working individuals less than 50 years old.

While there were significant gradients across SES (as defined by annual household
income) for both of the generic HRQL measures and the HAQ after adjustment for RA
severity, similar findings were not observed for the RAQoL. Despite significant univariate
gradients across SES as defined by annual household income and education, the RAQoL did
not display a clear SES gradient in the multiple linear regression analysis. We postulate that
the reason for this is that the RAQoL is capturing items that are so germane to RA that the
variance in its score is explained mostly by the objective and subjective disease severity
measures. Indeed, the addition of annual household income had a negligible impact on the
model R² in the multiple linear regression analysis of the RAQoL whereas it improved the
model fit in all the other analyses.

Finally, it can be argued that the results using OLS regression only reveal an
association between self-reported annual household income and HRQL or functional status
without the ascertainment of directionality (i.e. is it the low income that is causing the low
HRQL/functional status or vice versa?). We utilized TSLS regression to account for this and
found no evidence to support that the low income was “caused” by the low HRQL or
functional status (i.e. the beta coefficients achieved by OLS were not biased). This finding
likely makes sense in our sample since most participants were elderly and retired and their
current annual household income was likely not influenced by their current HRQL or functional status.

Our study shows that even in a country such as Canada with universal access to health care, the impact of RA on self-reported health is strongly associated with SES as measured by annual income even after adjusting for disease severity. Because self-reported health has been strongly associated with mortality and morbidity, there are important implications for intervention. In addition, these findings should be considered in the context of cost-utility analysis to prevent biasing of utility values obtained from preference-based instruments. In the event that studies do not investigate or report SES or if the SES in the study sample differs significantly from the population of interest, the results of the analysis may have poor generalizability. Further research should focus on the mediating factors that contribute to this social gradient in self-reported health outcomes in RA.
6.6 REFERENCES


13. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health Qual Life Outcomes. 2003;11:4-12.


<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (yrs)</td>
<td>61.5</td>
<td>25.9</td>
</tr>
<tr>
<td>RA Duration (yrs)</td>
<td>13.87</td>
<td>11.41</td>
</tr>
<tr>
<td>Pain VAS (mm) 0 to 100</td>
<td>43.12</td>
<td>27</td>
</tr>
<tr>
<td>Tender Joint Count 0 to 50</td>
<td>15.09</td>
<td>12</td>
</tr>
<tr>
<td>Swollen Joint Count 0 to 50</td>
<td>9.14</td>
<td>9.67</td>
</tr>
<tr>
<td>Erythrocyte Sedimentation Rate (mm/hr)</td>
<td>24.71</td>
<td>21.01</td>
</tr>
<tr>
<td>HAQ Disability Index 0 to 3.0</td>
<td>1.1</td>
<td>0.77</td>
</tr>
<tr>
<td>RAQoL Score 0 to 28</td>
<td>12.73</td>
<td>8.48</td>
</tr>
<tr>
<td>HUI-2 Global Utility Score -0.03 to 1.00</td>
<td>0.71</td>
<td>0.19</td>
</tr>
<tr>
<td>HUI-3 Global Utility Score -0.36 to 1.00</td>
<td>0.53</td>
<td>0.29</td>
</tr>
<tr>
<td>EQ-5D Global Utility Score -0.59 to 1.00</td>
<td>0.67</td>
<td>0.24</td>
</tr>
<tr>
<td>SF-6D Global Utility Score 0.31 - 1.00</td>
<td>0.63</td>
<td>0.13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-Reported RA Severity, n %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Mild</td>
<td>9</td>
<td>3%</td>
</tr>
<tr>
<td>Mild</td>
<td>34</td>
<td>11%</td>
</tr>
<tr>
<td>Moderate</td>
<td>120</td>
<td>38%</td>
</tr>
<tr>
<td>Severe</td>
<td>110</td>
<td>35%</td>
</tr>
<tr>
<td>Very Severe</td>
<td>27</td>
<td>9%</td>
</tr>
<tr>
<td>Self-Reported RA Control, n %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very Well Controlled</td>
<td>33</td>
<td>11%</td>
</tr>
<tr>
<td>Well Controlled</td>
<td>76</td>
<td>24%</td>
</tr>
<tr>
<td>Adequately Controlled</td>
<td>123</td>
<td>39%</td>
</tr>
<tr>
<td>Not Well Controlled</td>
<td>61</td>
<td>19%</td>
</tr>
<tr>
<td>Not Controlled At All</td>
<td>7</td>
<td>2%</td>
</tr>
<tr>
<td>Hospitalized For RA in Last 12 Months, n%</td>
<td>45</td>
<td>15%</td>
</tr>
<tr>
<td>Missed Work or School Due to RA in Last 12 Months, n%</td>
<td>59</td>
<td>19%</td>
</tr>
<tr>
<td>Purchased or Rented Equipment for RA in Last 12 Months, n%</td>
<td>72</td>
<td>23%</td>
</tr>
<tr>
<td>Used Allied Health Professional/Home Care Services in Last 12 months, n %</td>
<td>129</td>
<td>42%</td>
</tr>
<tr>
<td>Concomitant Chronic Illness Other Than RA, n%</td>
<td>192</td>
<td>62%</td>
</tr>
</tbody>
</table>
### TABLE 6.2. PROPERTIES OF THE MEASURES OF SOCIOECONOMIC STATUS (SES) IN OUR SAMPLE

<table>
<thead>
<tr>
<th>Contextual SES Measures</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighbourhood Median Income</td>
<td>20040</td>
<td>4313</td>
</tr>
<tr>
<td>Bachelor’s Education (%)</td>
<td>15.2</td>
<td>10.2</td>
</tr>
<tr>
<td>Neighbourhood Unemployment Rate (%)</td>
<td>8.7</td>
<td>2.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Self-Reported SES Measures (Proximate)</th>
<th>Number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education Completed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than High School/Trade</td>
<td>75</td>
<td>24%</td>
</tr>
<tr>
<td>High School/Trade</td>
<td>169</td>
<td>54%</td>
</tr>
<tr>
<td>Bachelor’s</td>
<td>52</td>
<td>17%</td>
</tr>
<tr>
<td>Missing</td>
<td>17</td>
<td>5%</td>
</tr>
<tr>
<td>Annual Household Income</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$20,000</td>
<td>50</td>
<td>16%</td>
</tr>
<tr>
<td>$20,000 to $50,000</td>
<td>115</td>
<td>37%</td>
</tr>
<tr>
<td>&gt;$50,000</td>
<td>90</td>
<td>28%</td>
</tr>
<tr>
<td>Missing</td>
<td>58</td>
<td>19%</td>
</tr>
</tbody>
</table>
TABLE 6.3 UNIVARIATE ASSOCIATIONS WITH THE GENERIC HRQL MEASURES (THE SF-6D AND THE HUI3)

<table>
<thead>
<tr>
<th>Factor</th>
<th>SF-6D Global Utility Score</th>
<th>HUI3 Global Utility Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Regression Coefficient (SE)</td>
<td>p-value</td>
</tr>
<tr>
<td>DEMOGRAPHICS</td>
<td>Age</td>
<td>-0.0002 (0.0006)</td>
</tr>
<tr>
<td></td>
<td>Gender (Female is reference)</td>
<td>0.05 (0.02)</td>
</tr>
<tr>
<td>RA SEVERITY VARIABLES</td>
<td>Years since diagnosis</td>
<td>-0.002 (0.0006)</td>
</tr>
<tr>
<td></td>
<td>No. of other chronic diseases</td>
<td>-0.02 (0.005)</td>
</tr>
<tr>
<td></td>
<td>Erythrocyte sedimentation rate</td>
<td>-0.002 (0.0006)</td>
</tr>
<tr>
<td></td>
<td>Tender joint count</td>
<td>-0.006 (0.0005)</td>
</tr>
<tr>
<td></td>
<td>Swollen joint count</td>
<td>-0.006 (0.0007)</td>
</tr>
<tr>
<td></td>
<td>Global pain VAS</td>
<td>-0.003 (0.0002)</td>
</tr>
<tr>
<td></td>
<td>Patient global assessment VAS</td>
<td>0.003 (0.0002)</td>
</tr>
<tr>
<td></td>
<td>HAQ disability index score</td>
<td>-0.12 (0.08)</td>
</tr>
<tr>
<td></td>
<td>Hospitalization in last year*</td>
<td>-0.03 (0.02)</td>
</tr>
<tr>
<td></td>
<td>Home/Health services for RA*</td>
<td>-0.05 (0.02)</td>
</tr>
<tr>
<td></td>
<td>Purchase/rent RA equipment*</td>
<td>-0.05 (0.02)</td>
</tr>
<tr>
<td></td>
<td>Missed work/school in last year*</td>
<td>-0.07 (0.02)</td>
</tr>
<tr>
<td></td>
<td>RA self-reported severity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Very mild</td>
<td>0.33 (0.04)</td>
</tr>
<tr>
<td></td>
<td>Mild</td>
<td>0.24 (0.03)</td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>0.15 (0.02)</td>
</tr>
<tr>
<td></td>
<td>Severe</td>
<td>0.08 (0.02)</td>
</tr>
<tr>
<td></td>
<td>Very severe</td>
<td>Ref</td>
</tr>
<tr>
<td></td>
<td>RA self-reported control</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Very well controlled</td>
<td>0.28 (0.05)</td>
</tr>
<tr>
<td></td>
<td>Well controlled</td>
<td>0.24 (0.04)</td>
</tr>
<tr>
<td></td>
<td>Adequately controlled</td>
<td>0.17 (0.04)</td>
</tr>
<tr>
<td></td>
<td>Not well controlled</td>
<td>0.07 (0.04)</td>
</tr>
<tr>
<td></td>
<td>Not controlled at all</td>
<td>Ref</td>
</tr>
<tr>
<td>PROXIMATE SES FACTORS</td>
<td>Education Completed</td>
<td>-0.06 (0.02)</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>-0.04 (0.02)</td>
</tr>
<tr>
<td></td>
<td>High school/Trade</td>
<td>Ref</td>
</tr>
<tr>
<td></td>
<td>Bachelors education</td>
<td>0.005 (0.003)</td>
</tr>
<tr>
<td></td>
<td>Yrs. Post-secondary education</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Annual household income</td>
<td></td>
</tr>
<tr>
<td></td>
<td>≤ $20,000</td>
<td>-0.09 (0.02)</td>
</tr>
<tr>
<td></td>
<td>&gt; $50,000</td>
<td>-0.06 (0.02)</td>
</tr>
<tr>
<td>CONTEXTUAL SES FACTORS</td>
<td>Median Neighborhood Income†</td>
<td>0.017 (0.018)</td>
</tr>
<tr>
<td></td>
<td>% Bachelors Education</td>
<td>0.0002 (0.0008)</td>
</tr>
<tr>
<td></td>
<td>Neighbourhood Unemployment</td>
<td>-0.0005 (0.002)</td>
</tr>
</tbody>
</table>

* Reference category is "no"
† For categories of $10,000
TABLE 6.4 UNIVARIATE ASSOCIATIONS WITH THE GENERIC HRQL MEASURES (THE HUI2 AND THE EQ-5D)

<table>
<thead>
<tr>
<th>Factor</th>
<th>HUI2 Global Utility Score</th>
<th></th>
<th></th>
<th>EQ-5D Global Utility Score</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Regression Coefficient (SE)</td>
<td>p-value</td>
<td>Regression Coefficient (SE)</td>
<td>p-value</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DEMOGRAPHICS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.01 (0.0009)</td>
<td>NS</td>
<td>0.008 (0.001)</td>
<td>NS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender (Female is reference)</td>
<td>0.04 (0.03)</td>
<td>NS</td>
<td>0.05 (0.03)</td>
<td>NS</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RA SEVERITY VARIABLES</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years since diagnosis</td>
<td>0.002 (0.001)</td>
<td>0.04</td>
<td>-0.003 (0.001)</td>
<td>0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of other chronic diseases</td>
<td>-0.03 (0.008)</td>
<td>0.0008</td>
<td>-0.02 (0.01)</td>
<td>0.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tender joint count</td>
<td>-0.007 (0.0008)</td>
<td>&lt;0.0001</td>
<td>-0.008 (0.001)</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Erythrocyte sedimentation rate</td>
<td>-0.002 (0.0007)</td>
<td>0.01</td>
<td>-0.002 (0.0008)</td>
<td>0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swollen joint count</td>
<td>-0.008 (0.001)</td>
<td>&lt;0.0001</td>
<td>-0.01 (0.001)</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global pain VAS</td>
<td>-0.004 (0.0003)</td>
<td>&lt;0.0001</td>
<td>-0.005 (0.0004)</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient global assessment VAS</td>
<td>0.004 (0.0003)</td>
<td>&lt;0.0001</td>
<td>0.005 (0.0004)</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospitalization in last year*</td>
<td>-0.077 (0.032)</td>
<td>0.02</td>
<td>-0.05 (0.04)</td>
<td>0.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Home/Health services for RA*</td>
<td>-0.09 (0.02)</td>
<td>&lt;0.0001</td>
<td>-0.11 (0.03)</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purchase/rent RA equipment*</td>
<td>-0.12 (0.03)</td>
<td>&lt;0.0001</td>
<td>-0.12 (0.03)</td>
<td>0.0003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missed work/school in last year*</td>
<td>0.08 (0.03)</td>
<td>0.009</td>
<td>0.13 (0.04)</td>
<td>0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RA self-reported severity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very mild</td>
<td>0.31 (0.03)</td>
<td>&lt;0.0001</td>
<td>0.41 (0.08)</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mild</td>
<td>0.27 (0.07)</td>
<td>&lt;0.0001</td>
<td>0.36 (0.05)</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>0.18 (0.04)</td>
<td>&lt;0.0001</td>
<td>0.24 (0.04)</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severe</td>
<td>0.05 (0.04)</td>
<td>NS</td>
<td>0.12 (0.04)</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very severe</td>
<td>Ref</td>
<td></td>
<td>Ref</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RA self-reported control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very well controlled</td>
<td>0.40 (0.07)</td>
<td>&lt;0.0001</td>
<td>0.56 (0.08)</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Well controlled</td>
<td>0.37 (0.07)</td>
<td>&lt;0.0001</td>
<td>0.52 (0.08)</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adequately controlled</td>
<td>0.31 (0.06)</td>
<td>&lt;0.0001</td>
<td>0.42 (0.08)</td>
<td>&lt;0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not well controlled</td>
<td>0.13 (0.07)</td>
<td>0.05</td>
<td>0.22 (0.08)</td>
<td>0.006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not controlled at all</td>
<td>Ref</td>
<td></td>
<td>Ref</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PROXIMATE SES FACTORS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education completed</td>
<td></td>
<td>NS</td>
<td></td>
<td></td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>-0.07 (0.04)</td>
<td>0.04</td>
<td>-0.11 (0.04)</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school/trade</td>
<td>-0.05 (0.03)</td>
<td>NS</td>
<td>-0.02 (0.04)</td>
<td>NS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bachelors education</td>
<td>Ref</td>
<td></td>
<td>Ref</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yrs. Post-secondary education</td>
<td>0.006 (0.005)</td>
<td>NS</td>
<td>0.002 (0.006)</td>
<td>NS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual household income</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>≤ $20,000</td>
<td>-0.11 (0.03)</td>
<td>0.006</td>
<td>-0.15 (0.04)</td>
<td>0.0004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; $50,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CONTEXTUAL SES FACTORS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median neighborhood income†</td>
<td>0.000002 (0.000003)</td>
<td>NS</td>
<td>0.000008</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% bachelors education</td>
<td>0.0001 (0.001)</td>
<td>NS</td>
<td>0.001 (0.001)</td>
<td>NS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neighbourhood unemployment</td>
<td>-0.003 (0.004)</td>
<td>NS</td>
<td>-0.008 (0.005)</td>
<td>0.08</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Reference category is "no"
† For categories of $10,000
TABLE 6.5 UNIVARIATE ANALYSIS WITH THE DISEASE-SPECIFIC MEASURES
(THE HAQ AND THE RAQOL)

<table>
<thead>
<tr>
<th>Factor</th>
<th>HAQ Score</th>
<th>RAQoL Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Regression Coefficient (SE)</td>
<td>p-value</td>
</tr>
<tr>
<td>DEMOGRAPHICS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.007 (0.004)</td>
<td>0.03</td>
</tr>
<tr>
<td>Gender (Female is reference)</td>
<td>-0.35 (0.11)</td>
<td>0.001</td>
</tr>
<tr>
<td>RA SEVERITY VARIABLES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years since diagnosis</td>
<td>0.02 (0.003)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>No. of other chronic diseases</td>
<td>0.10 (0.03)</td>
<td>0.002</td>
</tr>
<tr>
<td>Tender joint count</td>
<td>0.03 (0.003)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Erythrocyte sedimentation rate</td>
<td>0.008 (0.003)</td>
<td>0.007</td>
</tr>
<tr>
<td>Swollen joint count</td>
<td>0.03 (0.004)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Global pain VAS</td>
<td>0.02 (0.001)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Patient global assessment VAS</td>
<td>-0.01 (0.001)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Hospitalization in last year*</td>
<td>0.33 (0.13)</td>
<td>0.008</td>
</tr>
<tr>
<td>Home/Health services for RA*</td>
<td>0.43 (0.09)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Purchase/rent RA equipment*</td>
<td>0.48 (0.10)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Missed work/school in last year*</td>
<td>0.42 (0.12)</td>
<td>0.0005</td>
</tr>
<tr>
<td>RA SEVERITY VARIABLES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very mild</td>
<td>-1.40 (0.26)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Mild</td>
<td>-1.26 (0.17)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Moderate</td>
<td>-0.68 (0.14)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Severe</td>
<td>-0.31 (0.15)</td>
<td>0.03</td>
</tr>
<tr>
<td>Very severe</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>RA self-reported control</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very well controlled</td>
<td>-1.39 (0.28)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Well controlled</td>
<td>-1.25 (0.27)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Adequately controlled</td>
<td>-0.94 (0.26)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Not well controlled</td>
<td>-0.40 (0.27)</td>
<td>0.13</td>
</tr>
<tr>
<td>Not controlled at all</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>PROXIMATE SES FACTORS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education completed</td>
<td>0.05</td>
<td>NS</td>
</tr>
<tr>
<td>None</td>
<td>0.30 (0.13)</td>
<td>0.02</td>
</tr>
<tr>
<td>High school/trade</td>
<td>0.26 (0.12)</td>
<td>0.03</td>
</tr>
<tr>
<td>Bachelors education</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>Yrs. Post-secondary education</td>
<td>-0.03 (0.02)</td>
<td>NS</td>
</tr>
<tr>
<td>Annual household income</td>
<td>&lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>≤ $20,000</td>
<td>0.61 (0.13)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>$20,000 - $50,000</td>
<td>0.50 (0.10)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>$50,000</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>CONTEXTUAL SES FACTORS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median neighborhood income†</td>
<td>0.1 (0.1)</td>
<td>NS</td>
</tr>
<tr>
<td>% bachelors education</td>
<td>0.0004 (0.002)</td>
<td>NS</td>
</tr>
<tr>
<td>Neighbourhood unemployment</td>
<td>-0.0005 (0.002)</td>
<td>NS</td>
</tr>
</tbody>
</table>

* Reference category is "no"   
† For categories of $10,000
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Self-Reported Annual Family Income</th>
<th>Self-Reported Education Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt; 20K</td>
<td>20 - 50K</td>
</tr>
<tr>
<td></td>
<td>N=49</td>
<td>n=113</td>
</tr>
<tr>
<td>RA Duration (years)</td>
<td>16.7 (13.7)</td>
<td>13.1 (9.7)</td>
</tr>
<tr>
<td>Number of people per household</td>
<td>1.7 (1.1)</td>
<td>2.1 (0.8)</td>
</tr>
<tr>
<td>Erythrocyte sedimentation rate</td>
<td>26.5 (24.2)</td>
<td>25.9 (18.6)</td>
</tr>
<tr>
<td>Number of swollen joints (0-54)</td>
<td>9.6 (10.7)</td>
<td>9.2 (9.3)</td>
</tr>
<tr>
<td>Number of tender joints (0-54)</td>
<td>15.4 (13.3)</td>
<td>15.5 (11.1)</td>
</tr>
<tr>
<td>Self-reported RA severity (1-5)</td>
<td>3.8 (0.8)</td>
<td>3.4 (0.9)</td>
</tr>
<tr>
<td>Self-reported RA control (1-5)</td>
<td>2.9 (1.1)</td>
<td>2.7 (1.0)</td>
</tr>
<tr>
<td>Disease activity - patients' global assessment (0-100)</td>
<td>53.6 (26.4)</td>
<td>58.8 (25.2)</td>
</tr>
<tr>
<td>Hospitalization in last year (%)*</td>
<td>2.5</td>
<td>6.2</td>
</tr>
<tr>
<td>Home/Health services for RA (%)*</td>
<td>8.9</td>
<td>17.5</td>
</tr>
<tr>
<td>Purchase/rent RA equipment (%)*</td>
<td>6.4</td>
<td>15.2</td>
</tr>
<tr>
<td>Missed work/school in last year (%)*</td>
<td>4.0</td>
<td>7.2</td>
</tr>
<tr>
<td>Pain VAS (0-100)</td>
<td>46.5 (29.0)</td>
<td>45.1 (26.7)</td>
</tr>
<tr>
<td>RAQoL Score (0 – 30)</td>
<td>14.5 (8.4)</td>
<td>13.2 (8.2)</td>
</tr>
<tr>
<td>HAQ Disability Index (0.0 - 3.0)</td>
<td>1.4 (0.7)</td>
<td>1.3 (0.8)</td>
</tr>
<tr>
<td>SF-6D Global Score (0.30 - 1.00)</td>
<td>0.58 (0.13)</td>
<td>0.61 (0.13)</td>
</tr>
<tr>
<td>HUI3 Global Score (-0.36 -1.00)</td>
<td>0.41 (0.31)</td>
<td>0.52 (0.28)</td>
</tr>
<tr>
<td>HUI2 Global Score (-0.03 -1.00)</td>
<td>0.64 (0.21)</td>
<td>0.69 (0.18)</td>
</tr>
<tr>
<td>EQ-5D Global Score (-0.59 – 1.00)</td>
<td>0.58 (0.28)</td>
<td>0.65 (0.23)</td>
</tr>
</tbody>
</table>

214
The points indicate the least squares means and the lines indicate their 95% confidence intervals. HRQL measures were adjusted for RA duration, pain VAS, self-reported RA control and severity, tender joint count, RAQoL score, and number of people in the household. Higher scores indicate better HRQL.
FIGURE 6.2: GENERIC HRQL BY SELF-REPORTED ANNUAL INCOME (HUI2 AND EQ-5D)

The points indicate the least squares means and the lines indicate their 95% confidence intervals. HRQL measures were adjusted for RA duration, pain VAS, self-reported RA control and severity, tender joint count, RAQoL score, and number of people in the household. Higher scores indicate better HRQL.
FIGURE 6.3: RAQoL SCORE AND HAQ DISABILITY INDEX BY SELF-REPORTED INCOME

The points indicate the least squares means and the lines indicate their 95% confidence intervals. HRQL measures were adjusted for RA duration, pain VAS, self-reported RA control and severity, tender joint count, RAQoL score, and number of people in the household. Higher scores indicate better HRQL.
CHAPTER 7

ARE INDIRECT UTILITY MEASURES RELIABLE AND RESPONSIVE IN RHEUMATOID ARTHRITIS PATIENTS?

7.1 FOREWORD

This manuscript is currently under review under the same title in the *Quality of Life Research*. The candidate is the first author of the manuscript with is co-authored by Daphne Guh who provided statistical support; Dr. Amir Adel Rashidi who assisted with data entry, database construction, and data manipulation; Dr. Jacek Kopec, a member of the candidate’s committee who supplied analytical advice; Dr. Michal Abrahamowicz who invented the polytomous regression techniques for responsiveness and instructed the candidate in their application; and Dr. John Brazier who developed the SF-6D and provided methodological advice. Drs. Aslam Anis and John Esdaile, co-supervisors of the candidate, were also co-authors on the manuscript.

The candidate’s role in the manuscript involved the conception of the research question, development of the primary hypothesis and methodology, all statistical analyses, and the writing of the final manuscript.

7.2 INTRODUCTION

Improvement in health quality of life (HRQL) is considered to be one of the most important goals in the management of rheumatoid arthritis (RA)\(^1\). As such, HRQL and health status measures have often been used as outcomes in clinical trials and studies assessing a
A variety of instruments that assess RA-specific HRQL (for example, the Arthritis Impact Measurement Scales (AIMS), the Rheumatoid Arthritis Quality of Life questionnaire (RAQoL)) or generic HRQL or function (such as the Short Form 36 (SF-36)) have been applied to the assessment of RA.\textsuperscript{2,6,7}

Preference-based or indirect utility measures are generic HRQL measures that are often used in clinical and observational studies as the scores that they generate can be utilized to calculate quality adjusted life-years (QALYs) and can thus be integrated into cost-utility analyses.\textsuperscript{8} Examples of these instruments include the Health Utilities Index Mark 2 (HUI2) and Mark 3 (HUI3), EuroQol (EQ-5D), and the Short Form 6D (SF-6D). All of these instruments have been previously applied in the assessment of patients with RA.\textsuperscript{9-11}

Responsiveness is often defined as the ability of an instrument to measure change;\textsuperscript{12} however, there are multiple definitions of responsiveness that exist in the literature. These definitions can be divided into three categories:\textsuperscript{13} 1) ability of an instrument to detect change in general (also referred to as “sensitivity” by Liang et al.)\textsuperscript{14}; 2) ability of an instrument to detect clinically important change; and 3) the ability of an instrument to detect real changes in the concept being measured.

There has been little work in the evaluation and comparison of responsiveness (using any definition) of the indirect utility instruments. A recent study by Conner-Spady et al.,\textsuperscript{11} examined the responsiveness of three preference-based measures of HRQL (EQ-5D, HUI3, and the SF-6D) in a sample of patients with at least one of several types of rheumatological conditions. To our knowledge, there have been no evaluations of the responsiveness of the RAQoL in RA in North American populations although one has been published in a Swedish sample.\textsuperscript{7} Therefore, there remains a need for more research to assess the responsiveness of
these measures, to compare their characteristics, and to determine how their properties compare to disease-specific measures. Finally, since the indirect utility measures are often used as the source of weightings used for QALYs in cost-utility studies in RA, it is important that they are determined to be reliable, valid and responsive in this disease state.

Therefore, the primary purpose of this study was to examine the reliability and responsiveness of the indirect utility instruments and the RAQoL and the HAQ from baseline to six months in a sample of rheumatoid arthritis patients. A secondary purpose was to examine the reliability and validity of using a patient completed transition question as the external criterion to assess responsiveness in RA.

7.3 METHODS

7.3.1 Study Sample

To be included, subjects had to have a rheumatologist-confirmed diagnosis of RA (as defined by the American College of Rheumatology diagnostic criteria)\textsuperscript{15}, receive rheumatology care within the province of British Columbia, consent to answer the questionnaires, be sufficiently proficient in English to answer the questionnaires, and be willing to participate in follow-up surveys. Recruitment of RA patients began in October 2001 and ended in September 2002. Ethical approval for this study was obtained through the University of British Columbia’s Behavioural Ethics Committee and informed consent was obtained from each of the participants.

Eight private rheumatologists’ offices from the study areas referred subjects into the sample during their interactions in routine clinical practice. In addition, two of the eight rheumatologists’ practices sent letters to all of their patients with RA inviting them to
participate in the survey. All patient questionnaires were self-administered, self-completed and submitted via mail. The study rheumatologists' offices supplied additional information from the patients' health record.

7.3.2 Measures

Participants were asked to complete a questionnaire at baseline and three and six months thereafter. The questionnaire consisted of sections devoted to socio-economic, clinical and functional status and quality of life assessment instruments.

7.3.2.1 Clinical

Participants were asked questions regarding their RA and medication history including adverse reactions over the past three months. Other self-reported clinical variables included swollen joint count (SJC) and tender joint count (TJC) (using the mannequin-based 42 joint count methodology)\textsuperscript{16}, a 10 cm pain visual analogue scale (VAS), a patient global assessment of disease activity (10 cm VAS)\textsuperscript{1}, and RA severity and RA control (both using a 5 point Likert scale). Erythrocyte sedimentation rate (ESR) values closest to the date of completion of the questionnaire (within 1 month) were extracted from the patient's chart for those patients whose rheumatologist used this measure for patient monitoring. In addition, the attending rheumatologists were asked to complete a physician global assessment of disease activity (10 cm VAS) for each patient.\textsuperscript{1}

In addition, for the six-month questionnaire, participants were asked to complete a five point Likert scale that assessed changes in their RA since answering the baseline
questionnaire. The question asked was “Overall, how would you describe changes in your rheumatoid arthritis since answering the FIRST questionnaire (i.e. about 6 months ago?)”). Response choices included “Much Worse”, “Somewhat Worse”, “The Same”, “Somewhat Better” and “Much Better”. These questions are referred to as “patient transition questions” for the remainder of the manuscript. To increase the number of patients in each category, responses to these questions were collapsed into three categories as follows: (1) worse (included responses “much worse and somewhat worse”); (2) the same; and (3) better (included “much better and somewhat better) which is a similar approach adopted by other investigators.9,12,14

The sample of RA patients in our study experienced “natural” courses of their disease over time rather than exposure to a treatment of known efficacy, administered in a randomized design. In group level analyses, average change scores can mask the proportion of patients with follow-up scores that differ from those at baseline. Because of this, we carried out separate analyses for each of the distribution-based responsiveness measures according to our collapsed transition question criteria (“worse”, “the same”, or “better”). This is the same approach used by other investigators.11,17

7.3.2.2 Health Status and HRQL Measures

7.3.2.2.1 Health Assessment Questionnaire (HAQ) Disability Index

The HAQ is a measure of physical disability that assesses ability to complete everyday tasks in areas such as dressing and grooming, rising, eating, walking, personal hygiene, reach, grip and other activities (such as getting into and out of a car). Each of these areas is assigned a section score that is further adjusted to account for the use of any aids,
devices or help from another person. Section scores are then summed and averaged to give an overall score between 0.0 (best possible function) to 3.0 (worst function). A HAQ score difference of 0.25 is said to represent the minimally important difference (MID).\textsuperscript{18,19}

7.3.2.2.2 Rheumatoid Arthritis Quality of Life Questionnaire (RAQoL)

The RAQoL consists of 30 questions (answered by yes/no) that assess such aspects of RA as moods and emotions, social life, hobbies, everyday tasks, personal and social relationships, and physical contact. The RAQoL is scored by assigning a point for each affirmative response and no points for negative responses. Thus, scores range from 0 (least severity) to 30 (highest severity). To date, the MID for the RAQoL has been estimated to be approximately 2.00.\textsuperscript{20}

7.3.2.2.3 Preference Based Measures – MAUT Instruments

The indirect utility assessment instruments used in the questionnaire were the HUI2, HUI3, SF-6D, and the EQ-5D. In a cross-sectional analysis in patients with RA, the MID for the overall utility scores was determined to be 0.03 to 0.04 for the HUI2, 0.06 to 0.07 for the HUI3, and 0.03 to 0.05 for the SF-6D and the EQ-5D.\textsuperscript{20} In another analysis of seven longitudinal studies examining SF-6D global utility scores, investigators estimated that the MID to be 0.033 (95% CI: 0.029 to 0.037).\textsuperscript{10} A recent comprehensive review of the similarities and differences across these instruments is available and is beyond the scope of this research paper.\textsuperscript{21}
7.3.3 Data Analysis

7.3.3.1 Reliability

The reliability of a transition question to assess changes in health status has not previously been studied. To determine reliability of the patient transition question and the other questionnaires, a second questionnaire was sent to a randomly selected group of 50 patients immediately after receipt of their follow-up questionnaire with the instructions to complete and return within 5 weeks. The five week period was chosen as this was determined a priori to be the time window in which changes (either improvement or deterioration) in their RA would be unlikely. Patients were instructed to answer the transition question in relation to their baseline questionnaire.

Another approach to test-retest reliability was the "stable groups" approach comparing scores from patients who reported that they remained stable from 0 to 6 months. For all analyses, intraclass correlation coefficients (two-way mixed effect model such that the subject effect was random and the instrument effect was fixed) were calculated for the overall scores from the two time periods.

7.3.3.2 Validity of the Transition Question

Similarly, no assessments have been conducted to determine the validity of the transition question with respect to assessing rheumatoid arthritis. To determine agreement with the results from the collapsed transition question and changes in the HRQL instruments (divided into "worse", "same" and "better" categories using MID values from the literature as cut-off points), a weighted kappa (using quadratic weights) was calculated where 1.00 signifies perfect agreement and 0.00 signifies no agreement. We also estimated the MID

224
using values that were calculated from this study using those who scored either “somewhat better” or “somewhat worse” on the transition question assuming that the change experienced by these patients was equivalent to the MID as noted by other investigators. Both anchor- and distribution-based approaches to assessing MID values have been shown to yield similar values in many situations.

To further examine the validity of using the transition question as the external criteria, we examined the Spearman’s correlation between the transition question and changes in variables that were found to exhibit strong correlation with the generic and disease-specific HRQL measures in cross-sectional analyses (patient global assessment of disease severity, pain visual analogue scale (VAS), and the HAQ disability score). Spearman’s rho of > 0.50 or < -0.50 were considered be strong, while values between -0.49 to -0.30 or 0.30 to 0.49 were considered moderate and values between -0.30 and 0.30 were considered to be weak. For purposes of this analysis, a correlation of moderate or greater was considered to be evidence of validity.

7.3.3.3 Measures of Responsiveness

Our analysis focused on the assessment of responsiveness to change in RA for the indirect utility measures (the HUI2, HUI3, SF-6D and the EQ-5D), the RAQoL and the HAQ. Analysis for responsiveness was completed for the baseline and six month pairs of responses. For each patient who had data on all instruments at each of the pair of visits, the difference between the two corresponding indirect utility, RAQoL, and HAQ scores was calculated. In the primary analysis of responsiveness, the results were stratified into patients classified as “better”, “the same”, or “worse” according to the collapsed transition question.
In addition, in a secondary analysis, utilizing the patient global assessment of disease activity (called “patient global” hereafter), the percentage improvement over baseline was calculated utilizing the following formula:

\[
\left(\frac{6\text{mos}\.patient\text{global} - \text{baseline}.patient\text{global}}{\text{baseline}.patient\text{global}}\right) \times 100
\]

According to this formula, patients were classified as: 1) “better” if the patient global had changed by ≥ 20%, 2) “the same” if the patient global had changed > -20% and < 20%; and 3) “worse” if the patient global had changed < -20%. Spearman correlation coefficients were calculated to determine the correlation between this classification criteria and the collapsed transition question. All the indices of responsiveness (as described below) were calculated for the subgroups defined by this criterion.

Five distribution-based approaches were employed to assess responsiveness:

1) the effect size (ES)\textsuperscript{26} using the following formula:

\[
\frac{\text{mean}(x_1 - x_2)_{total\text{group}}}{SD_{total\text{group}}}
\]

where:

\( x_1 = \text{mean score at 6 months for the entire group} \)

\( x_2 = \text{mean score at baseline for the entire group} \)

\( SD_{total\text{group}} = \text{standard deviation at baseline for the entire group} \)
An effect size of 1 indicates a change in magnitude equivalent to one standard deviation. We adopted the criteria of Cohen, where absolute values of effect sizes (d) can be categorized as small (< 0.5), medium (0.5 to 0.8), or large (> 0.8). Positive values reflect improvement while negative values reflect worsening for the indirect utility instruments while the converse is true for the HAQ and the RAQOL.

2) the standardized response mean (SRM)\textsuperscript{13} using the following formula:

\[
\frac{\text{mean}(x_1 - x_2)_{\text{total group}}}{SD(x_1 - x_2)_{\text{total group}}}
\]

where

\(x_1 =\) mean score at 6 months for the entire group

\(x_2 =\) mean score at baseline for the entire group

SD \((x_1 - x_2)_{\text{total group}} =\) the standard deviation (SD) for the change in scores in the entire group.

The absolute values of the SRM are regarded as either small (<0.5), medium (0.5 to 0.8) or large (>0.8) and the signs (either positive or negative) are interpreted as for the ES.\textsuperscript{28}

3) the control standardized response mean (CSRM)\textsuperscript{17} using the following formula:

\[
\frac{\text{mean}(x_1 - x_2)_{\text{total group}}}{SD(x_1 - x_2)_{\text{no change}}}
\]
where the mean change in the total group refers to the mean change in the subgroups of “worse”, “same”, and “better” (as per the external criteria) and the standard deviation is taken from the subgroup reporting no change. The criterion for the size of the CSRM is the same as for the ES and SRM.

4) the relative efficiency statistic (RE)\textsuperscript{29} using the following formula:

$$\left( \frac{t_{\text{comparison}}}{t_{\text{goldstandard}}} \right)^2$$

Given the information on the superior responsiveness of disease-specific over generic measures,\textsuperscript{30} we selected the RAQoL as the “gold standard” which to compare each of the instruments. The measure with the highest RE has the highest power for a given sample size, or requires fewer patients, to achieve a given level of statistical power.\textsuperscript{12}

5) paired sample t-tests reported as a p-value\textsuperscript{13}

Since the standard errors of the distribution-based approaches are not defined, we used bootstrap methods to estimate 95% confidence intervals (CI) for the ES, SRM, and the CSRM.\textsuperscript{10} Rather than conduct a large number of statistical tests, the 95% CIs were investigated to determine the degree of overlap between the values generated across the HRQL measures. Also, since it is well-known that the results of these indices sometimes generate conflicting results,\textsuperscript{12,31} we ranked the order of the values according to the responsiveness statistic and calculated the overall median value across the responsiveness statistics to determine the overall rank.
The distribution-based methods described above do not provide answers to practical questions such as, for example, how likely is a decrease in a specified amount in the utility score (as measured by the indirect instruments) to represent actual deterioration? Thus, we utilized a flexible polytomous regression model to assign probabilities of patient’s improvement, status quo, or deterioration (as defined by the transition question) to different levels of change in the indirect utility and disease specific HRQL measures. The results of this polytomous regression are presented in a graph of 3 curves, each of which describes how the estimated probability of a respective outcome (improvement, no change, or worsening as defined by the collapsed transition question or the patient global assessment of disease activity question), changes as a function of the difference in two consecutive scores.

Finally, we examined associations between changes in either the unweighted domain scores of the EQ-5D and the SF-6D (as these instruments do not typically calculate single-attribute utility values) or the single-attribute utility scores of the HUI2 and HUI3 with the external criteria. The purpose of these analyses was to investigate which domains/single attributes were most likely to change in response to improvement or worsening in RA (as defined by the external criteria). Statistical analysis using Kruskal-Wallis was employed. Conservatively, we defined a clear association if the test for was significant for the domain or single attribute with both external criteria.

7.4 RESULTS

7.4.1 Demographics and Missing Values

Of the 320 RA patients who returned the baseline questions, 239 returned the six month questionnaires for a 75% follow-up response rate. Characteristics of our baseline
sample have been described in detail elsewhere. Baseline characteristics of those who completed the six month questionnaires compared to those who did not are shown in Table 7.1. For most of the variables examined there were no differences between the baseline characteristics of those who completed the baseline questionnaire as compared to those who did not. However, for all of the instrument scores, those who completed the six month questionnaires appeared to have poorer baseline mean HRQL scores than those who did not (with the exception of the HUI2) but this relationship was statistically significant only for the HAQ. Other variables that differed between the subgroups were self-reported severity and proportion who worked outside the home in the past 12 months (both favouring those only completing the baseline questionnaire).

7.4.2 Reliability

Test-retest reliability for the collapsed categories of the transition question ("worse", "same", and "better") using the follow-up questionnaire and a subsequent questionnaire within 5 weeks of these responses yielded 38 valid (received within the stated time frame) responses and perfect agreement in 36 patients. In the two patients not showing agreement, both returned their reliability questionnaire 14 days after the follow-up questionnaire, one assessed change in his/her RA at 3 months as "better" but "worse" 14 days later, while the other assessed change at 3 months as "worse" but "better" 14 days later. Therefore, the ICC value for the collapsed transition question was 0.80 (95% CI 0.64 to 0.89) with these two responses included and 1.00 if these are eliminated. The results for the test-retest reliability approach for the generic and disease specific instruments are shown in Table 7.2. Using the stable groups approach (i.e. those reporting no change from 0 to 6 months), we also
determined ICC values to examine the reliability for the generic and disease-specific instruments (Table 7.3). Results were similar to the test-retest reliability approach in that reliability of the EQ-5D overall score appeared to be the lowest while the RAQoL and the HAQ displayed the highest reliability.

7.4.3 Validity of the Transition Questionnaire

For the 0 to 6 months transition question, 96 (40%) reported improvement, 85 (36%) reported no change and 58 (24%) reported worsening. Of these, 222 patients had pairs of answers on all questionnaires to permit comparisons (89 reporting improvement, 77 reporting status quo and 56 reporting worsening). For the secondary external criterion (as defined by categorization of the patient global assessment of disease severity VAS) for these 222 pairs, results of the patient global scores were available and were classified as follows: 65, 118, and 39 reporting improvement, status quo and worsening using criterion described in the Methods section. The two external criteria had fairly low agreement (weighted kappa 0.30, 95% CI 0.20 to 0.41).

To examine the agreement between results of the collapsed transition question (improved, status quo, worsened) and the HRQL measures categorized based upon the literature-based MID values, we plotted the results as bar graphs and calculated weighted kappa values (Figure 7.1). For all the instruments, agreement between the transition question and the categories of the HRQL values was relatively low (weighted kappa ranging from 0.15 to 0.28). MID values calculated from our longitudinal sample using the anchor-based approach yielded values somewhat smaller than those reported in the literature (Table 7.4).
Spearman's correlations between the transition question responses and changes in the patient global assessment of disease activity VAS, the pain VAS, and the HAQ disability score are shown in Table 7.5. Correlations between the transition question responses and the RA outcome measures were similar in magnitude to those between the RA outcome measures. Changes in the patient global assessment of disease activity VAS and the HAQ score displayed moderate correlation with the transition question.

7.4.4 Responsiveness

The mean change scores for each of the instruments between six months and baseline are shown for the entire sample and stratified by results of the transition question in Table 7.6 and for the categories defined by the patient global assessment of disease activity in Table 7.7. As hypothesized, for many of the instruments, since the sample was experiencing "natural" changes in their disease over time, the change scores for the entire sample tended to obscure the changes in the subgroups. Scatterplots of the indirect utility scores over time (from three measurements at baseline, 3 and 6 months) are presented in Figures 7.2 to 7.5 with ordinary least squares regression lines depicting the overall trends. Most of these lines had slopes in the hypothesized direction (positive for "better" and negative for "worse" as defined by the collapsed transition question). For those who reported their RA as "the same", slopes of the regression line tended to be positive across the indirect utility measures. Also of note, within each instrument, the average scores at baseline between the three groups as defined by the collapsed transition question were different with those stating that their RA had worsened having either lower baseline indirect utility scores (for the HUI2, HUI3, SF-6D, and the EQ-5D) and higher (and therefore, worse) RAQoL and HAQ scores. Those
reporting that their RA was “the same” after six months tended to have better HRQL scores for all the instruments at baseline than the “better” and “worse” categories. For each of the measures, on average, changes for those reporting “better” or “worse” were in the appropriate direction (i.e. for the indirect utility scores, positive and negative, respectively; whereas, for the RAQoL and the HAQ, this was reversed). These findings were similar when the external criterion for change was changed to categories based upon changes in the patient global assessment of disease severity VAS (Table 7.7).

The indices of responsiveness (ES, SRM, CSRM, paired t-test and the RE) and their associated 95% CI for those who responded as better, the same or worse according to the transition question are presented in Table 7.6, according to the patient global rating of disease severity VAS (Table 7.7) and the rankings of the various responsive statistics are shown in Table 7.8. Generally, the results of the various responsiveness statistics tended to agree within each of the instruments (Table 7.8) and there was little overlap between their 95% CI. Overall, the RAQoL was the most consistently responsive of the instruments tested regardless of which of the external criteria were applied. Depending on whether the change was classified as either “worse” or “better” and which of the external criteria were applied, the indirect utility instruments and the HAQ displayed varying degrees of responsiveness. For example, the EQ-5D appeared to be responsive in those who were classified as “worse” irrespective of which external criteria were applied but unresponsive in those classified as “better”. The HAQ appeared to be relatively responsive in both those classified as better or worse using the patient transition question to define the groups, but less responsive (in relation to the other instruments) when the patient global assessment of disease severity criterion was applied. The HUI3 appeared to be relatively unresponsive except in those
classified as "better" by the patient global assessment of disease severity. The HUI2 was consistently ranked among the middle in responsiveness and the SF-6D appeared to be more responsive in those classified as "better" (by either criterion) than those classified as "worse".

7.4.5 Flexible Polytomous Regression Techniques

Results from the flexible polytomous regressions exploring responsiveness are shown in Figures 7.6 to 7.17. The curves on each figure correspond to the 3 types of outcome (worse, same, better) as defined by the external criteria (patient transition question or the patient global assessment of disease activity). Each curve shows how the estimated probabilities of a specific response vary depending on the observed change in the scores of the instruments.

In general, the results of using the patient global assessment of disease activity VAS appear to be better able to discriminate between those patients whose RA has improved, worsened or stayed the same than the transition question. This is evident in all of the graphs as there is a sharper delineation between the three curves (worse, better and same) in Figures 7.12 to 7.17 than in the corresponding Figures 7.6 to 7.11 (for example, comparing Figure 7.6 to Figure 7.12, of which both examine changes in the HUI2). Overall, the RAQoL appeared to be most responsive in both Figure 7.10 and Figure 7.16 as compared to the other instruments using the same external criterion. For example, in Figure 7.16, there is very good discrimination between the three curves as shown by their degree of separation. The probability of being classified as "the same" is high (approximately 60%) if the difference between the two scores is zero. Similarly, this probability decreases as we move in either
direction and becomes extremely small when the difference is ± 20. As the difference in the scores gets larger in the positive direction (recall that larger values in the RAQoL reflect worse HRQL), the probability of being classified as “worse” grows to > 80% when the difference in scores is approximately 15 and almost 100% when the difference is 20. These values are similar to those displayed for negative values (reflecting improvement) in the RAQoL and the dashed curve labeled as “better”.

For the indirect utility instruments for the graphs using the patient transition question as the external criteria for change, there was generally fairly poor discrimination between the curves with significant overlap between the probabilities of being classified “better”, “worse” and “same” across the range of difference scores (Figures 7.6 to 7.10). Using the patient global assessment of disease activity VAS criteria, the curves for all the indirect utility instruments showed much better discrimination between those classified as “better” and “worse” (Figures 7.12 -7.15). However, for those classified as the “same”, there was considerable overlap between these probabilities and the probabilities for “better” and “worse”. The HUI3 appeared to be the best able to discriminate in this regard (Figure 7.13). Thus, it would seem that although these instruments can discriminate change well (according to the external criterion) in those who improve or worsen, those that stay the same yield somewhat problematic difference scores. This finding could be a property of the instruments or may be a reflection of the cut-off values of our external criterion.

Similarly, for the HAQ, the patient global assessment of disease severity VAS criterion appeared to result in better discrimination between the curves; however, as with the indirect utility measures, there was considerable overlap between the “same” category and the other categories.
7.4.6 Change in Unweighted Domain Scores (EQ-5D, SF-6D) and Single Attribute Utilities (HUI2, HUI3)

The associations between the instrument unweighted domains (EQ-5D and SF-6D) and the single attribute scores (HUI2 and HUI3) and the external criteria of change are shown in (Table 7.9). For the EQ-5D, pain/discomfort, anxiety/depression and self-care, and for the SF-6D, physical, and social functioning, role limitations and pain met our criteria for statistical significance. For the single attributes from the HUI systems, ambulation, emotion, and pain (from the HUI3) and mobility, emotion and pain (from the HUI2) met the criteria. Of note, there were more significant associations between the domains/single attributes and the changes defined by the patient global assessment of disease severity categories than the patient transition question responses. For example, with the EQ-5D there was a significant association between the mobility domain in the patient global assessment of disease severity VAS defined changes but not for the other external criterion. For the SF-6D, HUI3, and HUI2 there were significant associations for the vitality domain, the dexterity single attribute, and the sensation single attribute, respectively, using the patient global assessment of disease severity VAS defined changes. Of note, for the self-care single attribute in the HUI2, there was a significant association between the patient transition defined changes but not the other criterion.

7.5 DISCUSSION

This study is the first to compare the reliability and longitudinal changes in scores obtained with four indirect utility instruments (HUI3, HUI2, EQ-5D, SF-6D), a disease-
specific measure (the RAQoL), and a disability measure (the HAQ) in a sample of patients with rheumatoid arthritis. Our results demonstrate that while the generic, preference-based measures yielded scores that were generally reliable, they had lower responsiveness (as assessed by multiple methodologies) in RA than the disease-specific RAQoL. The indirect utility measures did, however, yield moderate responsiveness statistics when the patient global assessment of disease severity was applied as the external criterion for change. The domains and attributes of the indirect utility instruments that were commonly associated with the external criteria for change in RA tended to be pain, ambulation/physical functioning, and emotional/mental health.

We also examined the reliability and validity of utilizing a patient-completed transition question to function as the primary external criteria of change which is a common approach. However, the main concern with this approach is recall bias. Some literature suggests that retrospective estimates of the initial state are often highly correlated with the present state and uncorrelated with the initial state. Another concern deals with the starting point of individuals who are rating their health changes. For example, an individual starting at a lower point in health or function may rate a small change as significant where a person of higher function may regard change of the same magnitude as insignificant. Despite these concerns, there has been little work in evaluating the reliability and validity of transition questions. In this study, we found that the reliability of the transition question was acceptable using the test-retest approach. However, from the validity standpoint, we found that the responses from the transition question were only lowly to moderately correlated with commonly accepted clinical variables used to assess RA (such as the pain VAS, the patient global assessment of disease severity VAS, and the HAQ). The patient
transition defined changes also had low agreement with previously defined MID values and the changes defined by the patient global assessment of disease severity VAS. In addition, we found that there were fairly large mean differences in the instruments between the time points for individuals who were classified as being the "same" from their RA perspective (sometimes the change in this category was of similar magnitude as those classified as "worse" or "better"). This point was illustrated in the polytomous regression plots where there was considerable overlap between the "same" and "better" or "worse" curves. While this finding could be the result of shortcomings of the instrument in assessing changes in RA, these findings were not observed when a different external criterion was applied (categories based upon the patient global assessment of disease activity VAS). Also, several single attributes that were expected to have significant associations with changes in RA were significantly associated with changes in the patient global assessment VAS and not the patient transition question changes (mobility (EQ-5D), vitality (SF-6) and dexterity (HUI3)). Therefore, categorization of the patient global assessment of disease activity VAS appears to be a superior external criterion for RA than the patient transition question as it was expected that these domains/single attributes would be associated with changes in RA. Therefore, we would hypothesize that these are the main factors that are driving the observed changes in the global utility scores.

Generally, dividing the sample into "worse", "same" and "better" using the patient global assessment of disease severity VAS categories seemed to more accurately define these groups than the patient transition question. This point is illustrated by the larger responsiveness statistics for all of the instruments, the smaller amount of change in all of the instruments in those classified as having their RA being the "same" as at baseline, and a
greater magnitude of change (either negative or positive) in those classified as having their RA “worse” or “better” than baseline. Using the transition question as the external criterion resulted in small ES, SRM and CSRM statistics for virtually all of the instruments and non-significant p-values on the paired t-tests for those who reported to have improved or worsened from baseline for many of the indirect utility measurements (Table 7.6). Conversely, when applying the classification according the patient global assessment of disease severity VAS (Table 7.7), many of the responsiveness statistics for those classified having their RA improved or worsened over baseline can be interpreted as moderate or large, and all of the paired t-tests for those who improved or worsened were significant for all of the instruments.

The indirect utility instruments displayed different properties in this study. Reliability was acceptable for all of the scores except for the EQ-5D (ICC 0.46 to 0.52 depending on methodology employed). This finding is considerably lower than previously reported in rheumatoid arthritis (ICC of 0.73 using the stable groups approach and 0.78 using test-retest reliability). The differences in these two findings may be due to the five week window for resubmission of the reliability questionnaires in our study compared to two weeks in the other analysis. In the longer time frame, it is possible that there was a higher probability for change. This change may have penalized the EQ-5D much more than the other scales as there is a term in the EQ-5D scoring function (N3) that subtracts 0.269 if a score of the lowest level (3) occurs on at least one domain. Thus, a one category change (from “2” to “3”) in response in a single domain can have profound implications for reducing the EQ-5D utility score. However, other instruments which were found to be more responsive than the EQ-5D were stable (the RAQoL and the HAQ) over this time frame.
The HUI2 and the HUI3 generally had low responsiveness statistics utilizing the patient transition question as the external criteria and moderate responsiveness statistics when the categories of the patient global assessment of disease activity VAS were applied. Their relative rankings were towards the middle or bottom for all of the instruments regardless of the external criteria applied accept for the “better” category as defined by the patient global assessment of disease activity. For this category, the HUI3 had the highest responsiveness statistics in three categories (the ES, SRM, and the paired sample t-test). This was likely due to the observation that the mean change in this category was quite large (0.17) which was almost half of the baseline score. In the polytomous regression plots, the HUI3 appeared to have less overlap between the same and the better or worse curves than the other indirect utility instruments (i.e. Figure 7.13) which may make it more responsive in RA. As expected, the sensation attribute (HUI2), the vision, hearing and speech attributes (HUI3) and the cognition attributes (both scales) were not associated with the external criteria defined change in RA. Of note, although one would have expected dexterity (HUI3) and self-care (HUI2) to be consistently associated with changes in RA, each was only significant for only one of the external criteria.

The SF-6D generally had low responsiveness statistics utilizing the patient transition question as the external criteria and moderate responsiveness statistics when the categories of the patient global assessment of disease activity VAS were applied. This latter finding was especially true for the “better” category. In the rankings of the responsiveness statistics, the SF-6D had much higher rankings for those classified as improved (median rankings of 3 for both external criteria) compared to those classified as worsened (median rankings of 5 and 6). One of the problems with the responsiveness of the SF-6D when using our external
criteria was the amount of change experienced by those categorized as the “same”. Both paired t-tests for this category using each of the external criteria were significant indicating a large degree of mean change (0.04 in Table 7.6 which was as large as those reporting improvement and 0.02 in Table 7.7). These results are further illustrated in the Figures 7.9 and 7.14 with the probability of being scored as the same being somewhat constant over the range of SF-6D change scores.

As anticipated, the RAQoL was the most responsive to changes in both positive and negative directions which are in agreement with other research comparing disease-specific to generic HRQL instruments. The responsiveness statistics were generally moderate to large irrespective of the external criteria of change applied and were consistently in the top 2 in the rankings (Table 7.8). In addition, the results of the polytomous regressions reveals well delineated curves for same, better and worse without a large degree of overlap (Figures 7.10 and 7.15).

Results for the HAQ revealed that this instrument performed approximately equivalently for both of the external criteria with responsiveness statistics of similar magnitude. However, when compared to the other instruments, the HAQ rankings were among the highest for responsiveness statistics calculated from categories defined by the patient transition question but were either in the middle (for those categorized as worse) or at the bottom (for those categorized as better) for responsiveness statistics calculated from categories defined by the patient global assessment of disease severity VAS. Although the reason for this finding is not obvious, perhaps the patient transition question is capturing mostly changes in elements of disability (as measured by the HAQ) rather than other aspects/domains of RA which are being captured by the other instruments.
In summary, the RAQoL was consistently the most responsive of the tested instruments. Among the indirect utility instrument’s overall utility scores, the EQ-5D appeared to be the most responsive to worsening but not to improvement. Conversely, the HUI3 and SF-6D were superior in detecting improvement but the SF-6D detected changes in those classified as the “same”. Thus, in RA clinical trial situations where a known effective intervention is to be applied and there is a large probability of positive change, the SF-6D and the HUI3 would be superior to the other instruments. However, changes in the SF-6D might be larger as many patients classified as the same by other criteria would, in fact, improve using this scale. The HUI2 appeared to be fairly non-responsive in RA in comparison to the other measures.

We located two other studies that compared the responsiveness of indirect utility instruments in longitudinal sample of patients with a variety of musculoskeletal diseases. In the study that was conducted exclusively in RA, investigators examined the reliability and responsiveness of the SF-36, SF-6D, EQ-5D, standard gamble (SG), the modified HAQ, and a pain VAS in two groups of RA patients (Group 1 consisted of 24 patients with stable RA and Group 2 consisted of 60 patients beginning infliximab therapy). Patients in group 2 were assessed prior to being initiated on infliximab therapy and after 14 weeks of infliximab treatment.

Test-retest reliability was estimated for each instrument in the stable patient group using the ICC whereas responsiveness was assessed by using the paired t-test, effect size (ES) and standardized response mean (SRM). For all the measures, the ICC ranged from 0.50 (role emotional domain from the SF-36) to 0.92 (physical functioning domain from the SF-36). The preference-based measures had moderate reliability (ICCs of: EQ-5D 0.66, SF-
However, the sample from which these results were derived was very small (n=24) and thus, these estimates are not stable. In terms of responsiveness, for Group 2; all the overall scores and domain scores for the SF-36 detected significant changes from baseline to the second measurement. Standardized response means (SRM) and ES were the largest for the pain VAS, the EQ-5D VAS, the SF-36 physical component scores, and the SF-36 vitality domain. In terms of the preference-based measures, the SRM and ES values were 0.67 and 0.64 for the EQ-5D, 1.40 and 0.87 for the SF-6D, and 0.49 and 0.43 for the SG. Despite the fact that the change described by the EQ-5D system was twice that described by the SF-6D, the responsiveness statistics were much smaller mainly due to the larger SD of the baseline and change scores of the EQ-5D. The authors concluded that the SF-6D might be a preferable to the EQ-5D in measuring clinically-relevant improvement in RA.

In the other study, conducted in patients who had one of many rheumatic diseases, investigators utilized a five-point transition question for patient's to self-report their health changes from baseline to 12 months and then subsequently collapsed the transition question to three categories (better (n=40), same (n=30), and worse (n=28)). The results of the responsiveness analyses were somewhat different than ours. The ES for both those categorized as better or worse was moderate (0.53 and -0.58, respectively) for the EQ-5D and were much larger than those reported for the HUI3 and the SF-6D. Possible reasons for these differences include the sample size (n=98 pairs vs. n=222 pairs in our sample), sample characteristics (only half had RA while the balance was a mixture of other rheumatologic conditions with potentially more or less propensity for change), the recall period for the transition question (12 months vs. 6 months), and the lack of validity testing of the transition
question (i.e. we found that the transition question may not be the most valid external criteria to categorize the sample).

Blanchard et al.,\textsuperscript{12} examined the responsiveness of the HUI2 and the HUI3 (and other, disease-specific measures) in a sample of osteoarthritis patients (n=90) undergoing total hip arthroplasty. These investigators found that, although the disease-specific measures were the most responsive to this known, effective intervention, the preference-based measures had responsiveness statistics that were moderate to large in magnitude and were able to detect change resulting from THA. In addition, the pain (both HUI2 and HUI3), ambulation (HUI3) and self-care (HUI2) single-attributes had moderate to large ES.

Although the findings of the present study may provide valuable information on the application of these instruments in the assessment of changes experienced by patients with RA, there are some limitations to be considered when interpreting the results. Firstly, the results are specific to RA patients undergoing “natural” changes in their disease and may not generalize to other populations. Secondly, while the patient transition question appeared to suffer from several limitations in assigning categories of change to RA patients over six months, this finding might not generalize to other disease states/processes or shorter time frames of recall. In utilizing the secondary external criteria of change, we adopted ± 20% from baseline to be out cut-points in assigning categories of change. However, although this criterion is similar to what the American College of Rheumatology has used in determining the ACR 20 criteria\textsuperscript{1} and it appeared to perform better than the patient transition question, other cut-points may yield better results. Further studies are necessary to determine such cut-points.
We conclude that the reliability of all the instruments (except the EQ-5D) and the patient transition question was acceptable. The patient transition question might not be a valid external criterion for assigning categories of change to patients with RA and categories defined by the patient global assessment of disease severity appeared to perform better in this regard. The RAQoL was the most responsive although all the instruments were capable of detecting change to some degree. The HUI3 and the SF-6D may be the best indirect utility instruments to use in clinical trials of RA where a known effective intervention is to be applied.
REFERENCES


10. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health Qual Life Outcomes. 2003;11:4-12


TABLE 7.1: BASELINE CHARACTERISTICS OF THOSE SUPPLYING BOTH BASELINE AND SIX MONTH QUESTIONNAIRES COMPARED TO THOSE WHO ONLY COMPLETED THE BASELINE QUESTIONNAIRE

<table>
<thead>
<tr>
<th>Completed both baseline and six month</th>
<th>Completed only baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>239</td>
</tr>
<tr>
<td>Age, y (SD)</td>
<td>62.1 (13.1)</td>
</tr>
<tr>
<td>Duration of RA, y (SD)</td>
<td>13.8 (11.2)</td>
</tr>
<tr>
<td>Tender joints, range 0 – 28 (SD)</td>
<td>14.9 (11.5)</td>
</tr>
<tr>
<td>Swollen joints, range 0 – 28 (SD)</td>
<td>8.8 (8.5)</td>
</tr>
<tr>
<td>Pain VAS, range 0 – 100mm (SD)</td>
<td>38.5 (26.2)</td>
</tr>
<tr>
<td>Patient Global VAS of disease activity, range 0 – 100mm (SD)</td>
<td>59.0 (25.2)</td>
</tr>
<tr>
<td>Other chronic diseases</td>
<td>1.2 (1.3)</td>
</tr>
<tr>
<td>HUI2 utility score (SD)</td>
<td>0.71 (0.19)</td>
</tr>
<tr>
<td>HUI3 utility score (SD)</td>
<td>0.52 (0.29)</td>
</tr>
<tr>
<td>EQ-5D utility score (SD)</td>
<td>0.65 (0.24)</td>
</tr>
<tr>
<td>SF-6D utility score (SD)</td>
<td>0.62 (0.13)</td>
</tr>
<tr>
<td>RAQoL score (SD)</td>
<td>12.90 (8.1)</td>
</tr>
<tr>
<td>HAQ Score, range 0 – 3.0</td>
<td>1.2 (0.8)</td>
</tr>
<tr>
<td>Female (%)</td>
<td>79.2</td>
</tr>
<tr>
<td>Hospitalized for RA in last 12 months (%)</td>
<td>16.5</td>
</tr>
<tr>
<td>Used allied health professional/home care services for RA in last 12 months (%)</td>
<td>44.3</td>
</tr>
<tr>
<td>Self-reported RA severity (%)‡</td>
<td></td>
</tr>
<tr>
<td>Very Mild</td>
<td>1.3</td>
</tr>
<tr>
<td>Mild</td>
<td>10.2</td>
</tr>
<tr>
<td>Moderate</td>
<td>42.2</td>
</tr>
<tr>
<td>Severe</td>
<td>37.4</td>
</tr>
<tr>
<td>Very Severe</td>
<td>8.9</td>
</tr>
<tr>
<td>Self-reported RA control (%)</td>
<td></td>
</tr>
<tr>
<td>Very Well Controlled</td>
<td>10.2</td>
</tr>
<tr>
<td>Well Controlled</td>
<td>25.8</td>
</tr>
<tr>
<td>Adequately Controlled</td>
<td>42.6</td>
</tr>
<tr>
<td>Not Well Controlled</td>
<td>18.7</td>
</tr>
<tr>
<td>Not Controlled At All</td>
<td>2.7</td>
</tr>
<tr>
<td>Working outside the home in last 12 months (%)</td>
<td>17.4</td>
</tr>
</tbody>
</table>

SD = standard deviation
* p=0.04 by student t-test
† p=0.04 by Chi-square (df=4, Chi-square = 9.8)
‡ p=0.01 by Chi-square (df=1, Chi-square =6.7)
TABLE 7.2: TEST – RETEST RELIABILITY

<table>
<thead>
<tr>
<th>Instrument Attribute/Domain</th>
<th>ICC</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUI2</td>
<td>0.77</td>
<td>0.59 - 0.88</td>
</tr>
<tr>
<td>Sensation</td>
<td>0.81</td>
<td>0.63 – 0.90</td>
</tr>
<tr>
<td>Mobility</td>
<td>0.84</td>
<td>0.71 – 0.91</td>
</tr>
<tr>
<td>Emotion</td>
<td>0.72</td>
<td>0.52 – 0.84</td>
</tr>
<tr>
<td>Cognition</td>
<td>0.75</td>
<td>0.57 – 0.87</td>
</tr>
<tr>
<td>Self-care</td>
<td>0.36</td>
<td>0.05 – 0.60</td>
</tr>
<tr>
<td>Pain</td>
<td>0.52</td>
<td>0.25 – 0.72</td>
</tr>
<tr>
<td>HUI3</td>
<td>0.81</td>
<td>0.66 – 0.90</td>
</tr>
<tr>
<td>Vision</td>
<td>0.88</td>
<td>0.78 – 0.94</td>
</tr>
<tr>
<td>Hear</td>
<td>0.23</td>
<td>-0.05 – 0.47</td>
</tr>
<tr>
<td>Speech</td>
<td>1.00</td>
<td>-</td>
</tr>
<tr>
<td>Ambulation</td>
<td>0.82</td>
<td>0.68 – 0.90</td>
</tr>
<tr>
<td>Dexterity</td>
<td>0.78</td>
<td>0.63 – 0.88</td>
</tr>
<tr>
<td>Emotion</td>
<td>0.66</td>
<td>0.43 – 0.81</td>
</tr>
<tr>
<td>Cognition</td>
<td>0.46</td>
<td>0.15 – 0.68</td>
</tr>
<tr>
<td>Pain</td>
<td>0.59</td>
<td>0.34 – 0.76</td>
</tr>
<tr>
<td>SF-6D</td>
<td>0.89</td>
<td>0.79 – 0.94</td>
</tr>
<tr>
<td>Physical functioning</td>
<td>0.81</td>
<td>0.67 – 0.90</td>
</tr>
<tr>
<td>Role limitations</td>
<td>0.73</td>
<td>0.55 – 0.85</td>
</tr>
<tr>
<td>Pain</td>
<td>0.82</td>
<td>0.68–0.90</td>
</tr>
<tr>
<td>Mental health</td>
<td>0.77</td>
<td>0.60–0.87</td>
</tr>
<tr>
<td>Vitality</td>
<td>0.74</td>
<td>0.55–0.85</td>
</tr>
<tr>
<td>Social functioning</td>
<td>0.80</td>
<td>0.64–0.88</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>0.46</td>
<td>0.18 – 0.68</td>
</tr>
<tr>
<td>Mobility</td>
<td>0.74</td>
<td>0.55–0.85</td>
</tr>
<tr>
<td>Self-care</td>
<td>0.51</td>
<td>0.23–0.71</td>
</tr>
<tr>
<td>Usual activities</td>
<td>0.71</td>
<td>0.50 – 0.84</td>
</tr>
<tr>
<td>Pain/Discomfort</td>
<td>0.72</td>
<td>0.52 – 0.84</td>
</tr>
<tr>
<td>Anxiety/Depression</td>
<td>0.73</td>
<td>0.55 – 0.85</td>
</tr>
<tr>
<td>HAQ</td>
<td>0.97</td>
<td>0.93 – 0.98</td>
</tr>
<tr>
<td>RAQoL</td>
<td>0.93</td>
<td>0.86 – 0.96</td>
</tr>
</tbody>
</table>

Questionnaire results compared to results within 35 days of global scores, unweighted domain scores (EQ-5D and SF-6D), and single attribute utility scores (HUI2 and HUI3). Results are intraclass correlation coefficients (ICC) with 95% confidence intervals (CI)
TABLE 7.3: INTRACLASS CORRELATION COEFFICIENT VALUES FOR GENERIC AND DISEASE-SPECIFIC HRQL MEASURES FOR THOSE REPORTING NO CHANGE IN THEIR RHEUMATOID ARTHRITIS BETWEEN 0 AND 6 MONTHS

<table>
<thead>
<tr>
<th>Measure</th>
<th>ICC*</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUI2</td>
<td>0.72</td>
<td>0.60 – 0.81</td>
</tr>
<tr>
<td>HUI3</td>
<td>0.78</td>
<td>0.67 – 0.85</td>
</tr>
<tr>
<td>SF-6D</td>
<td>0.78</td>
<td>0.68 – 0.86</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>0.52</td>
<td>0.35 - 0.66</td>
</tr>
<tr>
<td>HAQ</td>
<td>0.90</td>
<td>0.84 – 0.93</td>
</tr>
<tr>
<td>RAQoL</td>
<td>0.86</td>
<td>0.78 – 0.91</td>
</tr>
</tbody>
</table>

*ICC = intraclass correlation coefficient (two-way mixed effect model such that the subject effect was random and the instrument effect was fixed)
TABLE 7.4: MINIMALLY IMPORTANT DIFFERENCES REPORTED IN THE LITERATURE AND DERIVED FROM THE SAMPLE USING ANCHOR-BASE APPROACHES

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Literature-derived MID* Values</th>
<th>Author, Reference #</th>
<th>Anchor-based MID* Values (SD)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Worse</td>
<td>Better</td>
<td></td>
</tr>
<tr>
<td>HUI2</td>
<td>0.03 – 0.04</td>
<td>8</td>
<td>-0.05 (0.19)</td>
<td>0.04 (0.13)</td>
<td></td>
</tr>
<tr>
<td>HUI3</td>
<td>0.06 – 0.07</td>
<td>8</td>
<td>-0.03 (0.25)</td>
<td>0.02 (0.22)</td>
<td></td>
</tr>
<tr>
<td>EQ-5D</td>
<td>0.03 – 0.05</td>
<td>8</td>
<td>-0.06 (0.21)</td>
<td>0.03 (0.22)</td>
<td></td>
</tr>
<tr>
<td>SF-6D</td>
<td>0.033</td>
<td>9</td>
<td>-0.02 (0.07)</td>
<td>0.02 (0.09)</td>
<td></td>
</tr>
<tr>
<td>RAQoL</td>
<td>2.00</td>
<td>8</td>
<td>1.39 (5.12)</td>
<td>-1.72 (5.56)</td>
<td></td>
</tr>
<tr>
<td>HAQ</td>
<td>0.25</td>
<td>6, 7</td>
<td>0.12 (0.50)</td>
<td>-0.10 (0.37)</td>
<td></td>
</tr>
</tbody>
</table>

*MID = minimally important differences
TABLE 7.5: CORRELATIONS BETWEEN THE TRANSITION QUESTION AND CHANGES IN RHEUMATOID ARTHRITIS OUTCOME VARIABLES FROM 0 TO 6 MONTHS

<table>
<thead>
<tr>
<th>Transition Question*</th>
<th>Patient Global Assessment of Disease</th>
<th>Pain VAS (100mm)</th>
<th>Health Assessment Questionnaire (HAQ) Score (0 – 3.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transition Question*</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient Global</td>
<td>0.41</td>
<td>-0.44</td>
<td>1.00</td>
</tr>
<tr>
<td>Assessment of Disease</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity VAS (100mm)</td>
<td>0.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain VAS (100mm)</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health Assessment</td>
<td>-</td>
<td>-0.29</td>
<td>0.39</td>
</tr>
<tr>
<td>Questionnaire (HAQ)</td>
<td>0.31</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>Score (0 – 3.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Transition question – 5 point scale with responses “Much Worse”, “Somewhat Worse”, “The Same”, “Somewhat Better”, “Much Better”

All correlations p ≤ 0.0001
### TABLE 7.6: DIFFERENCES AND RESPONSIVENESS STATISTICS FROM BASELINE TO 6 MONTHS

**STRATIFYING THE SAMPLE BY THE TRANSITION QUESTION**

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Baseline Mean</th>
<th>SD</th>
<th>Range</th>
<th>Mean Change</th>
<th>SD</th>
<th>Effect Size</th>
<th>95% CI</th>
<th>SRM</th>
<th>95% CI</th>
<th>CSRM</th>
<th>95% CI</th>
<th>Paired test t-RE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HUI3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>222</td>
<td>0.52</td>
<td>0.29</td>
<td>-0.16 to 1.00</td>
<td>0.03</td>
<td>0.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worse</td>
<td>56</td>
<td>0.44</td>
<td>0.3</td>
<td>0.16 to 0.95</td>
<td>-0.03</td>
<td>0.25</td>
<td>-0.10</td>
<td>-0.31 to 0.13</td>
<td>-0.12</td>
<td>-0.56 to 0.08</td>
<td>-0.18</td>
<td>-0.53 to 0.22</td>
<td>0.4</td>
</tr>
<tr>
<td>Same</td>
<td>77</td>
<td>0.6</td>
<td>0.26</td>
<td>-0.05 to 1.00</td>
<td>0.03</td>
<td>0.17</td>
<td>0.12</td>
<td>-0.03 to 0.26</td>
<td>0.18</td>
<td>-0.14 to 0.31</td>
<td>0.18</td>
<td>-0.14 to 0.31</td>
<td>0.13</td>
</tr>
<tr>
<td>Better</td>
<td>89</td>
<td>0.51</td>
<td>0.3</td>
<td>-0.15 to 0.97</td>
<td>0.07</td>
<td>0.24</td>
<td>0.23</td>
<td>0.08 to 0.41</td>
<td>0.29</td>
<td>0.01 to 0.40</td>
<td>0.41</td>
<td>0.13 to 0.74</td>
<td>0.005</td>
</tr>
<tr>
<td><strong>HUI2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>222</td>
<td>0.71</td>
<td>0.19</td>
<td>0.13 to 1.00</td>
<td>0.02</td>
<td>0.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worse</td>
<td>56</td>
<td>0.66</td>
<td>0.21</td>
<td>0.14 to 0.97</td>
<td>-0.03</td>
<td>0.19</td>
<td>-0.14</td>
<td>-0.41 to 0.10</td>
<td>-0.16</td>
<td>-0.39 to 0.16</td>
<td>-0.18</td>
<td>-0.71 to 0.16</td>
<td>0.22</td>
</tr>
<tr>
<td>Same</td>
<td>77</td>
<td>0.76</td>
<td>0.16</td>
<td>0.16 to 1.00</td>
<td>0.03</td>
<td>0.17</td>
<td>0.19</td>
<td>-0.05 to 0.27</td>
<td>0.18</td>
<td>-0.05 to 0.39</td>
<td>0.18</td>
<td>-0.05 to 0.39</td>
<td>0.12</td>
</tr>
<tr>
<td>Better</td>
<td>89</td>
<td>0.7</td>
<td>0.2</td>
<td>0.18 to 1.00</td>
<td>0.06</td>
<td>0.15</td>
<td>0.30</td>
<td>0.16 to 0.47</td>
<td>0.40</td>
<td>0.10 to 0.52</td>
<td>0.35</td>
<td>0.23 to 0.81</td>
<td>0.0001</td>
</tr>
<tr>
<td><strong>EQ-5D</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>222</td>
<td>0.65</td>
<td>0.24</td>
<td>-0.03 to 1.00</td>
<td>-0.001</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worse</td>
<td>56</td>
<td>0.6</td>
<td>0.25</td>
<td>-0.03 to 1.00</td>
<td>-0.04</td>
<td>0.21</td>
<td>-0.16</td>
<td>-0.44 to 0.06</td>
<td>-0.19</td>
<td>-0.66 to -0.02</td>
<td>-0.21</td>
<td>-0.55 to 0.06</td>
<td>0.04</td>
</tr>
<tr>
<td>Same</td>
<td>77</td>
<td>0.72</td>
<td>0.19</td>
<td>0.08 to 1.00</td>
<td>-0.02</td>
<td>0.19</td>
<td>-0.11</td>
<td>-0.36 to 0.16</td>
<td>-0.11</td>
<td>-0.41 to 0.02</td>
<td>-0.11</td>
<td>-0.41 to 0.02</td>
<td>0.48</td>
</tr>
<tr>
<td>Better</td>
<td>89</td>
<td>0.62</td>
<td>0.26</td>
<td>0.02 to 1.00</td>
<td>0.04</td>
<td>0.2</td>
<td>0.15</td>
<td>0.01 to 0.31</td>
<td>0.20</td>
<td>0.12 to 0.59</td>
<td>0.21</td>
<td>-0.01 to 0.45</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>SF-6D</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>222</td>
<td>0.62</td>
<td>0.13</td>
<td>0.31 to 1.00</td>
<td>0.03</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worse</td>
<td>56</td>
<td>0.59</td>
<td>0.13</td>
<td>0.31 to 0.95</td>
<td>-0.01</td>
<td>0.08</td>
<td>-0.08</td>
<td>-0.24 to 0.08</td>
<td>-0.13</td>
<td>-0.44 to 0.15</td>
<td>-0.13</td>
<td>-0.39 to 0.14</td>
<td>0.26</td>
</tr>
<tr>
<td>Same</td>
<td>77</td>
<td>0.64</td>
<td>0.11</td>
<td>0.42 to 0.94</td>
<td>0.04</td>
<td>0.08</td>
<td>0.36</td>
<td>0.19 to 0.56</td>
<td>0.50</td>
<td>0.31 to 0.70</td>
<td>0.50</td>
<td>0.31 to 0.70</td>
<td>0.0001</td>
</tr>
<tr>
<td>Better</td>
<td>89</td>
<td>0.62</td>
<td>0.13</td>
<td>0.31 to 1.00</td>
<td>0.04</td>
<td>0.11</td>
<td>0.31</td>
<td>0.11 to 0.49</td>
<td>0.36</td>
<td>0.16 to 0.58</td>
<td>0.50</td>
<td>0.19 to 0.78</td>
<td>0.0001</td>
</tr>
<tr>
<td><strong>RAQoL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>222</td>
<td>13.1</td>
<td>8.1</td>
<td>0 to 30</td>
<td>-1.26</td>
<td>5.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worse</td>
<td>56</td>
<td>14.9</td>
<td>7.7</td>
<td>0 to 30</td>
<td>1.43</td>
<td>4.25</td>
<td>0.19</td>
<td>0.04 to 0.33</td>
<td>0.34</td>
<td>-0.10 to 0.45</td>
<td>0.38</td>
<td>0.08 to 0.70</td>
<td>0.02</td>
</tr>
<tr>
<td>Same</td>
<td>77</td>
<td>10.8</td>
<td>7.4</td>
<td>0 to 27</td>
<td>-1.23</td>
<td>3.74</td>
<td>-0.17</td>
<td>-0.19 to 0.05</td>
<td>-0.33</td>
<td>-0.39 to 0.07</td>
<td>-0.33</td>
<td>-0.39 to 0.07</td>
<td>0.004</td>
</tr>
<tr>
<td>Better</td>
<td>89</td>
<td>13.8</td>
<td>8.4</td>
<td>0 to 30</td>
<td>-3</td>
<td>5.94</td>
<td>-0.36</td>
<td>-0.51 to -0.20</td>
<td>-0.51</td>
<td>-0.22 to 0.60</td>
<td>-0.80</td>
<td>-0.42 to -1.20</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>HAQ</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>222</td>
<td>1.15</td>
<td>0.77</td>
<td>0 to 2.63</td>
<td>-0.06</td>
<td>0.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worse</td>
<td>56</td>
<td>1.28</td>
<td>0.74</td>
<td>0 to 2.63</td>
<td>0.16</td>
<td>0.48</td>
<td>0.22</td>
<td>0.04 to 0.38</td>
<td>0.33</td>
<td>0.06 to 0.65</td>
<td>0.46</td>
<td>0.08 to 0.84</td>
<td>0.009</td>
</tr>
<tr>
<td>Same</td>
<td>77</td>
<td>0.99</td>
<td>0.79</td>
<td>0 to 2.38</td>
<td>-0.07</td>
<td>0.35</td>
<td>-0.09</td>
<td>-0.28 to 0.02</td>
<td>-0.20</td>
<td>-0.56 to -0.10</td>
<td>-0.20</td>
<td>-0.56 to 0.10</td>
<td>0.09</td>
</tr>
<tr>
<td>Better</td>
<td>89</td>
<td>1.21</td>
<td>0.75</td>
<td>0 to 2.50</td>
<td>-0.18</td>
<td>0.46</td>
<td>-0.24</td>
<td>-0.38 to -0.11</td>
<td>-0.39</td>
<td>-0.69 to -0.30</td>
<td>-0.51</td>
<td>-0.84 to -0.23</td>
<td>0.0001</td>
</tr>
<tr>
<td>TABLE 7.7: DIFFERENCES AND RESPONSIVENESS STATISTICS FROM BASELINE TO 6 MONTHS STRATIFYING THE CATEGORIES CREATED FROM PATIENT GLOBAL ASSESSMENT OF DISEASE SEVERITY VAS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Baseline</td>
<td>SD</td>
<td>Range</td>
<td>Mean Change</td>
<td>SD</td>
<td>Effect Size</td>
<td>95% CI</td>
<td>SRM</td>
<td>95% CI</td>
<td>CSRM</td>
<td>95% CI</td>
<td>Paired t-test</td>
</tr>
<tr>
<td>HUI3</td>
<td>Overall</td>
<td>222</td>
<td>0.52</td>
<td>0.29</td>
<td>-0.16 to 1.00</td>
<td>0.03</td>
<td>0.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Worse</td>
<td>39</td>
<td>0.47</td>
<td>0.28</td>
<td>-0.16 to 0.85</td>
<td>-0.10</td>
<td>0.21</td>
<td>-0.36</td>
<td>-0.04 to -0.65</td>
<td>-0.46</td>
<td>-0.07 to -0.88</td>
<td>-0.53</td>
<td>-0.11 to -0.82</td>
</tr>
<tr>
<td></td>
<td>Same</td>
<td>118</td>
<td>0.63</td>
<td>0.25</td>
<td>-0.07 to 0.97</td>
<td>0.01</td>
<td>0.19</td>
<td>0.05</td>
<td>-0.04 to 0.24</td>
<td>0.07</td>
<td>-0.06 to 0.31</td>
<td>0.07</td>
<td>-0.06 to 0.31</td>
</tr>
<tr>
<td></td>
<td>Better</td>
<td>65</td>
<td>0.37</td>
<td>0.28</td>
<td>-0.15 to 1.00</td>
<td>0.17</td>
<td>0.23</td>
<td>0.60</td>
<td>0.28 to 0.72</td>
<td>0.73</td>
<td>0.29 to 0.80</td>
<td>0.89</td>
<td>0.32 to 1.02</td>
</tr>
<tr>
<td>HUI2</td>
<td>Overall</td>
<td>222</td>
<td>0.71</td>
<td>0.19</td>
<td>0.13 to 1.00</td>
<td>0.02</td>
<td>0.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Worse</td>
<td>39</td>
<td>0.69</td>
<td>0.19</td>
<td>0.19 to 0.94</td>
<td>-0.06</td>
<td>0.14</td>
<td>-0.33</td>
<td>-0.63 to -0.07</td>
<td>-0.44</td>
<td>-0.10 to -0.80</td>
<td>-0.46</td>
<td>-0.13 to -0.89</td>
</tr>
<tr>
<td></td>
<td>Same</td>
<td>118</td>
<td>0.77</td>
<td>0.17</td>
<td>0.18 to 1.00</td>
<td>0.02</td>
<td>0.14</td>
<td>0.10</td>
<td>-0.08 to 0.18</td>
<td>0.13</td>
<td>-0.12 to 0.24</td>
<td>0.13</td>
<td>-0.12 to 0.24</td>
</tr>
<tr>
<td></td>
<td>Better</td>
<td>65</td>
<td>0.63</td>
<td>0.18</td>
<td>0.24 to 0.97</td>
<td>0.09</td>
<td>0.17</td>
<td>0.49</td>
<td>0.39 to 0.83</td>
<td>0.52</td>
<td>0.48 to 1.01</td>
<td>0.64</td>
<td>0.56 to 1.25</td>
</tr>
<tr>
<td>EQ-SD</td>
<td>Overall</td>
<td>222</td>
<td>0.65</td>
<td>0.24</td>
<td>-0.03 to 1.00</td>
<td>-0.001</td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Worse</td>
<td>39</td>
<td>0.65</td>
<td>0.21</td>
<td>0.08 to 1.00</td>
<td>-0.12</td>
<td>0.19</td>
<td>-0.55</td>
<td>-0.16 to -0.52</td>
<td>-0.63</td>
<td>-0.19 to -0.85</td>
<td>-0.69</td>
<td>-0.70 to -0.68</td>
</tr>
<tr>
<td></td>
<td>Same</td>
<td>118</td>
<td>0.73</td>
<td>0.19</td>
<td>0.08 to 1.00</td>
<td>-0.02</td>
<td>0.17</td>
<td>-0.09</td>
<td>0.17 to 0.01</td>
<td>-0.10</td>
<td>-0.34 to 0.01</td>
<td>-0.10</td>
<td>-0.12 to -0.10</td>
</tr>
<tr>
<td></td>
<td>Better</td>
<td>65</td>
<td>0.53</td>
<td>0.21</td>
<td>0.02 to 0.88</td>
<td>0.09</td>
<td>0.22</td>
<td>0.36</td>
<td>0.16 to 0.52</td>
<td>0.43</td>
<td>0.29 to 0.73</td>
<td>0.56</td>
<td>0.55 to 0.57</td>
</tr>
<tr>
<td>SF-6D</td>
<td>Overall</td>
<td>222</td>
<td>0.62</td>
<td>0.13</td>
<td>0.31 to 1.00</td>
<td>0.03</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Worse</td>
<td>39</td>
<td>0.60</td>
<td>0.11</td>
<td>0.37 to 0.95</td>
<td>-0.03</td>
<td>0.07</td>
<td>-0.24</td>
<td>-0.02 to -0.49</td>
<td>-0.35</td>
<td>-0.04 to -0.87</td>
<td>-0.30</td>
<td>-0.03 to -0.64</td>
</tr>
<tr>
<td></td>
<td>Same</td>
<td>118</td>
<td>0.66</td>
<td>0.13</td>
<td>0.37 to 1.00</td>
<td>0.02</td>
<td>0.09</td>
<td>0.18</td>
<td>0.05 to 0.31</td>
<td>0.26</td>
<td>0.09 to 0.45</td>
<td>0.26</td>
<td>0.09 to 0.45</td>
</tr>
<tr>
<td></td>
<td>Better</td>
<td>65</td>
<td>0.57</td>
<td>0.11</td>
<td>0.31 to 0.89</td>
<td>0.06</td>
<td>0.09</td>
<td>0.54</td>
<td>0.32 to 0.79</td>
<td>0.62</td>
<td>0.41 to 0.85</td>
<td>0.68</td>
<td>0.39 to 1.03</td>
</tr>
<tr>
<td>RAQoL</td>
<td>Overall</td>
<td>222</td>
<td>13.1</td>
<td>8.1</td>
<td>0 to 30</td>
<td>-1.26</td>
<td>5.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Worse</td>
<td>39</td>
<td>13.2</td>
<td>8.2</td>
<td>1 to 29</td>
<td>2.72</td>
<td>4.87</td>
<td>0.33</td>
<td>0.21 to 0.83</td>
<td>0.56</td>
<td>0.20 to 0.67</td>
<td>0.71</td>
<td>0.30 to 0.97</td>
</tr>
<tr>
<td></td>
<td>Same</td>
<td>118</td>
<td>10.4</td>
<td>7.3</td>
<td>0 to 29</td>
<td>-1.03</td>
<td>3.85</td>
<td>-0.14</td>
<td>-0.08 to -0.28</td>
<td>-0.27</td>
<td>-0.08 to -0.29</td>
<td>-0.27</td>
<td>-0.08 to -0.29</td>
</tr>
<tr>
<td></td>
<td>Better</td>
<td>65</td>
<td>17.1</td>
<td>7.7</td>
<td>0 to 30</td>
<td>-4.29</td>
<td>6.20</td>
<td>-0.56</td>
<td>-0.18 to -0.75</td>
<td>-0.69</td>
<td>-0.27 to -1.08</td>
<td>-1.11</td>
<td>-0.23 to -1.59</td>
</tr>
<tr>
<td>HAQ</td>
<td>Overall</td>
<td>222</td>
<td>1.15</td>
<td>0.77</td>
<td>0 to 2.63</td>
<td>-0.06</td>
<td>0.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Worse</td>
<td>39</td>
<td>1.06</td>
<td>0.71</td>
<td>0 to 2.63</td>
<td>0.25</td>
<td>0.46</td>
<td>0.34</td>
<td>0.11 to 0.44</td>
<td>0.50</td>
<td>0.28 to 0.88</td>
<td>0.63</td>
<td>0.26 to 0.92</td>
</tr>
<tr>
<td></td>
<td>Same</td>
<td>118</td>
<td>0.97</td>
<td>0.77</td>
<td>0 to 2.63</td>
<td>-0.07</td>
<td>0.39</td>
<td>-0.08</td>
<td>-0.06 to -0.25</td>
<td>-0.17</td>
<td>-0.12 to -0.46</td>
<td>-0.17</td>
<td>-0.12 to -0.46</td>
</tr>
<tr>
<td></td>
<td>Better</td>
<td>65</td>
<td>1.53</td>
<td>0.63</td>
<td>0 to 2.50</td>
<td>-0.22</td>
<td>0.43</td>
<td>-0.35</td>
<td>-0.32 to -0.76</td>
<td>-0.50</td>
<td>-0.48 to -0.92</td>
<td>-0.57</td>
<td>-0.44 to -1.24</td>
</tr>
</tbody>
</table>
TABLE 7.8: RANKINGS OF RESPONSIVENESS OF MEASURES ACCORDING TO THE RESPONSIVENESS STATISTIC AND THE EXTERNAL CRITERIA OF CHANGE (EITHER RESPONSES TO THE PATIENT TRANSITION QUESTION OR TO THE PATIENT GLOBAL ASSESSMENT OF DISEASE ACTIVITY VAS)

<table>
<thead>
<tr>
<th>Measure</th>
<th>ES</th>
<th>SRM</th>
<th>CSRM</th>
<th>Paired t-test</th>
<th>RE</th>
<th>Median Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUI2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>HUI3</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>SF-6D</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>RAQoL</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>HAQ</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure</th>
<th>ES</th>
<th>SRM</th>
<th>CSRM</th>
<th>Paired t-test</th>
<th>RE</th>
<th>Median Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUI2</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>HUI3</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>SF-6D</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>RAQoL</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HAQ</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

ES = Effect Size; SRM = Standardized Response Mean; CSRM = Control Standardized Response Mean; RE = Relative efficiency; Ties in rankings are possible.
TABLE 7.9: ASSOCIATIONS BETWEEN INSTRUMENT UNWEIGHTED DOMAINS / SINGLE ATTRIBUTE SCORE CHANGES AND SELF-REPORTED CHANGE FROM 0 TO 6 MONTHS

<table>
<thead>
<tr>
<th>Instrument Changes in Domain/SA</th>
<th>Patient Transition Question</th>
<th>P-value</th>
<th>Patient Global Assessment of Disease Severity VAS Categories</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EQ-5D</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usual activities</td>
<td></td>
<td>0.41</td>
<td></td>
<td>0.22</td>
</tr>
<tr>
<td>Pain/discomfort</td>
<td></td>
<td>0.03</td>
<td></td>
<td>0.003</td>
</tr>
<tr>
<td>Mobility</td>
<td></td>
<td>0.19</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Anxiety/Depression</td>
<td></td>
<td>0.05</td>
<td></td>
<td>0.002</td>
</tr>
<tr>
<td>Self-care</td>
<td></td>
<td>0.003</td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>SF-6D</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical functioning</td>
<td></td>
<td>0.003</td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Role limitations</td>
<td></td>
<td>0.001</td>
<td></td>
<td>0.008</td>
</tr>
<tr>
<td>Social functioning</td>
<td></td>
<td>&lt;0.001</td>
<td></td>
<td>0.002</td>
</tr>
<tr>
<td>Pain</td>
<td></td>
<td>0.003</td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Mental Health</td>
<td></td>
<td>0.08</td>
<td></td>
<td>0.19</td>
</tr>
<tr>
<td>Vitality</td>
<td></td>
<td>0.08</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td><strong>HUI3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vision</td>
<td></td>
<td>0.92</td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td>Hearing</td>
<td></td>
<td>0.77</td>
<td></td>
<td>0.50</td>
</tr>
<tr>
<td>Speech</td>
<td></td>
<td>0.84</td>
<td></td>
<td>0.97</td>
</tr>
<tr>
<td>Ambulation</td>
<td></td>
<td>0.01</td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Dexterity</td>
<td></td>
<td>0.41</td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Emotion</td>
<td></td>
<td>0.003</td>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td>Cognition</td>
<td></td>
<td>0.63</td>
<td></td>
<td>0.58</td>
</tr>
<tr>
<td>Pain</td>
<td></td>
<td>0.007</td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>HUI2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensation</td>
<td></td>
<td>0.99</td>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>Mobility</td>
<td></td>
<td>0.02</td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Emotion</td>
<td></td>
<td>0.004</td>
<td></td>
<td>0.004</td>
</tr>
<tr>
<td>Cognition</td>
<td></td>
<td>0.36</td>
<td></td>
<td>0.99</td>
</tr>
<tr>
<td>Self-care</td>
<td></td>
<td>0.009</td>
<td></td>
<td>0.23</td>
</tr>
<tr>
<td>Pain</td>
<td></td>
<td>0.003</td>
<td></td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

258
FIGURE 7.1: AGREEMENT BETWEEN THE PATIENT TRANSITION QUESTION AND CHANGES USING MID CUTOFFS FOR THE GENERIC AND DISEASE-SPECIFIC INSTRUMENTS

HUI2 differences between 6 months and baseline

Weighted Kappa = 0.24 (95% CI 0.13 - 0.35)

HUI3 differences between 6 months and baseline

Weighted Kappa = 0.17 (95% CI 0.06 - 0.28)
FIGURE 7.1: AGREEMENT BETWEEN THE PATIENT TRANSITION QUESTION AND CHANGES USING MID CUTOFFS FOR THE GENERIC AND DISEASE-SPECIFIC INSTRUMENTS

HUI2 differences between 6 months and baseline
Weighted Kappa = 0.24 (95% CI 0.13 - 0.35)

HUI3 differences between 6 months and baseline
Weighted Kappa = 0.17 (95% CI 0.06 - 0.28)
EQ-5D difference between 6 months and baseline

Weighted Kappa = 0.19 (0.09 - 0.28)

SF-6D differences between 6 months and baseline

Weighted Kappa = 0.15 (95% CI 0.04 - 0.25)
RAQoL difference between 6 months and baseline

Weighted Kappa = 0.28 (95% CI 0.17 - 0.38)

HAQ difference between 6 months and baseline

Weighted Kappa = 0.23 (95% CI 0.12 - 0.33)
FIGURE 7.2: SCATTERPLOT OF HUI2 UTILITY SCORES OVER TIME STRATIFIED BY THE RESULTS OF THE COLLAPSED TRANSITION QUESTION

The lines represent a least squares regression fit to characterize the trend for change over time by category.
FIGURE 7.3: SCATTERPLOT OF HUI3 UTILITY SCORES OVER TIME
STRATIFIED BY THE RESULTS OF THE COLLAPSED
TRANSITION QUESTION

The lines represent least squares regression fit to characterize the trend for change by
category.
FIGURE 7.4: SCATTERPLOT OF EQ-5D UTILITY SCORES OVER TIME STRATIFIED BY THE RESULTS OF THE COLLAPSED TRANSITION QUESTION

The lines represent least squares regression fit to characterize the trend for change by category.
FIGURE 7.5: SCATTERPLOT OF SF-6D UTILITY SCORES OVER TIME STRATIFIED BY THE RESULTS OF THE COLLAPSED TRANSITION QUESTION

The lines represent least squares regression fit to characterize the trend for change by category.

HUI2: Change from 0 to 6 months

![Diagram showing the relationship between change in HUI2 and transition question response probabilities.](image-url)

EQ-5D - Change from 0 to 6 Months

Estimated probability of a response on the transition question

Worse

Same

Better

Change in EQ-5D score: 6 month minus baseline score

0.0 0.2 0.4 0.6 0.8 1.0

-0.6 -0.4 -0.2 0.0 0.2 0.4 0.6

SF-6D - Change from 0 to 6 months

Change in SF-6D: 6 mos. minus baseline scores

RAQoL: Change from 0 to 6 months

Estimated probability of a response on the transition question

Change in RAQoL score: 6 mos. minus baseline score

HAQ - Change from 0 to 6 Months

Estimated probability of a response on the transition question

Change in HAQ Score: 6 mos. minus baseline score

n = 226 df = 3

Change in HUI3: 0 to 6 months

Estimated probability of a response according to the patient global VAS

Change in HUI3: 6 month minus baseline score
FIGURE 7.14: RESULTS OF THE MULTI-RESPONSE MODEL OF THE
ASSOCIATION BETWEEN A CHANGE IN THE EQ-5D AND THE
PATIENT GLOBAL ASSESSMENT OF DISEASE ACTIVITY

EQ-5D: Change from 0 to 6 months

Estimated probability of a response according to the patient global VAS

Worse

Same

Better

Change in EQ-5D: 6 months minus baseline score
FIGURE 7.15: RESULTS OF THE MULTI-RESPONSE MODEL OF THE ASSOCIATION BETWEEN A CHANGE IN THE SF-6D AND THE PATIENT GLOBAL ASSESSMENT OF DISEASE ACTIVITY

SF-6D: Change from 0 to 6 months

HAQ: Change from 0 to 6 months

Estimated probability of a response according to the patient global VAS

Better

Same

Worse

Change in HAQ: 6 months minus baseline score

-1.0 -0.5 0.0 0.5 1.0

0.0 0.2 0.4 0.6 0.8 1.0

0.0 0.2 0.4 0.6 0.8 1.0

0.0 0.2 0.4 0.6 0.8 1.0
CHAPTER 8

GENERAL DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

8.1 SUMMARY OF STUDY FINDINGS

The results of this study provided significant insights into the limits of interchangeability of indirect utility assessment instruments in the assessment of HRQL in patients with rheumatoid arthritis (RA). In addition, it provided evidence that although each of these instruments displayed cross-sectional, construct validity in RA, they yielded quite different results and hence could lead to different policy decisions when utilized in an economic evaluation for RA interventions. As well, annual household income appeared to impact on the scores achieved with these instruments which is inconsistent with the assumptions underlying utility measurements. Finally, for longitudinal construct validity, as expected and consistent with previous research findings, the disease-specific measure (the RAQoL) performed the best. However, the HUI3 and SF-6D appeared to be suitable for measuring RA improvement, whereas the EQ-5D was the most responsive of the preference-based measures in terms of detecting worsening.

In terms of the interchangeability of the scores from the instruments, it was found that, on a cross-sectional basis, mean scores were significantly different. In addition, when agreement was analyzed, the scores from the instruments had low to moderate agreement and agreement tended to be better at the higher end of the utility scores (towards perfect health) and tended to be much worse at the lower end (as scores approached states judged to be near or worse than death). In comparison to the other instruments, it appeared that the EQ-5D
tended to suffer from limitations previously described in the literature – namely gaps in its distribution especially at the mid-point of its scale.\textsuperscript{2,3} These differences are likely due to the various domains, levels of choice within each domain, scoring algorithms, and the valuation techniques that are intrinsic to each system. However, despite the differences in scores, in our RA sample, each instrument appeared to be measuring mostly physical functioning and pain. However, the HUI2 and HUI3 systems were also assessing cognition while the EQ-5D and SF-6D were measuring emotional/mental health.

In terms of construct validity in RA, the preference-based instruments all appeared to perform as hypothesized and almost as well as the RAQoL and the HAQ. All of the scores from the indirect utility assessment instruments were lower in naturally occurring dichotomous groups (such as use of equipment for RA in the past twelve months – “yes” or “no”) thought to have higher RA severity. In addition, there was an observed gradient of scores in the hypothesized direction across self-reported disease-severity and control with moderate to strong correlations. Of note, although the measures of association tended to be slightly higher for the RAQoL and the HAQ, the preference-based measures performed almost as well in their ability to discriminate between groups of known RA severity or symptoms. The overall scores and relevant attributes (for example, mobility, self-care and pain within the HUI2) within each of the instruments performed as expected with moderate to large, significant correlations with RA clinical variables. In addition, weak associations were found between attributes not thought to be directly affected by RA (such as sensation, vision, and hearing) and RA clinical variables. Minimally important differences in the instrument scores were found to be somewhat higher than had been hypothesized or observed.
in other studies of the HUI systems,\textsuperscript{4,5} and approximately the same for the SF-6D,\textsuperscript{6} EQ-5D and the HAQ.\textsuperscript{7} Other investigators have noted that differences observed in the mean and individual scores across the instruments might result in radically different estimations of QALYs and, thus, outcomes of economic evaluations.\textsuperscript{8-10} When the different utility scores were applied to a decision-analytic, Markov-based, cost-effectiveness analysis of infliximab plus MTX versus MTX alone, results presented in Chapter 5 substantiated these findings and hypotheses. For example, between the lowest (using HUI3-derived QALYs) and highest (using SF-6D derived QALYs) of the incremental cost-effectiveness ratios generated through the use of scores from the HUI2, HUI3, EQ-5D and SF-6D as QALY weights, there was a difference of over 100\% ($33,092 per QALY, 95\% CI $30,887 to $35,436 for the HUI3 vs. $67,005 per QALY, 95\% CI $62,773 to $71,540 for the SF-6D). Considering that the ceiling ratio commonly cited for policy-makers is $50,000 per QALY,\textsuperscript{11} different funding decisions could be made depending on which indirect utility assessment instrument was utilized.

The impact on the different utility instruments on cost-effectiveness analysis is not limited to their differences in scores. As Sculpher and O'Brien wrote, scores on preference-weighted measures, if truly reflecting society's preferences for health states, should not be influenced by income.\textsuperscript{12} However, there is evidence that scores on other generic, nonpreferenced-based measures vary by income and other measures of socioeconomic status in RA.\textsuperscript{13} Results presented in Chapter 6 clearly show that prior to adjustment for RA severity, each of the scores from any of the HRQL instruments (the RAQoL, HAQ, and the indirect utility instruments) had a significant gradient within categories of income (Table 6.5). After adjustment for RA severity which could be inherently biasing these finding
(sicker people have more severe RA and are poorer) and number of people in the household, there still was a clear gradient across income for all of the indirect utility assessment instrument scores (however, only the HUI3 and SF-6D scores were still significantly associated with income in our sample). Of note, the HAQ also demonstrated this relationship whereas the RAQoL did not.

Other crucial properties of any HRQL instrument are reliability and longitudinal construct validity (which may also be referred to as responsiveness). Little work has been done in this regard on any of the indirect, utility assessment instruments (both within and outside of RA) and for the RAQoL. In addition, although patient transition questions (where patients have rated changes in their health retrospectively over a specified time period as “better”, “worse” or “the same”) have been utilized as an external criterion of change in responsiveness studies, there is little information about the reliability and validity of this technique in RA. As documented in Tables 7.2 and 7.3, we found that all of the indirect assessment instruments were reliable, with exception of the EQ-5D. This finding is both in agreement and in contrast with previous studies conducted in RA. Consistently, using various methodologies and external criteria for change, the RAQoL was determined to be the most responsive out of the tested instruments. The indirect utility instruments were poorly to moderately responsive with the SF-6D and the HUI3 being the most responsive for RA improvement and the EQ-5D being the most responsive for RA worsening. The HUI2 was the least responsive to changes in RA and the performance of the HAQ was dependent on the external criterion of change that was applied. In terms of using the patient transition question as the external criterion of change, it was found to be moderately correlated with other RA outcome measures but poorly associated with changes in the HRQL measures. Using
categories of the patient global assessment of disease severity question appeared to yield much more valid results.

8.2 UNIQUE CONTRIBUTIONS, IMPACT, AND IMPLICATIONS

The manuscripts comprising this dissertation contribute significantly to the current body of literature both methodologically and through the findings as they relate to disease specific HRQL and generic measures that often are utilized as QALY weightings in the cost-utility framework. Thus, the findings in this study have significant implications both in the assessment of HRQL and in the application of the indirect utility assessments in economic evaluations of interventions for RA.

In terms of the feasibility of these instruments in RA, all of the instruments tested appear to have been accepted by the RA patients and were almost entirely successfully completed. In addition, they all appeared to be appropriate measures for RA in that they were able to discriminate between levels of disease severity and were strongly correlated with RA clinical measures. Therefore, in terms of these properties for the generic measures, there was little evidence that one should be chosen over the other to assess HRQL in RA. As a disease-specific measure, the RAQoL was slightly better than the generic measures in terms of construct validity and thus should be a preferred measure for the assessment of RA specific HRQL.

In terms of the application of their scores as QALY weights to be used in economic evaluation, we found a large impact on the incremental cost-effectiveness ratio by utilizing the different indirect utility assessment instrument scores. Considering the lack of advice on QALY weighting standardization either within the rheumatology research community or
from economic evaluation guidelines,21 these results are extremely useful in illustrating the magnitude of this problem. As such, these finding should guide either the standardization of QALY weighting measures within economic evaluations for RA or make it implicit that sensitivity analyses that covers the wide range of utility values that are possible with these instruments are performed. In addition, these results will be useful for informing economic evaluations occurring outside of RA and rheumatological conditions as this issue is likely also relevant in different disease states. Thus, an effort should be made for QALY weighting standardization both within and across disease states, especially if comparisons are to be made between economic evaluations while allocating limited health care resources.

The finding that annual household income was associated with scores on the preference-based measures was problematic as utility scores should be neutral with respect to income.12 As such, this research provides empiric evidence for the concern raised by Sculpher and O’Brien12 and will serve as an impetus for future work in this area to address this limitation of the current instruments or prevent double-counting in economic evaluations. These results also show that generic quality of life, despite RA severity, was lower in those who are poorer which was similar to other investigators’ findings.13 Since self-reported health (including HRQL) is strongly and independently associated with morbidity and mortality,22 there should be new focuses for interventions in RA that do not focus solely on clinical or severity measures. Another benefit of this research was the discovery that population-based, census-level or post-code based variables (neighbourhood median income, percentage with baccalaureate degrees and neighbourhood unemployment) are poorly correlated with RA patient self-reported socioeconomic data. As such, this
research can inform other studies conducted in this area and prevent erroneous results based upon incorrect assumptions.

With respect to responsiveness, our results showed that not only does the RAQoL have cross-sectional construct validity but it also was reliable and responsive to change. Thus, the RAQoL appeared to be a feasible, reliable, valid, and responsive instrument — all properties necessary for an excellent tool.¹⁴ There were no reasons why this new measure should not be integrated as an outcome measure in the assessment of RA in both clinical trials and in observational studies. With respect to the indirect utility assessment instruments, the responsiveness analysis provided guidance about which measures should be adopted in certain situations. For example, in the application of a known, effective intervention, the SF-6D or HUI3 would be suitable choices considering their cross-sectional and longitudinal construct validity (for the latter, their ability to detect change is superior in measuring improvement in RA rather than worsening). The EQ-5D may be better suited when a preference-based measure is needed to detect worsening. The HUI2 appeared to relatively unresponsive in patients with RA.

8.3 STUDY STRENGTHS AND LIMITATIONS

8.3.1 Strengths

This study has a number of specific strengths that enhance the credibility of the results. First of all, it was conducted in a relatively large sample of patients in a homogeneous, well-delineated disease group, RA. This fact was important as there should not be differential response between the instruments if some domains were more valid or responsive within different diseases. In addition, although recruitment occurred within one
province, due to the cosmopolitan nature of British Columbia, many of the participants came from racially diverse backgrounds, thus increasing the applicability of our results.

Another advantage with our methodology was the inclusion and comparison of the four most popular indirect utility assessment instruments (the HUI2, HUI3, EQ-5D and SF-6D) as well as a disease-specific (RAQoL) and a well-established disability measure (HAQ). Thus, we were able to perform between instrument comparisons within the indirect utility assessment classification and also compare their individual and collective performances relative to disease-specific measures. These additional data generated are improvements over many of the comparative studies which have only examined two or three of the instruments leaving many questions unanswered (see Chapter 2).

Due to our prospective sampling and questionnaire methodology, we were able to directly ask questions to the RA patients that could not otherwise have been assessed through administrative data or retrospective chart reviews. As such, we could assess construct validity based upon other variables (such as outlined in Table 4.6) rather than limiting our analysis to commonly collected clinical data. In addition, we could ascertain many sociodemographic variables including annual household income directly from respondents rather than relying on census level data, which, as we have shown, could have lead to inaccurate results (Chapter 6).

For the cost-effectiveness analysis of infliximab (Chapter 5), we were fortunate to have access to some very large and comprehensive datasets that increased the validity of the results. For example, we modeled all-cause mortality based upon more than 1900 consecutive RA patients followed since 1974 at the Wichita (Kansas) Arthritis Center, an outpatient rheumatology clinic.23,24 We also had access to the dataset which formed the basis
of the only comprehensive, Canadian RA cost-of-illness study that has been published to date.\textsuperscript{25} Resource utilization and cost data were collected during a longitudinal study of 1063 Canadian RA patients who reported semi-annually on their health services utilization over the preceding 6 months during 1983 and 1994.\textsuperscript{25} These direct costs of RA care were comprehensive and were comprised of long-term care, rehabilitation, nursing homes, health professional visits, medications, diagnostic tests, acute hospitalization, emergency department visits, ambulance services, dialysis and outpatient surgeries. Finally, to estimate indirect costs due to productivity losses, we had access to data from a prospective longitudinal sample of 120 employed RA patients recruited in Ontario, Canada from September 1999 to December 2001.\textsuperscript{26} In the self-report questionnaire, participants were asked the number of days missed due to RA in the past 6 months and their regular weekly working hours.

We were also able to divide the HAQ into smaller discrete health states for the state-transition model than have been previously used by investigators utilizing HAQ-based Markov models.\textsuperscript{27,28} The problem with dividing the HAQ scores into four to six discrete health states is that much smaller changes in this instrument have been shown to be predictive of large changes in resource utilization,\textsuperscript{29} work productivity\textsuperscript{30} and mortality.\textsuperscript{24} Thus, in these models, the loss of transitions between smaller changes in HAQ resulted in a less accurate assessment of costs and outcomes. As such, the fact that we were able to create a 25 state transition model (based on an underlying continuous process) with costs and outcomes associated with each of the discrete states, allowed for more accurate modeling of the progression of the disease and prediction of economic outcomes.
In the reliability and responsiveness analyses, we were able to assess these properties using more than one method. For reliability assessment, we used two versions of test-retest methodology which produced very similar intraclass correlation coefficients adding to the consistency of our results. In addition, since it has been demonstrated that the various responsiveness statistics often give different results, we used multiple methods to assess this property and examined the consistency of the results generated by these analyses. Besides using commonly recommended responsiveness statistics (ES, SRM, CSRM, paired t-tests, and RE), we also utilized a new graphical method developed by Abrahamowicz and colleagues. This new method is potentially advantageous for the clinical translation of responsiveness as clinicians can examine the probability of true improvement (as assessed by the external criterion) based on a given change in the HRQL score. Therefore, our recommendations for responsiveness were based upon multiple criteria and are likely considerably more robust than if only one methodology was utilized. The rankings across the different statistics, were, for the most part, in agreement and thus strengthened our conclusions. In addition, we also utilized two external criteria for change – the patient transition question and a categorization of the patient global assessment of disease severity visual analogue scale. Therefore, we examined consistency across results achieved with both external criteria in our assessment of responsiveness.

### 8.3.2 Limitations

As with any study, this one was not without its limitations, none of which significantly affected the findings. In each of the stand-alone manuscripts, limitations have been stated but, nonetheless, they merit review.
With respect to the questionnaire process, there were a few potential threats to the internal validity of the study. For example, since this was a natural history study without a known, effective intervention or a control group, people's HRQL was assumed to change (either positively or negatively) or stay the same over time due to their RA. However, this assumption was only true if all extraneous variables that might affect health status could be held constant over time. Of course, at the individual level, events may have occurred that influenced their HRQL that were independent of their RA (i.e. development of or worsening of co-morbidities, personal problems). However, on a group level, it would be unlikely for these events to influence the changes in health status captured by the HRQL instruments.

It was possible that the methods used to administer the questionnaire may have influenced the results to some extent. The ordering of the HRQL instruments measures was not randomized because of the self-completed nature of the study and due to the observation, during piloting, that respondents were completing the questionnaires in different orders rather than how they were presented. Also, because respondents were permitted to take the questionnaire home with them (or, were mailed the questionnaire at home) and were instructed to mail it, once completed, back to the investigator, it was possible that some of the respondents answered the questions over several days. Thus, it was conceivable that some patients filled out one HRQL measure on a given day when their health status was different from when they completed other HRQL measures. For instruments such as the EQ-5D with an immediate recall period, this could result in potentially biased results during cross-sectional comparisons with the other instruments.

Another area of concern with longitudinal studies is that the experience of people lost to follow-up may be systematically different than those who decide to stay 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 (Table 7.1). Patient characteristics and HRQL scores between the groups were generally very similar. The few statistically significant differences between the two groups tended to favour the group that completed only the baseline questionnaires as they were found to have significantly better mean baseline HAQ scores, lower self-reported RA severity, and a higher percentage were working outside the home. The retention rate for both those who completed both the baseline and 6 month assessment was 239 out of the 320 (75%). 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 problem with any study assessing the responsiveness of HRQL instruments has to do with the definitions of this property that exist in the literature.\textsuperscript{15,16} For the most part, consensus is lacking on what constitutes a responsive measure, and how responsiveness should be quantified. In addition, attempts to further refine and define this property of HRQL measures have led to the creation of new and, sometimes conflicting, definitions.\textsuperscript{15,16,37}

Also, with repeated administration of the questionnaires over time, there was the possibility that respondents may have undergone what is deemed “response shift”.\textsuperscript{34} This process involves a change in the calibration or self-valuation of a particular health state over time. This finding (the adaptation to long-term conditions leading to bias in self-assessments of HRQL) has recently been illustrated.\textsuperscript{35,36} Another competing theory has been deemed “implicit theory of change”.\textsuperscript{37} In this theory, patients who respond to a questionnaire regarding health where there is retrospective recall (i.e. “Overall, how would you describe
changes in your rheumatoid arthritis since answering the FIRST questionnaire? (i.e. about six months ago”) reconstruct the previous state such that the prospective judgment is more valid. For example, patients begin with their perceived present health state and then infer, based upon impression rather than a conscious comparison between two time points, what their health state must have been. Either of these conditions would have been a threat to validity in our study. These theories are especially applicable when one considers that one of our external criteria, the patient transition question, was based upon recall of a health state from baseline.

Because no ‘gold standard’ exists for establishing change in HRQL in RA, two external anchor-based criteria were used to evaluate responsiveness. One of which was the patient transition question, a 5-point Likert scale evaluating how the respondent’s RA had changed since the last questionnaire. However, there has been some criticism about this approach as recall of a previous state in patients with a chronic disease is subject to recall bias (implicit change theory) or could be influenced by response shift, as described above.38 However, in defense of this criterion of change, since this question was asked only at the follow-up and not at the baseline, error associated with its measurement will only occur once and would not be compounded with baseline measurement error.

The other anchor based criterion was a categorization of the patient global assessment of disease activity VAS. The cutoff values applied in this study were derived from the definitions of positive change associated with effective treatment as defined by the American College of Rheumatology criteria,39 but they may not be the optimal cutoff points for change in this natural history RA sample. In addition, the definition for “worse” associated with this
criterion (≥ 20% reduction in the VAS where “0” is anchored at “Very Poor” and “1” is anchored at “Very Well”) has not been widely used.

Comparisons of difference scores for patients whose RA was defined as “worse”, “same” or “better” across the two anchors assisted in distinguishing between problems that were associated with the external anchors of change (i.e. misclassification error). The direction and magnitude of mean difference scores of patients using the different external anchors of change provided insights into the appropriateness of these external anchors. The categorization of the patient global assessment of disease activity criterion demonstrated all 3 of the following characteristics that were intuitively desirable (using literature values for the MID): (1) positive mean change scores that equal or surpass the MID for a summary score based on patients classified as “better”; (2) negative mean change scores that equal or surpass the MID for a summary score in patients classified as “worse”; (3) change scores, either positive or negative, that were less than the MID for patients classified as “same” (Table 7.7). For the patient transition question, these characteristics did not apply, and there were some instances where the group classified as “same” had a greater mean difference score than the groups classified as either “worse” or “better” (Table 7.6). For example, changes in the “same” group for the HUI3, and HUI3 were of equal magnitude (although in the opposite direction) to those classified as “worse”. For the SF-6D, those classified as “same” had changes equal and in the same direction as those classified as “better”. This finding was especially problematic as this change was significant (p<0.0001). The RAQoL also had a significant amount of change (p=0.004) in the same direction (although not of equal magnitude) as those classified as better. For these and other reasons, the patient global assessment of disease activity categories was deemed to be a superior external criterion.
The choice of time periods over which to evaluate responsiveness was based upon the investigators estimates of the likelihood of change over a given time period. As such, to allow for the greatest opportunity for potential change in RA, responsiveness was evaluated between baseline and 6 months. However, further examination of responsiveness could occur between other time periods such as between baseline and 3 months, or even 3 and 6 months. While these analyses might capture smaller meaningful change, we believe that such a short time frame (3 months) might not be adequate for most patients to experience true changes in their disease in a natural history study of RA. Finally, although we conducted analyses to determine which of the domains/single attributes of the indirect utility assessment instruments were associated with changes as defined by both external criteria, we did not specifically evaluate the responsiveness of each of these domains/attributes.

In Chapter 5, as in all decision-analytic studies, there were some limitations and important assumptions that could influence the interpretation of the results. First of all, as the transition probability matrices were estimated from a randomized, controlled trial, the results of the analysis really reflected cost-efficacy and not cost-effectiveness. Wolfe et al. make this important distinction in a recent editorial pointing out that outcomes achieved with drug therapy in RA clinical trials are often quite different than observed in clinical practice. Secondly, we have assumed that the impact on HAQ of infliximab would impart a survival benefit similar to that observed with MTX. While this assumption is likely reasonable, it will be many years before this finding can be substantiated in an observational study. Also, although we estimated direct costs from a large Canadian sample followed longitudinally for a number of years, there are some limitations with these data. They were drawn from the province of Saskatchewan which has few rheumatologists and thus the costs of care may be
conservative. In addition, these costs come from an era when other new drug therapies (such as the COX-2 specific inhibitors) and other drug strategies (such as combination DMARD therapy) were not yet available or utilized thus further underestimating the costs likely to be incurred today. From the productivity database, in order to integrate the costs into the HAQ-based model, we had to find a relationship between scores on the HAQ and ability to work. Unfortunately, the entire HAQ questionnaire was not collected in this study and only certain HAQ questions and elements from the Multidimensional Functional Assessment Questionnaire were available. From these, we constructed a disability index that served as a proxy for the HAQ. On further testing in the prospectively collected sample, we found that there was high correlation between this proxy score and the actual HAQ.

Furthermore, we simulated 100,000 individual patients' results (50,000 for each treatment strategy) which may have reduced the confidence limits around the cost-effectiveness ratio through a reduction in variability. Therefore, with fewer simulated patients, the confidence intervals may have been wider with a larger degree of overlap with the results achieved by the different indirect utility assessment instruments. Finally, we used a ceiling ratio of $50,000 per QALY as the threshold for what decision-makers might consider to be a fundable program. However, despite widespread use of this ceiling value in the literature, there is evidence that this value might be too low. 

With respect to the paper outlining the effect of income on the scores achieved with the HRQL instruments (Chapter 6), there were a few limitations. Annual income data was collected in increments of $10,000 (starting with the first category of <$20,000) up to a maximum of $100,000 (the last category was >$100,000) and then further categorized to <$20,000, $20,000 to $50,000, and >$50,000. Some people declined to answer this question
resulting in 19% (n=58) missing values for this variable. It is possible that the income (and other characteristics) of these individuals was systematically different than those who answered the question resulting in a bias. However, when we compared all other available characteristics between these two groups, no differences were detected thus giving some confidence that no bias was introduced.

Collapsing a variable, as was done with income, can sometimes lead to a loss of information. However, when measuring socioeconomic status, rather than continuous income, definitions utilized in other studies have often been based on similar categories. Income, as an indicator of SES, is almost always classified depending on the number of subjects. For example in large analyses (such as the Black report and the National Population Health Survey analyses in Canada), income has been categorized. This categorization aids in interpretability as, instead of finding a linear relationship between continuous income and HRQL score, categorizing income allowed us to classify people as being of low, medium or high income and to calculate the gradient in HRQL scores associated with this classification. Also, with our sample's size, it was helpful to classify, as the classification decreased the noise. There is much greater noise by misclassification of a person when using $10K increments (i.e. a person can report their income as $40K/yr when, in fact, it's closer to $30K/yr. By classifying this way, one is more likely to classify an individual into the correct group even if there is some error in the reporting. This may not be the case if we were utilizing a more objective source of data (i.e. income tax reports) but is most likely with self-report information due to recall bias.

In the relative income hypothesis, health is related to one's income (or social status) relative to others in their population, not to their absolute income. This hypothesis facilitates
the division of the sample into categories (such as low/intermediate/high) of social status based upon income. Population-based studies (with very large sample sizes) inherently use quintiles or deciles. Since this study was not population-based and is based upon a relatively small sample (when compared to the population), we were restricted to categorizing our data into larger categories from which one can draw statistical inference. Finally, these categories of income have been applied successfully in another sample.\textsuperscript{46}

For many of the analyses in our study, we have utilized parametric statistics. However, there is some evidence that the scores obtained on HRQL instruments can be interpreted as ordinal. There are numerous publications that support the analysis of these types of data using conventional, parametric statistical techniques provided that some criteria are met. Particularly, in a recent publication, the authors have nicely demonstrated that, when an ordinal health-related quality of life scale has more than seven categories, it can be treated as continuous provided that most of categories are occupied.\textsuperscript{47} Also, where the distribution is spread over a number of categories, it can be assumed that the data were generated from a continuous distribution. The indirect utility assessment instruments have more than 100 possible scores, the HAQ has 25 categories and the RAQoL has 30. In our sample, the distributions the scores of the instruments are spread across many of the “categories”, and it appears that the underlying distribution is continuous. These findings are well known in the psychometric literature and have been described almost two decades ago by Bollen et al.,\textsuperscript{48} and Johnson and Creech.\textsuperscript{49}
8.4 RECOMMENDATIONS

The results from this study give rise to some important recommendations. First of all, since application of the different indirect utility assessment instruments in an economic evaluation result in incremental cost-effectiveness ratios that differ by as much as 100%, guidelines should be developed for use of the scores from these instruments as weightings for QALYs. Without standardization to the application of these instruments, it will be difficult to compare the results of economic evaluations both within and across disease groups. Thus, at the very least, sensitivity analysis should be employed which covers the full spectrum of QALYs that could be achieved with the different indirect utility assessment instruments. Without these changes, the limitation of not being able to compare across studies directly challenges one of the advantages to using the incremental cost per QALY approach - namely, the ability to compare across studies based on the same outcome metric.

However, having stated the above recommendation, within RA and likely other disease states, it is difficult to compare among the instruments based upon their properties as HRQL scales. For example, all the indirect utility assessment instruments displayed construct validity in RA and likely are suitable as generic HRQL measures. However, in terms of reliability, responsiveness and ability to measure positive changes in RA, the HUI3 and the SF-6D appeared to be superior to the EQ-5D.

The finding that scores of the indirect utility assessment instruments vary by income reveal the possibility that bias could be introduced into economic evaluations in a number of ways. In clinical trials where an economic evaluation is to be conducted alongside, care should be taken to insure that annual household income is balanced between the two or more arms that are being compared.
As an external criterion for change, the transition question that we adopted did not appear to be the best anchor-based approach to utilize. Categories based upon the patient global assessment of change VAS appeared to perform better and thus should be utilized in future studies when an external criterion is desired.

Finally, the disease-specific RAQoL displayed excellent properties as a HRQL instruments. Not only was it feasible, well received by patients, valid (as measured using the known-groups approach to construct validity), it was highly responsive, even in a natural history study of RA. As such, there should be consideration by investigators for this scale to be used as an outcome in clinical trials and observational studies in RA.

8.4.1 Further Research

Although we have recommended that there be a standardized approach to the application of scores from these indirect utility assessment instruments in economic evaluations, a lot of effort will be required to implement this task. Research should be performed in various disease states comparing the properties of these instruments and which performs the best. After this research is conducted, a systematic approach will need to be adopted to determine which of the instruments, when applied over several disease states, yields the most valid responses.

Further comparisons of preference-based measures within RA need to be performed. For example, responses obtained with the indirect utility measures, which reflect society’s preferences, should be compared with measures such as the standard gamble that can reflect patient preferences. Comparisons about how differences with application of scores from both approaches influence economic evaluation should be performed.
Economic evaluations utilizing measured changes over time in scores obtained with the indirect utility assessment instruments should be performed. We are in the process of applying this methodology to a cost-effectiveness analysis of a new pharmacotherapeutic intervention for RA.

Methods to remove the impact of income from the scores achieved with the indirect utility instruments should be investigated. In addition, with economic evaluations already performed using these instruments, it would be useful to know if, and to what extent, double-counting has occurred and if bias has been introduced.

The RAQoL should be compared to other RA-specific HRQL measures such as the AIMS systems to determine its relative performance. Also, further examination of the responsiveness of each of the domains/attributes of the indirect utility assessment instruments should be performed to further evaluate the scales' relative merits in the assessment of RA. Finally, further research is required to investigate the mechanisms of reduced HRQL in RA secondary to lower income despite adjustment for disease severity and other chronic diseases. Considering the association between self-reported health and morbidity and mortality, this finding could result in further improvements in health beyond those offered from effective drug therapies.

8.5 CONCLUSIONS

The choice of indirect utility assessment instruments in estimating QALYs in economic evaluations for RA appears to have been somewhat arbitrary and there has been little guidance in selecting which instrument should be used. However, due to their ease and low cost of administration, there has been an increased use of scores from indirect utility
assessment instruments as QALY weightings despite the lack of comparative data on their strengths and weaknesses. As with other diseases, these observations are relevant to RA where, due to a recent explosion in the availability of effective, expensive drug therapies, economic evaluations are becoming increasingly popular.

This study provides good evidence that the scores achieved by these instruments are not interchangeable, and when used to estimate QALYs in economic evaluation, could result in very different information for policy decision-makers. Through an exploration of their properties, it can be concluded that all the indirect utility assessment instruments are feasible, valid and, for the most part reliable (with the exception of the EQ-5D). However, there are important differences between the instruments in terms of which domains of health they assess, the scores that they achieve and their relative responsiveness to changes in RA. The HUI3 and the SF-6D, based on their properties, appear to be the instruments with the most desirable properties to be utilized in RA. However, when their scores were applied to estimate QALYs in an economic evaluation for a pharmacotherapeutic intervention for RA, the outcomes (in terms of incremental cost per QALY) using the SF-6D were more than twofold that of the HUI3 (with those achieved with the EQ-5D and the HUI2 being in between).

Other important findings of this study include that annual household income is positively associated with scores achieved with the indirect utility assessment instruments, The fact that income influences the scores achieved by all the indirect utility assessment instruments (but, in our sample, only significantly for the HUI3 and the SF-6D) implies that bias might be introduced into economic evaluations that utilized these scores as QALY weightings.
Another finding of this study included that the RAQoL had excellent psychometric properties thus making it a useful outcomes measurement tool for both clinical trials and observational studies in RA. In addition, in the assessment of responsiveness, we found that the use of a patient transition question as an anchor-based, external criterion for change in RA was not as useful as using categorizations of the patient global assessment of disease activity VAS. Finally, the finding that generic HRQL is lower in patients with lower income makes future research in determining the mechanism behind this observation imperative as self-reported health is associated with morbidity and mortality.

In conclusion, although all indirect utility assessment measures appear to demonstrate the minimal requirements to assess generic HRQL in RA, when used as quality weights in the calculation of QALYs in an economic evaluation, they give vastly different results that could result in different policy recommendations. The scores of these instruments could also be biased by income. The RAQoL displayed excellent properties and is a suitable disease-specific HRQL instrument for RA.
8.6 REFERENCES


6. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health and Quality of Life Outcomes 2003;1:4 (available at http://www.hqlo.com/content/1/1/4)


33. Fortin PR, Abrahamowicz, Clarke AE, Neville C, Du Berger R, et al.. Do lupus
disease activity measures detect clinically important changes? J Rheumatol
2000:27;1421-1428.

34. Brossart DF, Clay DL, Willson VL. Methodological and statistical considerations for
threats to internal validity in pediatric outcome data: response shift in self-report

35. Groot W. Adaptation and scale of reference bias in self-assessments of quality of

self-reported and observed physical function in the elderly: the influence of response

37. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing
responsiveness: a critical review and recommendations. J Clin Epidemiol

38. Norman G. Hi! How are you today? Response shift, implicit theories and differing

39. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al.. The
American Rheumatism Association 1987 revised criteria for the classification of

40. Pickard AS. Responsiveness of generic health status measures in stroke. Doctor of

41. Wolfe F, Michaud K, Pincus T. Do rheumatology cost-effectiveness analyses make
42. Ubel PA. What is the price of life and why doesn’t it increase at the rate of inflation? Ann Intern Med 2003;163:1637-41.


APPENDIX I

THE COST OF COX INHIBITORS: HOW SELECTIVE SHOULD WE BE?

Canada's national health insurance scheme covers the cost of outpatient physician visits and acute hospital care for all Canadians. Prescription drug costs are not universally covered and the extent of coverage varies by province. However, at the very least, those over 65 years and the financially indigent are covered in every province and some also provide universal coverage after an initial deductible/co-payment. It has been estimated that almost three-quarters of Canadians have some form of coverage (either public, private or both) for prescription medications.\(^1\) Thus, it is of concern that total expenditures on prescription drugs have increased more rapidly than any other health care component. In 1999, they accounted for over 15% of total health care expenditures.\(^2\) The significance of this is underscored by the fact that in 1993 the total expenditures on prescription drugs exceeded the total expenditures on physicians for the first time.\(^3\) Over the past decade, this increase in overall drug expenditures can be attributed to both an increase in the number of drugs being consumed and the higher average price of individual drugs. Since 1995, Canadians have spent on average 9% more each year on drugs. In 1999, they spent 11.4% more than they had in 1998.\(^1\)

For rheumatology, these trends in Canadian drug utilization and expenditures are even more pronounced than other fields. In 1999, prescription drugs used in the treatment of arthritic diseases accounted for approximately 3% ($231 million) of the 8.3 billion (approximately $270 per person) spent on prescription drugs.\(^4\) Although this figure is low compared to cardiovascular medications (15% of expenditures on prescription drugs or $1.2 billion), it represents one of the largest year over year increases in prescription drug purchases by therapeutic class. When compared to 1998, expenditures on arthritis drugs in 1999 increased by 27%. In addition, 4% (11,500) of all prescriptions in 1999 were for
arthritis treatment, a 15% increase over the number in 1998. These numbers are similar to those in the US where antiarthritics accounted for 3% ($3.7 billion) of the $112 billion prescription drug market (approximately $400 per person) and 4% (108,500) of all prescriptions. Despite the similarity in these percentages, the increase in prescription costs and numbers of prescriptions in the US (1999 vs 1998) were higher: 77% and 19%, respectively.

In part, this large increase in Canadian drug expenditures was due to the introduction of newly patented medications including the COX-2 selective inhibitors, celecoxib (Celebrex®) and rofecoxib (Vioxx®). Celecoxib was introduced to the Canadian market in April 1999 whereas rofecoxib was available only in October 1999. Consequently, due to its earlier availability, celecoxib had the larger impact on the Canadian prescription drug market during 1999. Specifically, purchases of celecoxib were almost $70 million compared to less than $3 million for rofecoxib. In addition, celecoxib was the 48th most commonly prescribed drug in Canada (1.8 million prescriptions in 1999) and the 14th highest in prescription drug expenditures. These drugs generated considerable advertising. For instance, in 1999, celecoxib was the 4th most widely promoted drug as measured by estimated advertising expenditures ($4 million), advertisement pages in professional journals (336 pages), health professional detailing (55,000), and samples distributed (47,000).

Of note, our data reveals that the COX-2 selective inhibitors have resulted in significant changes in the patterns and costs of both osteoarthritis and rheumatoid arthritis pharmacotherapy in Canada (Figures 1 and 2). For rheumatoid arthritis, there was an average of approximately 570 thousand prescriptions per year for NSAIDS from 1995 to 1998. For 1999 and 2000 (forecasted data), we have estimated that there will be approximately 505
thousand prescriptions for both selective (260 thousand) and non-selective (245 thousand) COX inhibitors. While the number of prescriptions may have decreased, the annual cost of these therapies has risen from an average of $13 million to approximately $21 million. Interestingly, the impact of the COX-2 selective inhibitors is even more pronounced in osteoarthritis. From 1995-1998, on average, there were about 1.7 million annual prescriptions for NSAIDS resulting in annual expenditures of $44 million. After the introduction of COX-2 selective agents, these figures increased to almost 1.9 million prescriptions and $78 million in expenditures. In that only the provinces of Alberta, Manitoba and Quebec were providing full reimbursement for at least one of the prescribed COX-2 selective agents during 1999 (other provinces providing partial reimbursement or restricted status coverage included Ontario, New Brunswick, Newfoundland, Nova Scotia, and Saskatchewan, while only British Columbia and Prince Edward Island currently do not provide coverage for either of these agents), these figures represent a dramatic shift in prescribing habits. Clearly, physicians have been convinced of the added benefit of the agents and Canadian patients, usually reluctant to pay for a drug when an alternative that is covered by the health care system exists, have 'bought into' this belief.

Authors of recent editorials have attempted to place the new specific COX-2 inhibitors into clinical perspective.\textsuperscript{5-7} The International Cox-2 Study Group outlined the specific instances where these authors believe that the COX-2 specific inhibitors should be used in favour of classic COX nonspecific NSAIDS.\textsuperscript{5} These recommendations, for the most part, were based on expert opinion derived from epidemiological studies that evaluated the risks of serious gastrointestinal complications in certain patient groups. Since these
recommendations are fairly restrictive, it is unlikely that they are being followed precisely, given the rapid increase in the use of COX-2 specific inhibitors seen in our data.

The COX-2 specific inhibitors possess equivalent efficacy as COX nonspecific NSAIDS in both RA \(^8\) and OA \(^9\). From the available evidence, it has been estimated that the number of patients that one would need to treat with a COX-2 specific inhibitor (instead of COX nonspecific NSAIDS) to avoid a serious gastrointestinal event and death is approximately 130 and 1300, respectively.\(^6\) Another editorial estimated the yearly incremental cost (in 1999 US dollars) of preventing one complicated ulcer with these new agents to be approximately $30,000.\(^7\) However, it remains to be determined if this reduction in ulcer-related hospitalizations, morbidity, and mortality leads to a reduction in overall treatment costs with the COX-2 specific inhibitors over traditional agents.

The data suggest that COX-2 specific inhibitors are replacing the traditional NSAIDs. Of particular interest is the rapid increase in the utilization of COX-2 specific inhibitors in OA. Due to the higher risk-benefit ratio associated with using NSAIDs, the first line management of osteoarthritis has historically been acetaminophen.\(^10\) However, NSAIDs have been shown to be moderately more effective than acetaminophen in osteoarthritis\(^11\) and these patients are less willing than rheumatoid arthritis patients to accept the risk of gastrointestinal toxicity associated with nonspecific COX inhibitors.\(^12\) Clinicians may feel that patients prescribed COX-2 specific inhibitors will realize the additional benefit of an NSAID over acetaminophen, without the increase in risk of toxicity.

In the past, treatment for arthritis has been inexpensive relative to other disease groups. A sufficiently large number of NSAIDs existed in generic versions to allow considerable choice, and most disease suppressive medications are inexpensive. This is changing. The
prospect of COX-2 selective antiinflammatory drugs being used for more than 50% of arthritis consumers, along with new second-line drugs such as leflunomide (Arava®), infliximab (Remicade®) and etanercept (Enbrel®) being used to treat rheumatoid arthritis, juvenile rheumatoid arthritis, spondyloarthritis and perhaps other inflammatory types of arthritis is rapidly modifying the prescribing scene. Even so, arthritis overall will remain relatively inexpensive to treat given the considerable burden of disease in the community and the consequences to the individual and to society of not treating these diseases with effective medications. Nonetheless, the cost-effectiveness of the new therapies has to be unequivocally established using rigorous scientific methods. Only then will large third party-payers such as health maintenance organizations, pharmacy benefit managers, and private insurers in the US and provincial drug plans in Canada be convinced that they represent "good value" for scarce health care dollars.
REFERENCES


FIGURE 1: EXPENDITURES ON PRESCRIPTION DRUGS BY THERAPEUTIC CLASS FOR OA AND RA COMBINED, IN CANADA OVER TIME

Costs: 2000 figures based on an annualized projection from 1st quarter data (January to March 2000)
FIGURE 2: NUMBER OF PRESCRIPTIONS BY THERAPEUTIC DRUG CLASS FOR RA AND OA COMBINED, IN CANADA OVER TIME

Costs: 2000 figures based on an annualized projection from 1st quarter data (January to March 2000)
APPENDIX II

PRACTICAL PHARMACOGENETICS: THE COST-EFFECTIVENESS OF SCREENING FOR THIOPURINE S-METHYLTRANSFERASE (TPMT) POLYMORPHISMS IN PATIENTS WITH RHEUMATOLOGICAL CONDITIONS TREATED WITH AZATHIOPRINE (AZA)

*Marra CA, Esdaile JM, Anis AH. J Rheumatol 2002;36:1851-5*
INTRODUCTION

The study of pharmacogenetics may assist in the determination of the considerable variation in respond to drug therapy by individuals. With the recent ability to determine genetic polymorphisms that are associated with the efficacy or toxicity of particular drugs prior to the initiation of therapy, the potential exists to tailor drug regimens for individuals. Therefore, the adoption of these new diagnostic techniques could lead to an enhancement of the beneficial effects and the reduction of adverse effects of therapeutic agents. However, as with any new, expensive technology, it is imperative to systematically assess the cost-effectiveness of this new approach in order to determine if scarce health care dollars should be allocated to its routine integration into the health care system. Thus, as the potential arises for genetic testing to be utilized to predict the outcomes associated with drugs used for rheumatological conditions, systematic evaluations must occur to assess their relative economic merit.

Azathioprine (AZA) interferes with the de novo synthesis of inosinic acid via feedback inhibition of 6-thiosinosinic acid. AZA also inhibits the interconversion of purine bases such as inosine to adenine and guanine ribonucleotides. AZA, as with all thioguanines, is metabolized to the active metabolites referred to collectively as 6-thioguanine nucleotides (6-TGN), which, in turn are inactivated by thiopurine-6-methyltransferase (TPMT). The active metabolites, 6-TGN, accumulate in tissues and are thought to be responsible for many of AZA’s serious toxicities. TPMT is a cytosolic enzyme that preferentially catalyzes the inactivation (S-methylation) of the active metabolites of the thiopurines (6-mercaptopurine, AZA and thioguanine). TPMT shows codominant genetic polymorphism with about 85-90% of
people exhibiting high TPMT activity and approximately 10-15% of individuals having intermediate TPMT activity caused by heterozygosity at the TPMT locus.\textsuperscript{5,6} About 1 in 300 individuals inherits TPMT deficiency as an autosomal recessive trait. Unfortunately, no phenotype allows for the detection of either TPMT deficiency or intermediate activity thus these conditions, when identified, are usually only detected after a severe adverse reaction to a thiopurine medication.\textsuperscript{5} Case reports, case series and case-control studies suggest that a reduction in TPMT activity is associated with the development of toxicity to thiopurine medications.\textsuperscript{3,5,7-11} Due to these findings, some investigators have advocated the prospective measurement of TPMT activity.\textsuperscript{5,8-11} Unfortunately, TPMT assays are not clinically available, are influenced by several exogenous factors (eg. recent red blood cell transfusion, diuretic therapy, alcoholism) and are expensive, thus precluding their use.\textsuperscript{5}

Yates CR et al. developed and validated polymerase chain reaction (PCR)-based methods for detection of TPMT mutations in the genomic DNA of humans.\textsuperscript{5} Specifically they established the genetic basis for TPMT polymorphisms and discovered that, at the genetic level, the three groups of TPMT activity can be classified as homozygous wild-type (high activity), heterozygous mutants (intermediate activity) and homozygous mutants (deficiency). This PCR test for TPMT activity yielded the following sensitivities and specificities: 1) high activity, 100% and 100%; 2) intermediate activity, 95% and 100%; 3) deficiency, 100% and 100%. In addition, since PCR is relatively inexpensive, widely available, rapid, not affected by exogenous factors, and widely used clinically in other areas, it is a useful diagnostic test for the determination of TPMT activity.

Decision analysis has been widely employed as a tool to systematically assist technology assessment when there is uncertainty about clinical or economic outcomes.\textsuperscript{12,13}
This technique has been used extensively to model both the outcomes and cost-effectiveness of diagnostic testing in a variety of disease states.\textsuperscript{14,15,16} However, to date, there has been no published cost-effectiveness analysis examining the cost-effectiveness of screening for enzyme polymorphism-based strategies.

The objective of this study was to model the cost-effectiveness of two alternate AZA treatment strategies in two common autoimmune rheumatologic conditions (rheumatoid arthritis and SLE). The two strategies consist of: 1) PCR testing strategy prior to initiation of AZA resulting in dosage reduction in cases of reduced TPMT activity and deficiency; and 2) No testing with usual therapy.

METHODS

\textit{Literature Review}

To determine the incidence of AZA induced adverse events (primarily agranulocytosis, leukopenia and pancytopenia), a systematic search of the English language literature using computerized MEDLINE, EMBASE, Pre-MEDLINE and Cochrane Systematic Review databases from 1966 to September 2000 and a manual search of references from retrieved articles were performed. Search terms included AZA, TPMT, polymorphisms, rheumatoid arthritis, rheumatology, systemic lupus erythematosus, leukopenia, or pancytopenia.

A recent case report outlined the occurrence of severe pancytopenia in a 14 year old girl with juvenile spondylarthritis on 3mg/kg/day of AZA.\textsuperscript{11} Further laboratory evaluation identified a homozygous deficiency of TPMT which led to a buildup of the AZA metabolite
6-thioguanine. After identification and withdrawal of the drug, the patient gradually improved over an eight-week course.

A prospective case-control study measured the TPMT enzyme activity in patients via HPLC methods. Cases were defined as longstanding RA patients on AZA (n= 33) while the two control groups were patients with early RA (n=24) and healthy volunteers (n=42). The authors found that 14 out of the 33 (42%) AZA-treated patients developed severe toxicity and had to be withdrawn from therapy. Of these, 7 (50%) had intermediate TPMT activity and the toxicity exhibited was gastrointestinal intolerance or hematological cytopenia. The relative risk of developing severe toxicity in patients with intermediate TPMT activity compared to those with high TPMT activity was 3.1 (95% CI of 1.6-6.2). The authors also postulated that intermediate TPMT activity was associated with both hematological and gastrointestinal toxicities.

A recent study has cast some doubt on whether the determination of patients’ TPMT activity assessed via genotyping has any predictive value in determining the adverse reactions to AZA. The investigators correlated the adverse hematological toxicities with the TPMT genotype in 120 SLE patients attending their rheumatology clinic. The authors did not find a relationship between TPMT genotype and AZA-induced toxicity except for a single homozygote-deficient patient. However, this study has several limitations. Although 120 patients were tested, only 78 had ever received AZA. Of the seven patients found to have TPMT polymorphisms, only four had received AZA. Of these, three were heterzygotes who received very small doses of AZA (0.4 and 0.6 mg/kg) and thus would not be expected to develop toxicity. In addition, it is unclear how the authors elucidated adverse reactions to AZA as no definitions or methods of collection (prospective or retrospective) are supplied.
In a recent Cochrane Systematic Review of the use of AZA in the treatment of rheumatoid arthritis, the efficacy and toxicity of AZA was evaluated. Specifically, in comparison to patients who received placebo, those that received AZA were found to be 4.6 times more likely to withdraw from therapy due to adverse effects over a six month treatment period. The most common adverse reactions were found to be gastrointestinal symptoms (15%), mucocutaneous reactions (26%) and hematological disturbances (9%). In addition, the incidence of leukopenia in these trials was 21%, but most cases resolved by lowering the dose of AZA.

Finally, Black et al. evaluated the incidence of polymorphic inactivation of AZA by thiopurine methyltransferase and the development of clinical toxicity in a cohort of 67 patients from two rheumatology units who were prescribed AZA. Polymerase chain reaction-based assays were used to detect mutations in TPMT. Six of the 67 patients (9%) were heterozygous for mutant TPMT alleles. Of these, five patients discontinued therapy within one month of starting treatment due to leukopenia (< 3.5 \times 10^9/L). On further questioning, the remaining patient admitted to having not taken any AZA.

**Decision Analytic Model**

A predictive, decision analytic model was created using Data for HealthCare by TreeAge Software (version 3.5, Williamstown, Mass.) to estimate the effects and costs of each of the strategies (Figure 1). Specifically, the strategies evaluated were: 1) a PCR test screening for TPMT reductions or deficiencies (as per the methods described by Yates et al., the PCR assays isolate and characterize two mutant alleles that are associated with TPMT deficiency, TPMT*2 and TPMT*3A). and a reduction in dose based on the results of the
test; or 2) usual full-dose therapy with AZA. Because TPMT reductions or deficiencies result in an accumulation of the active metabolites, modified dosing could be implemented in order to avoid toxicity without compromising therapeutic effect. The dosing guidelines by TPMT genotype are as follows: 1) homozygous wild type (normal TPMT activity): target dose of 2.0–2.5 mg/kg/day; 2) heterozygous (reduced TPMT activity): target dose of 1.0 mg/kg/day; and 3) homozygous mutant (deficient of TPMT activity): target dose of 0.25 mg/kg/day. The time horizon for the development of adverse events in the model was 6 months so discounting of costs and effects was not required. All costs were in 1999 Canadian dollars (1 Canadian dollar = 0.67 United States dollar).

Target Population

For the purposes of this analysis, only costs and outcomes of patients with rheumatological conditions (mainly RA and SLE) were included in the model.

Probabilities and Values Used in the Model

All probabilities and values that were used both in the base case and sensitivity analyses are detailed in Table 1. To estimate the incidence of hematological disturbances (agranulocytosis and/or leukopenia) that would be preventable through a prophylactic AZA dosage reduction in susceptible individuals, the results of the literature review previously described were evaluated. From the Cochrane Systematic Review of AZA in RA, it appears that approximately 9% of the RA patients experienced hematological toxicities in the pooled results of three published six month randomized, clinical trials and up to 21% experienced leukopenia. This figure is in agreement with the study by Black et al.
disturbance was not included in the model as due to the lack of consistent correlation with their development with TPMT, the idiosyncratic nature of gastrointestinal disturbances (i.e. in clinical trials, even placebo arms have high incidences of gastrointestinal events), and the lack of significant medical attention necessary to resolve most cases.

From the results of the case-control study, it also appears that the prospective evaluation of TPMT activity would not be sufficient to eliminate all cases of hematological toxicity. Thus, for the base case analysis, we assumed that 50% of the cases of hematological toxicity could be eliminated by the PCR-testing strategy through the identification of susceptible individuals and the implementation of a prophylactic dosage reduction.

To determine the incidence of true and false positive and true and false negative PCR test results, we utilized the sensitivity and specificity of the TPMT PCR genotype testing procedure outlined by Yates et al.. To convert these parameters into probabilities, we first converted the sensitivity and specificity values into both positive and negative likelihood ratios. The likelihood ratio expresses the odds that the test result occurs in patients with the disease versus those without the disease. Thus, the positive likelihood ratio equals sensitivity/(1-specificity) and the negative likelihood ratio equals (1-sensitivity)/specificity.

Decision probabilities were calculated from these likelihood parameters and the a priori probability of having a positive result using Baye’s Revision.

Costs used in the Model

All cost data that were used in the model are summarized in Table 2. Only direct medical costs were considered. Since the cost of the PCR genotyping test is not available in local clinical laboratories and is not currently being utilized in this manner, we estimated the
cost of this test from the costs of performing other PCR based tests that are clinically available. Specifically, we estimated that the base case cost of the PCR genotype testing would be $100 Canadian per person.

From interviews with a hospital-based, academic hematologist and an office-based, community hematologist, we estimated the costs associated with caring for individuals who developed a hematological disturbance. Specifically, the experts felt that approximately 50% of these individuals would need to be hospitalized due to the risks associated with agranulocytosis and the underlying immunosuppression that occurs with drug treatment of their underlying illness. Those ill enough to be hospitalized were assumed to require intravenous antibacterials and G-CSF. In addition, daily complete counts with differential and blood cultures if fever was present would be ordered. Average duration of hospitalization in those hospitalized was estimated to be 10 days. A fully-allocated hospital cost model was utilized to estimate the costs associated with resource utilization.\(^{23}\)

For those able to be managed as outpatients (estimated to be 50%), costs of physician visits and laboratory tests were determined from the Provincial Guide to Medical Fees and drug costs were determined from the IMS Health Canada database of provincial drug costs. It was assumed that those managed as outpatients would incur two office visits and have two complete counts with differential. Approximately 30% would receive G-CSF as outpatient therapy. Other outpatient costs would be incurred by those who were false positives with the PCR genotype testing. Since these individuals would be treated with a reduced dosage of AZA, the likelihood of treatment failure would be high. Thus, these individuals were assumed to incur an additional physician office visit and an additional prescription for a medication.
RESULTS

In the base case model, the normal dosing strategy cost $677 per patient whereas the genotype-directed dosing strategy cost $663 per patient. In the genotype dosing strategy, the number needed to treat to avoid one adverse event over six months was 20. Thus, the genotype-based dosing strategy dominated (it was both more effective and less costly) the usual dosing strategy.

Sensitivity analyses

The most influential parameters on the results of the model in univariate sensitivity analyses are summarized in Figure 2. Since the PCR test was not performed by clinical laboratories in our region, we did not have an accurate cost estimate for this parameter. Thus, we have utilized a wide range of test costs for the sensitivity analysis (Figure 3). As it can be seen from the results of this analysis, the outcomes of the model are very sensitive to the value of this parameter such that the threshold value was approximately $114. However, the model was quite robust to plausible changes in the sensitivity and specificity of the PCR test with threshold values of 0.929 and 0.947, respectively. A sensitivity analysis on the probability of the preventable hematological cytopenias revealed the threshold value to be 4.4%, above which the genotype based dosing strategy is less costly and below which the usual dosing is less costly. Finally, a sensitivity analysis on the probability receiving inpatient care for the hematological cytopenias revealed the threshold value to be 44%, above which the genotype based dosing strategy is less costly and below which the usual dosing is less costly.
DISCUSSION

We found that the prospective adoption of an AZA dosing strategy based on the molecular diagnosis of TPMT reduction and deficiency by genotypic results in a small cost reduction when using our base case estimates of costs and probabilities. In addition, this strategy results in avoidance of approximately half of all serious hematological toxicities associated with AZA therapy. The cost difference was sensitive to small changes in the cost of PCR and the probability of preventable hematological cytopenia. However, the clinical endpoint (the avoidance of hematological cytopenia) was robust across the range tested in our sensitivity analyses.

To our knowledge, this is the first economic evaluation to consider the impact of drug metabolizing enzymatic heterogeneity secondary to measurable genetic polymorphisms within the human genome on the disposition of a pharmaceutical agent and the subsequent development of adverse events. Other economic decision analyses have been completed that have evaluated the cost-effectiveness of determining the genotype of HIV and the impact on drug resistance.\(^{24,25}\) Similar to our study, these investigators found that the adoption of genetic diagnostic techniques can facilitate an improvement in the outcomes achieved with drug therapy.

One of the major impediments to implementing the prospective testing of TPMT activity through genotyping is the availability of the PCR test procedure in a clinical laboratory that permits a short turn-around time for the results. Currently, there are clinical sites where this test has been implemented to prospectively diagnose patients and allow for the reduction of thiopurine dosage and the avoidance of drug toxicity.\(^{4,5}\) We would therefore advocate the adoption of this technology should the necessary infra-structure (ie. clinical laboratory) and
expertise be available. In addition, it is possible, as with any diagnostic test, clinicians may not order the PCR test prior to initiating therapy but may use it as a confirmatory mechanism once toxicity has developed. Obviously, the cost-effectiveness of using the PCR test in this manner would be considerably different than determined from this study.

Our analysis has several limitations. First of all, we did not consider indirect costs as these were not available. Since the indirect costs would likely be higher in the strategy that included more adverse events, it is likely that the omission of these conservatively biases our results against the genotype based dosing strategy. In addition, we did not consider health-related quality of life for the various health states represented in our model (leukopenia without infection, leukopenia with infection, and healthy) as these were not available. Future analyses should be performed such that these outcomes can be integrated into the model.

In summary, we found that the costs of determining TPMT activity through genotyping and a corresponding AZA dose-reduction in susceptible individuals ranged from being slightly cost saving to relatively modest in light of its outcome benefit when compared with the usual approach to AZA dosing. Our analysis can act as a data-driven basis for policy decisions about genetic testing to avoid side effects associated with AZA in patients with rheumatic diseases.
REFERENCES


332
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Base Case</th>
<th>Range</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Upper</td>
<td>Lower</td>
</tr>
<tr>
<td>Hematological cytopenia</td>
<td>0.09</td>
<td>0.03</td>
<td>0.15</td>
</tr>
<tr>
<td>Sensitivity of PCR genotype test (%)</td>
<td>95.2</td>
<td>76.2</td>
<td>99.9</td>
</tr>
<tr>
<td>Specificity of PCR genotype test (%)</td>
<td>100</td>
<td>83.9</td>
<td>100</td>
</tr>
<tr>
<td>Positive likelihood ratio</td>
<td>$\sim\infty$</td>
<td>4.73</td>
<td>$\sim\infty$</td>
</tr>
<tr>
<td>Negative likelihood ratio</td>
<td>0.05</td>
<td>0.01</td>
<td>0.29</td>
</tr>
<tr>
<td>Inpatient Care</td>
<td>0.50</td>
<td>0.20</td>
<td>0.99</td>
</tr>
</tbody>
</table>
TABLE 2: COSTS USED IN THE DECISION MODEL

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Base Case</th>
<th>Range</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adverse event costs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inpatient</td>
<td>2679</td>
<td>2933</td>
<td>2592</td>
</tr>
<tr>
<td>Outpatient</td>
<td>790</td>
<td>990</td>
<td>590</td>
</tr>
<tr>
<td>PCR costs</td>
<td>100</td>
<td>50</td>
<td>200</td>
</tr>
<tr>
<td>Treatment failure cost</td>
<td>573</td>
<td>775</td>
<td>275</td>
</tr>
<tr>
<td>AZA Cost/50mg tablet</td>
<td>0.67</td>
<td>0.63</td>
<td>0.80</td>
</tr>
<tr>
<td>Pharmacist dispensing fee per episode</td>
<td>6.62</td>
<td>2.08</td>
<td>11.36</td>
</tr>
</tbody>
</table>

1 All costs in the model are 1999 Canadian Dollars

2 British Columbia Genome Sequencing Centre

3 Average Canadian cost per generic brand tablet as reported by IMS Health Canada
Decision Model Legend: SQUARE: decision node between the two strategies (normal and genotype dosing); CIRCLE: chance node from which probabilities of events emanate; TRIANGLE: terminal node denoting the final outcome and payoffs. Normal dosing refers to typical AZA dosing to a target dose of 2.0 - 2.5 mg/kg/day, Genotype dosing refers to the use of recommended reductions of AZA dosing according to genotype, pADR refers to the probability of hematological cytopenia in the normal dosing arm (the # is the corresponding probability of 1-pADR), the PCR testing properties (Test+, Test-, True+, False+, False-, True-) were derived from the sensitivity and specificity; All costs (cost 1 to cost 6) represent embedded formulae that calculate the costs associated with each decision path.
FIGURE 2: UNIVARIATE SENSITIVITY ANALYSES – MOST INFLUENTIAL PARAMETERS

Univariate sensitivity analyses of influential parameters

The dark vertical lines represent threshold points. The Y-axis of the graph refers to the expected value of the genotype dosing strategy.
FIGURE 3: SENSITIVITY ANALYSIS OF THE COST OF THE PCR TEST

Sensitivity Analysis on
cost of PCR

Threshold Values:
- Normal dosing
- Genotype dosing

EV = Expected or threshold value. This is the point when the value of the parameter being tested results in an equal cost in both strategies.
APPENDIX III

THE EFFECTIVENESS AND TOXICITY OF CYCLOSPORINE A IN RHEUMATOID ARTHRITIS: LONGITUDINAL ANALYSIS OF A POPULATION-BASED REGISTRY

INTRODUCTION

Rheumatoid arthritis (RA) is a chronic, inflammatory, systemic disease. Due to its progressive nature, extra-articular manifestations and the adverse effects of medications used for treatment, patients with RA have an increased risk of morbidity and mortality. Aggressive pharmacotherapy is recommended early in the course of the disease in order to prevent severe disability and death.

Methotrexate (MTX) has become the disease modifying antirheumatic drug (DMARD) of choice due to its superior efficacy and tolerability when compared to other agents. However, MTX, used as monotherapy or in combination with another DMARD, is rarely associated with sustained disease remissions and many patients do not respond to and/or tolerate this agent. Thus, newer treatments, such as cyclosporine A (CSA), have emerged as additional agents for patients who fail usual DMARD therapy.

Placebo-controlled or comparative, randomized trials of CSA in RA have demonstrated that CSA is more effective than placebo and is at least as effective as penicillamine and azathioprine. However, treatment with CSA is associated with an increased incidence of reversible renal dysfunction and hypertension. Other significant adverse reactions include headache, tremor, gastrointestinal disturbances and gum hyperplasia. Although these trials demonstrated efficacy, they were of short duration (2 to 12 months) and enrolled a selected group of patients willing to provide informed consent, thereby limiting somewhat their generalisability for the treatment of patients with RA. The objectives of the present study were to determine CSA’s long-term effectiveness and toxicity and determinants of such in a population-based inception cohort of patients with RA.
PATIENTS AND METHODS

Starting in January 1991, data were prospectively recorded for an inception cohort of all patients with RA treated with CSA in British Columbia (total population 3.4 million). To be eligible for reimbursement from the B.C. government for the costs of CSA, patients had to meet the revised criteria of the American College of Rheumatology for RA and had experienced an inadequate response to prior DMARD therapy including MTX. The same physicians and nurses evaluated all patients over time. CSA treatment was started at 2.5 - 3.0 mg/kg/day in 2 divided doses and was increased or decreased at monthly intervals until the minimum effective dose was achieved or toxicity developed. Discontinuation of CSA for ineffectiveness or toxicity was at the discretion of the attending rheumatologist in response to regular monitoring parameters described below. The reason for CSA discontinuation was determined prospectively. For the purposes of this study, data from January 1991 through December 1997 were analyzed.

Data included demographics, disease history (years of RA, extra-articular manifestations), RA treatment history (previous DMARD use, prednisone dose at CSA initiation), baseline clinical assessments, and details of CSA treatment (initiation dose, maximum dose, number of dose changes, dose at discontinuation, and the reason for discontinuation). Concomitant therapies including MTX were also recorded. Each patient was assessed monthly for the first three months and then at least every 3 months thereafter for CSA effectiveness (number of tender joints, duration of morning stiffness) and measurements of CSA toxicity. Patients were reviewed for toxicity (systolic (SBP) and
diastolic blood pressure (DBP), and serum creatinine) weekly for the first month and then were followed on a monthly basis.

**Statistical Analysis**

Kaplan-Meier survival analysis methods were used to estimate the probability of discontinuing CSA due to either lack of efficacy or toxicity. The log-rank test was used to test the null hypotheses that the survival time (i.e. total duration of CSA treatment) between categorical factors (sex, age groups, years of RA prior to CSA treatment divided into those with RA < or > 10 years, concomitant MTX use) was the same. Right censoring was used for patients lost to follow-up prior to discontinuation of CSA. Survival time was used as a measure of overall CSA effectiveness.

To develop a model that predicted how long a patient continued on CSA therapy, Cox proportional hazard models were used to evaluate CSA discontinuations, adjusting for the effects of age, sex, duration of RA, combination with MTX (yes/no), year of CSA initiation, and baseline measures of disease severity (tender joint count and duration of morning stiffness) in a forward stepwise fashion. Validity of the proportional hazards assumption was explored by inspecting plots of log[-log(proportion continuing CSA)] versus log(time) for categorical independent variables in the model. For both Kaplan Meier and Cox regression, significance was set at an α level of < 0.05 for the model and parameter estimates.

Generalized estimating equation (GEE) models were utilized to examine the relationship between each of the dependent variables (efficacy variables [joint count and duration of morning stiffness] and toxicity variables [serum creatinine, SBP, and DBP]) with a number of explanatory variables (sex, age, duration of rheumatoid arthritis, combination
treatment with MTX, number of previous DMARDs, number of extra-articular manifestations, and year of CSA start). GEE models were used because the repeated measures for each subject over time were assumed to be correlated. Joint count data were assumed to follow a Poisson distribution and thus the log-link function was applied. Morning stiffness, systolic blood pressure, diastolic blood pressure and serum creatinine were assumed to be normally distributed and thus the identity-link function was applied. An exchangeable correlation matrix was utilized for all analyses, which assumes that all observations are equally correlated. Significance was set at an $\alpha$ level of $< 0.05$ for the model and parameter estimates. Interactions between all significant covariates were tested in all regression models and residual plots were inspected to assess the validity of the models.

RESULTS

One hundred and thirty three patients were started on CSA treatment of which 100 (75%) were female. The median age was 58 years (range 28-82) with a median duration of RA of 13 years (range 2-54). Ninety-nine (74%) patients had extra-articular manifestations of RA (median 1, range 0-5) prior to CSA therapy with the most common being keratoconjunctivitis sicca (72 patients), rheumatoid nodules (68 patients), and vasculitis (15 patients).

Patients had received a median of 5 previous treatments with DMARDs [range 2-8] with inadequate response or toxicity. Previous DMARD use was as follows: MTX in 120 (90%), oral or parenteral gold products in 114 (86%), penicillamine in 90 (68%), azathioprine in 87 (65%), hydroxychloroquine in 83 (62%), sulfasalazine in 69 (52%), chlorambucil in 13 (10%), and cyclophosphamide in 6 (5%). At the time of CSA initiation,
102 patients (77%) were receiving prednisone (median dose 5 mg/day, range 2.5-60 mg) and 80 (60%) were receiving nonsteroidal anti-inflammatory drugs (NSAIDs). The median maximum CSA dose utilized was 4.1 mg/kg/day (range 2–7) whereas the median CSA dose at cessation was 3.3 mg/kg/day (range 1-6).

Twenty-seven patients (20%) received concomitant MTX therapy for at least a portion of their CSA treatment course. Of these, 26 (96%) had received previous MTX therapy prior to CSA initiation, 18 (67%) continued receiving MTX from the period prior to CSA initiation, and 9 (33%) initiated MTX treatment during CSA therapy. Neither baseline demographics nor treatment variables differed significantly between those treated with the combination therapy and those treated with CSA monotherapy, or between patients who were on MTX prior to initiating CSA and those who initiated MTX therapy during CSA.

Survival Analyses

Thirty-seven of the 133 patients (28%) discontinued CSA prematurely due to ineffectiveness (n=19) or toxicity (n=18). Of those discontinued due to toxicity, elevations in serum creatinine (n=10), hypertension (n=4), infections (n=3), and gingival hyperplasia (n=1) were the causes. On follow-up after CSA discontinuation, no patient developed elevated SCr or blood pressure secondary to CSA treatment. The remainder of the patients (n=96) was right censored due to the continuation of CSA therapy beyond the end of the study (n=61), loss to follow-up (n=24), death unrelated to RA/CSA (n=6), or cessation of CSA unrelated to ineffectiveness or toxicity (n=5). The median time to CSA discontinuation was 75 months (95% CI 38 to 112) assessed by Kaplan Meier analysis (Figure 1). In the initial 24 months of CSA treatment, nearly all of the discontinuations due to ineffectiveness
(18 out of 19) occurred. For the remaining months of CSA treatment, most discontinuations were due to toxicity (n=6) as opposed to ineffectiveness (n=1).

The evaluation of predictors of continued CSA treatment revealed that only combination with MTX (yes/no) was a significant predictor (p=0.012) (Figure 2). Gender, age (divided into 20 year increments), years of RA prior to CSA treatment (divided into those with RA < or > 10 years), and extra-articular manifestations (yes/no) were not significant predictors for CSA continuation.

*Cox Proportional Hazards Model and Generalized Estimating Equations Models*

The Cox proportional hazards model revealed that concomitant MTX and baseline affected joint count (divided by 10) were predictive of continued CSA therapy. MTX appears to be protective with respect to CSA discontinuation (relative hazard 0.22, 95% CI 0.10 to 0.94), whereas the baseline affected joint count was associated with an increased relative hazard of discontinuing CSA (1.28, 95% CI 1.00 to 1.64).

The results from all of the GEE multivariate models are presented in Table 1 and are discussed below.

*Effectiveness Variables*

Only the variables time on CSA and the use of MTX were significantly associated with a reduction in the joint count. According to the model, for patients who received the MTX/CSA combination therapy, the joint count was reduced by 21% (95% CI 1% to 37%) when compared to those receiving CSA monotherapy. In addition, for each month of continued treatment with CSA, the joint count was reduced by 1% (95% CI 0% to 2%).
For duration of morning stiffness, only the time on CSA therapy was found to be significant and was not influenced by measured potential confounders. Thus, for every month on CSA therapy, the duration of morning stiffness decreased by 2.0 minutes (95% CI 1.1 to 3.0).

**Toxicity Variables**

The number of previous DMARDs utilized prior to CSA therapy was significantly associated with increased SBP after adjusting for age and weight. With each additional DMARD utilized prior to starting CSA, SBP increased by 7.2 mmHg (95% CI 2.7 -11.7). Similarly, both the time on CSA therapy and the number of previous DMARDs utilized prior to CSA therapy were significantly associated with increased DBP after adjusting for the effects of age and weight. For each additional month on CSA therapy, the model predicted an increase in DBP of 0.07 mmHg (95% CI 0.02-0.09). Likewise, for each additional DMARD utilized prior to CSA therapy, the diastolic blood pressure increased by 3.8 mmHg (95% CI 3.0-6.4). As expected, both age and weight were significantly associated with both SBP and DBP. No interactions between predictors were found to be significant in either model.

Both the time on CSA therapy and the number of previous DMARDs utilized prior to CSA therapy were significantly associated with increased serum creatinine after adjusting for the effects of age, weight, and years of RA prior to CSA therapy. The number of previously tried DMARDs had a large impact in predicting nephrotoxicity as each additional DMARD resulted in an increase of 35 µmole/L in serum creatinine (95% CI 22 to 48). In addition, an interaction term between length of CSA treatment and the number of previous DMARDs was
also significant. The concomitant use of MTX was not significantly associated with increases in any of the toxicity variables.

**DISCUSSION**

We report the first, population-based, longitudinal experience with CSA in an inception cohort of RA patients. We note that CSA is both safe and effective over long term use and its effectiveness is enhanced when combined with MTX. Other major findings of our analyses include that the clinical improvement associated with CSA is enhanced by the duration of CSA treatment, that there is a specific discontinuation pattern for CSA, and that previous DMARD therapy is a significant predictor for the development of toxicity.

Importantly, the combination of MTX with CSA resulted in additional clinical improvement over CSA monotherapy without additional toxicity. The use of combination CSA/MTX was associated with longer CSA use and an improvement in joint counts, which is in general agreement with the only randomized, double-blind trial evaluating the efficacy of MTX plus CSA.\(^\text{16}\) In this multicenter trial, the combination of the two DMARDs was superior to MTX alone resulting in a clinically significant improvement that was maintained for up to 12 months without additional toxicity. It would appear that the benefit of combination therapy also occurs when MTX is added to CSA, and is maintained over the long-term use of CSA. Thus, combination therapy with MTX and CSA should be considered for all patients being placed on CSA. Further randomized, clinical trials could compare CSA/MTX to other recently successful combination therapies such as MTX plus infliximab\(^\text{17}\) to determine their relative efficacy and cost-effectiveness.

Another notable result was that there was no chronic or life-threatening toxicity associated with long-term use of CSA. No patient developed severe or irreversible side
effects although almost half of all CSA discontinuations were due to toxicity (mainly
transient increases in SCr and blood pressure). The duration of treatment with CSA and the
number of previous DMARDs patients were exposed to prior to CSA appeared to be
associated with increased SCr and elevated blood pressure. Previous DMARD use did not
correlate with duration of RA or with age. While we cannot explain this conclusively from
our data, we hypothesize that the number of previous DMARDs may be a surrogate marker
for the number or duration of NSAIDS utilized prior to CSA treatment. NSAIDS are both a
risk factor for the development of hypertension and renal dysfunction in patients with RA.\textsuperscript{18-20} Authors of a recent paper examining the factors that predict drug response in clinical trials
for RA found that, amongst other factors, previous DMARD use was associated with a
reduction in the likelihood of patient improvement (adjusted OR of 0.62).\textsuperscript{21} Therefore, from
our data, it is not clear if the number of previously used DMARDs is a confounder or an
independent risk factor for the development of toxicity from CSA. Of note, an interaction
term between duration of CSA and number of DMARDs was significant. Thus, it would
appear that renal toxicity is less likely to occur in patients who have been on CSA for a long
time and have been on numerous DMARDs. This observation is likely due to the early
discontinuation of CSA in people that are susceptible to nephrotoxicity thus selecting for
individuals who can tolerate the renal effects of this agent.

The time to discontinuation of CSA appeared to follow a specific pattern with a
higher discontinuation rate in the first 24 months of therapy followed by a plateau from 24 to
40 months, and then a more rapid rate of discontinuation after 40 months. This finding is
consistent with the data from the Australian and French experiences.\textsuperscript{13,14} For example, Johns
et al.\textsuperscript{13} found that patients with gastrointestinal disturbances, anxiety, tremors and hirsutism
withdrew early from treatment where later withdrawal was due to more serious toxicity (elevated serum creatinine or hypertension). The authors speculated that these early toxicities could have been managed with a dosage reduction rather than discontinuation. In our cohort, patients with these early toxicities were managed with a reduction in CSA dosage and, thus, many early discontinuations were due to ineffectiveness rather than toxicity. This finding supports the assumption in our study that duration of treatment is a measure of overall drug effectiveness.

Other clinic-based, retrospective analyses of CSA in RA have been published\textsuperscript{13-15} In these, withdrawal because of ineffectiveness occurred in 13% to 36%, while toxicity accounted for 11% to 33% of withdrawals. As in our study, the major toxicities were elevated serum creatinine and hypertension, although one study reported cancer in three patients.\textsuperscript{15} Strengths of our study include the relatively large sample size, the prolonged follow-up, and that it encompasses the entire population of RA patients treated with CSA in a defined geographic region (the province of B.C.). In addition, the same practitioners repeatedly assessed all RA patients in the cohort. Thus, many potential sources of selection bias and assessment were avoided. Our analysis contains patients treated with a combination MTX plus CSA and patients treated with CSA monotherapy such that comparisons of continuation time can be made. The utilization of Cox regression enabled the identification of determinants of CSA discontinuation and the use of GEE models allowed for the adjustment of repeated, correlated measures for each subject rather than analyzing the data at single points in time. The GEE regression models also allowed for the determination of variables significantly associated with favorable response and toxicity to CSA.
Although the data were collected prospectively, a limitation was that the analysis was completed retrospectively. Information that might have been of interest such as standardized measures of disability or longitudinal radiographic assessments were not available. Although we observed statistically significant differences between those patients treated with the combination of MTX and CSA and those on CSA monotherapy in terms of continuation time and clinical outcomes, treatment assignment was not random. Nonetheless, our results are consistent with those from a randomized trial and may augment the generalisability of the trial results. Finally, the cost-effectiveness of CSA relative to other DMARDs has been determined based upon the results of clinical trials with short follow-up. However, the long-term cost-effectiveness of CSA is still unknown and will be a focus of further studies.

In summary, CSA appears to be both safe and effective for long-term use in patients with severe RA who have failed on multiple other therapies. Combining CSA with MTX both prolongs the duration of CSA therapy and reduces the number of affected joints. CSA was reasonably well tolerated with no irreversible adverse events although a longer duration of CSA therapy was associated with the development of renal toxicity and increased DBP. The number of previous DMARDs used prior to CSA appears to be a determinant for the development of hypertension and renal dysfunction although it is unclear if this parameter is a confounder or an independent risk factor.
REFERENCES


### TABLE 1: RESULTS OF THE MULTIVARIATE ANALYSES

<table>
<thead>
<tr>
<th>Dependent Variable/Outcome Variables</th>
<th>Parameter estimate</th>
<th>95%CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint Count (divided by 10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>18.17*</td>
<td>14.73 to 22.42*</td>
</tr>
<tr>
<td>Time on CSA (months)</td>
<td>0.99*</td>
<td>0.98 to 0.99*</td>
</tr>
<tr>
<td>MTX combination (y/n)</td>
<td>0.79*</td>
<td>0.63 to 0.99*</td>
</tr>
<tr>
<td>Morning stiffness (minutes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>166.8</td>
<td>115.3 to 218.4</td>
</tr>
<tr>
<td>Time on CSA (months)</td>
<td>-2.0</td>
<td>-3.0 to -1.1</td>
</tr>
<tr>
<td>Systolic blood pressure (mmHg)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>75.9</td>
<td>61.0 to 90.8</td>
</tr>
<tr>
<td># of previous DMARDs</td>
<td>7.2</td>
<td>2.7 to 11.6</td>
</tr>
<tr>
<td>Age (years)</td>
<td>0.60</td>
<td>0.45 to 0.74</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>0.30</td>
<td>0.12 to 0.48</td>
</tr>
<tr>
<td>Diastolic blood pressure (mmHg)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>65.1</td>
<td>54.2 to 67.5</td>
</tr>
<tr>
<td>Time on CSA (months)</td>
<td>0.07</td>
<td>0.02 to 0.09</td>
</tr>
<tr>
<td># of previous DMARDs</td>
<td>3.8</td>
<td>3.0 to 6.4</td>
</tr>
<tr>
<td>Age (years)</td>
<td>0.07</td>
<td>0.07 to 0.19</td>
</tr>
<tr>
<td>Weight</td>
<td>0.16</td>
<td>0.10 to 0.22</td>
</tr>
<tr>
<td>Serum creatinine (μmol/L)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>31.1</td>
<td>16.6 to 45.6</td>
</tr>
<tr>
<td>Time on CSA (months)</td>
<td>0.63</td>
<td>0.42 to 0.84</td>
</tr>
<tr>
<td>Duration of RA (years)</td>
<td>-0.46</td>
<td>-0.77 to -0.15</td>
</tr>
<tr>
<td>Age (years)</td>
<td>0.65</td>
<td>0.45 to 0.85</td>
</tr>
<tr>
<td>Sex (male as reference group)</td>
<td>-14.8</td>
<td>-20.9 to -8.7</td>
</tr>
<tr>
<td># of previous DMARDs</td>
<td>34.8</td>
<td>22.1 to 47.7</td>
</tr>
<tr>
<td>Interaction between Time on CSA and</td>
<td></td>
<td></td>
</tr>
<tr>
<td># of previous DMARDs</td>
<td>-0.36</td>
<td>-0.47 to -0.26</td>
</tr>
</tbody>
</table>

* For joint count, the parameter estimates are rate ratios (RR) as the log-link was used. All other models are linear as the identity link was used.
FIGURE 1: SURVIVAL CURVE FOR CSA DISCONTINUATION

The crosses (+) represent those individuals who were censored.
FIGURE 2: SURVIVAL CURVE FOR CSA DISCONTINUATION FOR THOSE ON CSA ALONE (SOLID LINE) AND COMBINATION CSA/MTX (DOTTED LINE)

The crosses (+) represent those individuals who were censored.
APPENDIX IV

RHEUMATOID ARTHRITIS ASSESSMENT QUESTIONNAIRE: BASELINE
(SELF-ADMINISTERED)
SECTION I: RHEUMATOID ARTHRITIS and HEALTH CARE USE ASSESSMENT

You may decline to answer any question, however please remember that it is very important that we get the most accurate and complete information we can, and that all the information you provide is completely confidential.

HC1. What is your current marital status?
- Single
- Married
- Married and separated
- Common law
- Divorced
- Widowed

HC2. What type of health insurance coverage do you have?
[Please check all that apply]
- I don't currently have medical insurance
- Plan C (Social assistance)
- Plan E (Basic MSP) - self paid
- Plan E (Basic MSP) - employer paid
- Extended medical - self paid
- Extended medical - employer paid
- Prescription drug plan (3rd party coverage)
- Other

If OTHER, please specify:...........................................  

HC3. When were you first diagnosed (by a rheumatologist) with rheumatoid arthritis?

Appropriate date of rheumatoid arthritis diagnosis [month/year]
- I don't know
- I prefer not to answer this question

HC4. Do you currently, or have you previously, smoked cigarettes, cigars, or a pipe?
- Never smoked .............................................. [Go to Question HC7]
- Currently smoke
- Quit smoking
- Other
- I prefer not to answer this question
HC5. If you have quit smoking, how long has it been since you last smoked?
   • < 3 months
   • 3 - 6 months
   • 6 - 12 years
   • 1 - 5 years
   • > 5 years
   • I don't know

HC6. How much do you, or did you previously smoke?

   [ ] Amount smoked [per day]
   • I don't know
   • I prefer not answer this answer

HC7. Over the past year have you been admitted to hospital due to your rheumatoid arthritis
   (ie. Joint surgeries)?
   • Yes; How many times?..............................
   • No.
   • I don't know
   • I prefer not to answer this question

   [If NO, go to Question HC9]

HC8. What was the total number of days you spent in the hospital because of your rheumatoid arthritis
   (ie. joint surgeries) in the previous year?
   [ ] Total number of hospital days due to rheumatoid arthritis in the previous year
   • I don't know
   • I prefer not to answer this question

HC9. Over the past year, have you required any other services for your rheumatoid arthritis such as
   physiotherapy, occupational therapy, social work, diet/nutrition counselling, or in-home services
   (e.g. home care)?
   • Yes
   • No.............................................. [Go to Question HC11]
   • I don't know
   • I prefer not to answer this question
HC10. Could you specify the type of service(s) (e.g. physiotherapy, home care), the nature of the actual service(s) provided (e.g. physical conditioning, diet counseling) and number of visits?

<table>
<thead>
<tr>
<th>Type of service</th>
<th>Nature of Service</th>
<th>Number of visits</th>
</tr>
</thead>
</table>

HC11. Over the past year, have you had to rent or purchase any equipment (e.g. wheelchair, kitchen aids) related to your rheumatoid arthritis?

- Yes
- No
- I don't know
- I prefer not to answer this question

HC12. Can you describe this equipment and its cost?

<table>
<thead>
<tr>
<th>Type</th>
<th>Estimated cost (total per month)</th>
<th>Rented</th>
<th>Purchased</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

HC13. Has it ever been recommended that you should use one of these pieces of equipment described above, but you decided not to purchase or rent it?

- Yes
- No
- I don't know
- I prefer not to answer this question

HC14. What type or equipment was it, and why did you decide not to acquire it?

- I didn't think it would help
- I couldn't find one
- It was too expensive
- I don't know
- Other

If OTHER, please specify: ..............................................
HC15. Over the past year, have you utilized any complementary methods of health care (e.g., herbal medications, homeopathic medications, acupuncture, healing touch) for the management of your rheumatoid arthritis?

☐ Yes

☐ No. ........................................................................... [Go to Section II]

☐ I don't know

☐ I prefer not to answer this question

HC16. Can you list these complementary methods of care and their estimated cost to you over the last year?

<table>
<thead>
<tr>
<th>Type</th>
<th>Estimated cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>$</td>
</tr>
<tr>
<td>2.</td>
<td>$</td>
</tr>
<tr>
<td>3.</td>
<td>$</td>
</tr>
<tr>
<td>4.</td>
<td>$</td>
</tr>
<tr>
<td>5.</td>
<td>$</td>
</tr>
</tbody>
</table>
SECTION II: This section is made up of questions asking about yourself and anyone who might be living with you. This involves questions about education, income, and employment of yourself and anyone who might be sharing expenses with you.

P1. What was your main activity during the past 12 months? [Please check only one]
   □ Working at a job
   □ Looking for work
   □ Unable to work due to health reasons
   □ Going to school
   □ Keeping house
   □ Retired
   □ Other

If OTHER, please specify: .........................................................

P2. If your main activity was WORKING AT A JOB, what is your field of employment (e.g. secretary, nurse, laborer, store clerk, plumber, dentist)?

Brief description: .................................................................

P3. If you worked, even if your main activity wasn't WORKING AT A JOB, on average approximately how much did you work over the past 12 months?
   □ Full time (>35 hours / week)
   □ 30 - 35 hours / week
   □ 20 - 29 hours / week
   □ 10 - 19 hours / week
   □ < 10 hours / week
   □ Casual

P4. If someone in your field of employment lost the income from a missed work day, what would be the value of ONE day's lost income (before deductions)?

$ __________ Value of one day's lost income

□ I don't know
□ Not applicable
□ I prefer not to answer this question

P5. Over the past year, have you had to miss work or school because of your rheumatoid arthritis?
   □ Yes
   □ No ................................................................. [Go to Question P7]
   □ Not applicable
P6. Can you estimate how many days of work and/or school that you have missed over the past two weeks and the past year because of your rheumatoid arthritis?

<table>
<thead>
<tr>
<th></th>
<th>Over the past TWO WEEKS</th>
<th>Over the past 12 MONTHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Days of WORK missed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Days of SCHOOL missed</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P7. Over the past year, has anyone (e.g. spouse, partner, caregiver, friend) had to miss work or school because of your rheumatoid arthritis?

- [ ] Yes
- [ ] No

[Go to Question P9]

- [ ] I prefer not to answer this question

P8. Can you specify the relationship of these individuals to you, and how many days of WORK/SCHOOL they have missed in the past year because of your rheumatoid arthritis.

<table>
<thead>
<tr>
<th>Person</th>
<th>Relationship</th>
<th>Work days Missed</th>
<th>School days Missed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person 3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P9. Have you ever had to change jobs because of your rheumatoid arthritis?

- [ ] Yes; What was your previous job?
- [ ] No

P10. What is the highest grade (or year) of secondary (high school) or elementary school you have successfully completed?

- [ ] Number (1 - 13) of grades of secondary and/or elementary schools successfully completed
- [ ] Never attended school, or attended kindergarten only
- [ ] I prefer not to answer this question

P11. How many years of post-secondary education (after high school) have you completed?

- [ ] None
- [ ] Less than one year
P12. What certificates, degrees, or diplomas have you ever obtained? [Please check all that apply]

☐ None
☐ High school diploma
☐ Trades or non-university diploma
☐ Undergraduate (bachelor's) university diploma
☐ Degree in medicine, dentistry, vet medicine, optometry, chiropractic, etc.
☐ Master or Doctorate degree (M.A., M.Sc., Ph.D., D.Sc., D.Ed.)

P13. Is there another primary, non-dependent adult living in your household (e.g. a spouse or partner)?

☐ Yes
☐ No [Go to Question P21]

P14. What is this person's relationship to you?

☐ Spouse / Partner
☐ Partner / Guardian
☐ Sibling
☐ Roomate
☐ Other

If OTHER, please specify: ..................................................

P15. What was this person's main activity during the last 12 months?

☐ Working at a job
☐ Looking for work
☐ Unable to work due to health reasons
☐ Going to school
☐ Keeping house
☐ Retired
☐ Other
☐ I don't know
☐ I prefer not to answer this question

If OTHER, please specify: ..................................................

P16. If this person's main activity was WORKING AT A JOB, what is their field of employment (e.g. secretary, nurse, laborer, store clerk, plumber, dentist)?

Brief description ...............................
P17. If someone in this person's field of employment lost the income from a missed work day, what would be the value of ONE days lost income (before deductions)?

Value of one days lost income

- I don't know
- Not applicable
- I prefer not to answer this question

P18. What is the highest grade (or year) of secondary (high school) or elementary school this person has successfully obtained?

Number (1 - 13) of grades of secondary and/or elementary school successfully completed

- Never attended school, or attended kindergarten only
- I don't know
- I prefer not to answer this question

P19. How many years of post-secondary education (after high school) has this person completed?

Number of years of post-secondary education

- None
- Less than one year
- I don't know
- I prefer not to answer this question

P20. What certificates, degrees, or diplomas have they ever obtained?

[Please check all that apply]

- None
- High school diploma
- Trades or non-university certificate
- Undergraduate (bachelor's) university degree
- Degree in medicine, dentistry, optometry, chiropractic, etc.
- Masters or doctorate degree
- I don't know
- I prefer not to answer this question
P21. What was your approximate total household income from all sources for the previous year, before income tax deduction? [Including only your family members, not roommates that you don't share daily expenses with]

- less than $20,000
- $20,000 - $30,000
- $30,001 - $40,000
- $40,001 - $50,000
- $50,001 - $60,000
- $60,001 - $70,000
- $70,001 - $80,000
- $80,001 - $90,000
- $90,001 - $100,000
- $100,001 - $110,000
- greater than $100,000
- I don't know
- I prefer not to answer this question

P22. Do you have access to an automobile?

- Yes
- No

If YES: How many automobiles do you have?

- Total number of automobiles

P23. Do you have any children?

- Yes
- No

If YES: How many children do you have?

- Total number of children

How many children currently live with you?

- Total number of children currently living with you

P24. How would you describe the type of dwelling that you currently live in?

- Single detached home
- Duplex or Townhouse
- Apartment or Condominium
- Mobile home
- Boarding room / Hotel / Rooming house
- Other
- I prefer not to answer this question

If OTHER, please specify: ..................

365
P25. How many rooms does your current residence have, including kitchen, living room, bedrooms, finished rooms in the basement or attic, etc? (Do not count bathrooms, hallways, or rooms used exclusively for work).

[ ] Total number of rooms

P26. How many of these rooms are bedrooms?

[ ] Number of bedrooms

P27. How many adults and children (including yourself) live in your current residence?

[ ] Total number of adults

[ ] Total number of children

P28. Is your current residence:

- [ ] Owned by you or a member of household
- [ ] Rented [even if no rent is paid]
- [ ] Subsidized housing (you receive government rental assistance)
- [ ] Other
- [ ] I don't know
- [ ] I prefer not to answer this question

If OTHER, please specify: ..........................
SECTION III: RHEUMATOID ARTHRITIS MEDICATION USE

It is very important that we get the most accurate information that we possibly can. Some of the questions ask for very specific details, and deal with some things that may have happened over the past year, so please take your time and try and answer the questions as accurately as possible. Once again, please remember that any answers you give are completely confidential.

M1. Have you ever received a prescription for a rheumatoid arthritis medication from your doctor that you have not had filled out by a pharmacist?
   □ Yes; How many times? ................. [ ]
   □ No
   □ I don't know
   □ I prefer not to answer this question

[ If NO, go to Question M3 ]

M2. What was your reason for not filling the prescription?
   □ It was too expensive
   □ I didn't think I needed it
   □ I didn't think it would help
   □ I couldn't get to the pharmacy
   □ Other

If OTHER, please specify: .....................

M3. Other than your rheumatoid arthritis, do you have any other chronic diseases (such as high blood pressure, arthritis, diabetes, angina, depression) that have been diagnosed by your doctor?
   □ Yes
   □ No ...................................................  [Go to Question M5]
   □ I don't know
   □ I prefer not to answer this question

M4. If you have other chronic diseases in addition to your rheumatoid arthritis, can you list these diseases?

<table>
<thead>
<tr>
<th>1.</th>
<th>2.</th>
<th>3.</th>
<th>4.</th>
<th>5.</th>
<th>6.</th>
<th>7.</th>
<th>8.</th>
</tr>
</thead>
</table>

367
M7. Did you stop taking any ARTHRITIS MEDICATIONS during the PAST 3 MONTHS, regardless of reason?
- Yes
- No [Go to Question M8]

M7 (continued) IF YES, please complete ALL THE BLANKS ON THE LINE for any medications that you have stopped and tell us about the medicine you are taking instead (These medications should also be listed in question M6).

<table>
<thead>
<tr>
<th>Name of Medication You Stopped</th>
<th>If Stopped, Why?</th>
<th>Did You Start Another Medication to Replace it?</th>
<th>If Yes, Which Medication?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>[ ] Yes [ ] No</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[ ] Yes [ ] No</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[ ] Yes [ ] No</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[ ] Yes [ ] No</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[ ] Yes [ ] No</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[ ] Yes [ ] No</td>
<td></td>
</tr>
</tbody>
</table>

M8. Overall, how would you describe the severity of your rheumatoid arthritis? [Please check one]
- [ ] Very Mild
- [ ] Mild
- [ ] Moderate
- [ ] Severe
- [ ] Very Severe

M9. Overall, how would you classify the control of your rheumatoid arthritis? [Please check one]
- [ ] Very Well Controlled
- [ ] Well Controlled
- [ ] Adequately Controlled
- [ ] Not Well Controlled
- [ ] Not Controlled At All

M10. Below is a line with '0' at the left-hand end and a '1' at the right-hand end. The '0' represents "death", the '1' represents "perfect health", and the area in between represents a state of health somewhere in between. Make a mark on the line at the point that you feel represents how you feel today.

0

Death

1

Perfect

Health

368
SECTION IV: DRUG SIDE EFFECT

It is very important that we get the most accurate information that we possibly can. Some of the questions ask for very specific details, and deal with some things that may have happened over the three months, so please take your time and try and answer the questions as accurately as possible. Once again, please remember that any answers you give are completely confidential.

S1. Over the PAST THREE MONTHS have you had any side effects from your rheumatoid arthritis medications?

☐ Yes
☐ No [Go to SECTION V]

If YES, complete the rest of this section.

DIRECTIONS
1. Write in the name of the drug causing the side effect(s).
2. Indicate whether you stopped the drug.
3. List side effect(s) for each drug. Please list any abnormal laboratory findings such as low white blood count, protein in urine, low platelets, kidney problems, anemia, liver problems.
4. Check the severity of each side effect.
5. Please indicate how important the side effect was to you by making a mark on the scale from 0 to 1, where 0 is "Not at all" and 1 is "Very Much".
6. If you need more room, please use the back of this questionnaire.

A.
1. Drug Name: 

2. Did you stop taking the drug because of a side effect?

☐ Yes
☐ No

3. List side effect 

4. Severity of side effect
   ☐ Mild
   ☐ Moderate
   ☐ Severe

5. How important was this side effect TO YOU?

0
Not at all

1
Very
Much
B.
1. Drug Name: 

2. Did you stop taking the drug because of a side effect?
   □ 
   □ 

3. List side effect 

4. Severity of side effect
   □ Mild
   □ Moderate
   □ Severe

5. How important was this side effect TO YOU?

0 Not at all

1 Very Much

---

C.
1. Drug Name: 

2. Did you stop taking the drug because of a side effect?
   □ Yes
   □ No

3. List side effect 

4. Severity of side effect
   □ Mild
   □ Moderate
   □ Severe

5. How important was this side effect TO YOU?

0 Not at all

1 Very Much

370
SECTION V: RHEUMATOID ARTHRITIS CLINICAL MEASURES

This section asks questions regarding your rheumatoid arthritis that are commonly determined to see how your disease is being controlled. You may decline to answer any questions, however please remember that it is very important that we get the most accurate and complete information we can, and that all the information you provide is completely confidential.

C1. Below is a line with '0' at the left-hand end and a '1' at the right-hand end. The '0' represents "VERY WELL", the '1' represents "VERY POOR", and the area in between represents a state of health somewhere in between. Considering all the ways that your arthritis affects you, rate how you are doing on the following scale by placing a mark on the line.

```
<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Well</td>
<td>Very Poor</td>
</tr>
</tbody>
</table>
```

C2. Below is a line with '0' at the left-hand end and a '1' at the right-hand end. The '0' represents "NO PAIN", the '1' represents "SEVERE PAIN", and the area in between represents a state of pain somewhere in between. Make a mark on the line at the point that you feel represents how much pain have you had because of your arthritis IN THE PAST WEEK?

```
<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Pain</td>
<td>Severe Pain</td>
</tr>
</tbody>
</table>
```

C3. The following two pages are pictures of mannequins to help us determine your number of tender/painful joints (FIRST MANNEQUIN) and swollen joints (SECOND MANNEQUIN). You may decline to answer any question, however please remember that it is very important that we get the most accurate and complete information we can, and that all the information you provide is completely confidential.
Please indicate with an “X” in the circles below, any joints which are **PAINFUL** or **TENDER** at present. To test for pain, move your joints in a full range of motion and then squeeze the joint between your thumb and forefinger.
To test for **SWELLING**, notice if your joint appears larger or bulging and squeeze to see if it feels like it is full of liquid or like a sponge. Indicate with an “X” in the circles below any joints which are swollen at present.
Section VI

Please choose the response that applies best to you at this time. Check only one box for each statement.

1. I have to go to bed earlier than I would like to
   Yes □ No □

2. I'm afraid of people touching me
   Yes □ No □

3. It's difficult to find comfortable shoes that I like
   Yes □ No □

4. I avoid crowds because of my arthritis
   Yes □ No □

5. I have difficulty getting dressed
   Yes □ No □

6. I find it difficult walking from store to store
   Yes □ No □

7. Household chores take me a long time
   Yes □ No □

8. I sometimes have problems using the toilet
   Yes □ No □

9. I often get frustrated
   Yes □ No □

10. I frequently have to stop what I am doing to rest
    Yes □ No □

11. I have difficulty using a knife and fork
    Yes □ No □
Please choose the response that applies best to you at this time.

12. I find it hard to concentrate
   - Yes □
   - No □

13. Sometimes I just want to be left alone
   - Yes □
   - No □

14. I have difficulty walking very far
   - Yes □
   - No □

15. I try to avoid shaking hands with people
   - Yes □
   - No □

16. I often get depressed
   - Yes □
   - No □

17. I'm unable to join in activities with my family or friends
   - Yes □
   - No □

18. I have difficulty taking a bath/shower
   (Please answer for the one you usually use)
   - Yes □
   - No □

19. Sometimes I have a good cry because of my arthritis
   - Yes □
   - No □

20. My arthritis limits the places I can go
   - Yes □
   - No □

21. Any amount of activity I do makes me feel tired
   - Yes □
   - No □

22. I feel dependent on others
   - Yes □
   - No □

23. My arthritis is constantly on my mind
   - Yes □
   - No □

24. I often get angry with myself
   - Yes □
   - No □
Please choose the response that applies best to you at this time.

25. It's too much effort to go out and see people
   - Yes □
   - No □

26. I sleep poorly at night
   - Yes □
   - No □

27. I find it difficult to take care of the people I am close to
   - Yes □
   - No □

28. I feel unable to control my arthritis
   - Yes □
   - No □

29. I avoid physical contact
   - Yes □
   - No □

30. I'm limited in the clothes I can wear
   - Yes □
   - No □
This survey asks for your views about your health. This information will help you keep track of how you feel and how well you are able to do your usual activities.

Answer every question by selecting the answer as indicated. If you are unsure about how to answer a question, please give the best answer you can.

Please check the level that you feel represents your health in the following categories. (ie. Physical Functioning, Role Limitations, Social Functioning, Pain, Mental Health, Vitality)

### Physical Functioning

- Your health does not limit you in vigorous activities
- Your health limits you a little in vigorous activities
- Your health limits you a little in moderate activities
- Your health limits you a lot in moderate activities
- Your health limits you a little in bathing and dressing
- Your health limits you a lot in bathing and dressing

### Role Limitations

- You have no problems with your work or other regular daily activities as a result of your physical health or any emotional problems
- You are limited in the kind of work or other activities as a result of your physical health
- You accomplish less than you would like as a result of emotional problems
- You are limited in the kind of work or other activities as a result of your physical health and accomplish less than you would like as a result of emotional problems

### Social Functioning

- Your health limits your social activities none of the time
- Your health limits your social activities a little of the time
- Your health limits your social activities some of the time
- Your health limits your social activities most of the time
- Your health limits your social activities all of the time

### Pain

- You have no pain
- You have pain but it does not interfere with your normal work (both outside the home and housework)
- You have pain that interferes with your normal work (both outside the home and housework) a little bit
- You have pain that interferes with your normal work (both outside the home and housework) moderately
- You have pain that interferes with your normal work (both outside the home and housework) quite a bit
- You have pain that interferes with your normal work (both outside the home and housework) extremely

### Mental Health

- You feel tense or downhearted and low none of the time
- You feel tense or downhearted and low a little of the time
- You feel tense or downhearted and low some of the time
- You feel tense or downhearted and low most of the time
- You feel tense or downhearted and low all of the time

### Vitality

- You have a lot of energy all of the time
- You have a lot of energy most of the time
- You have a lot of energy some of the time
- You have a lot of energy a little of the time
- You have a lot of energy none of the time
EUROQOL (Page 1 of 2)

Your own health state today

By placing a tick in one box in each group below, please indicate which statements best describe your own health state 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 a health state 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 health state is.
**Health Utilities Index Mark 2 and Mark 3**

Since the HUI2 and HUI3 are proprietary, rather than the full instruments, the attributes and levels for each scale (with a description) has been included below.

**Health status classification system: HUI2**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SENSORY</td>
<td>1</td>
<td>Able to see, hear and speak normally for age</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Requires equipment to see or hear or speak</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Sees, hears, or speaks with limitations even with equipment</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Blind, deaf or mute</td>
</tr>
<tr>
<td>MOBILITY</td>
<td>1</td>
<td>Able to walk, bend, lift, jump and run normally for age</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Walks, bends, lifts, jumps or runs with some limitations but does not require help</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Requires mechanical equipment (such as canes, crutches, braces or wheelchair) to walk or get around independently</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Requires the help of another person to walk or get around and requires mechanical equipment as well</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Unable to control or use arms and legs</td>
</tr>
<tr>
<td>EMOTION</td>
<td>1</td>
<td>Generally happy and free from worry</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Occasionally fretful, angry, irritable, anxious, depressed, or suffering &quot;night terrors&quot;</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Often fretful, angry, irritable, anxious, depress or suffering &quot;night terrors&quot;</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Almost always fretful, angry, irritable, anxious, depressed</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Extremely fretful, angry, irritable or depressed usually requiring hospitalization or psychiatric institutional care</td>
</tr>
<tr>
<td>COGNITIVE</td>
<td>1</td>
<td>Learns and remembers school work normally for age</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Learns and remembers school work more slowly than classmates as judged by parents and/or teachers</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Learns and remembers very slowly and usually requires special educational assistance</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Unable to learn and remember</td>
</tr>
<tr>
<td>SELF-CARE</td>
<td>1</td>
<td>Eats, bathes, dresses and uses the toilet normally for age</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Eats, bathes, dresses or uses the toilet independently with</td>
</tr>
<tr>
<td></td>
<td>PAIN</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>----------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Free of pain and discomfort</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Occasional pain. Discomfort relieved by non-prescription drugs or self-control activity without disruption of normal activities</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Frequent pain. Discomfort relieved by oral medicines with occasional disruption of normal activities</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Frequent pain; frequent disruption of normal activities. Discomfort requires prescription narcotics for relief</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Severe pain. Pain not relieved by drugs and constantly disrupts normal activities</td>
<td></td>
</tr>
</tbody>
</table>

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
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.
<table>
<thead>
<tr>
<th>COGNITION</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Able to remember most things, think clearly and solve day to day problems.</td>
</tr>
<tr>
<td>2</td>
<td>Able to remember most things, but have a little difficulty when trying to think and solve day to day problems.</td>
</tr>
<tr>
<td>3</td>
<td>Somewhat forgetful, but able to think clearly and solve day to day problems.</td>
</tr>
<tr>
<td>4</td>
<td>Somewhat forgetful, and have a little difficulty when trying to think or solve day to day problems.</td>
</tr>
<tr>
<td>5</td>
<td>Very forgetful, and have great difficulty when trying to think or solve day to day problems.</td>
</tr>
<tr>
<td>6</td>
<td>Unable to remember anything at all, and unable to think or solve day to day problems.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PAIN</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Free of pain and discomfort.</td>
</tr>
<tr>
<td>2</td>
<td>Mild to moderate pain that prevents no activities</td>
</tr>
<tr>
<td>3</td>
<td>Moderate pain that prevents a few activities.</td>
</tr>
<tr>
<td>4</td>
<td>Moderate to severe pain that prevents some activities.</td>
</tr>
<tr>
<td>5</td>
<td>Severe pain that prevents most activities.</td>
</tr>
</tbody>
</table>

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>
**Health Assessment Questionnaire**

Please place an "X" in the box [ ] which best describes your usual abilities OVER THE PAST WEEK:

<table>
<thead>
<tr>
<th>DRESSING &amp; GROOMING</th>
<th>WITHOUT ANY DIFFICULTY</th>
<th>WITH SOME DIFFICULTY</th>
<th>WITH MUCH DIFFICULTY</th>
<th>UNABLE TO DO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are you able to:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dress yourself,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>including shoelaces</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>and buttons?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shampoo your hair?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| ARISING             |                        |                      |                      |              |
| Are you able to:    |                        |                      |                      |              |
| Stand up from a     |                        |                      |                      |              |
| straight chair?     |                        |                      |                      |              |
| Get in and out of   |                        |                      |                      |              |
| bed?                |                        |                      |                      |              |

| EATING              |                        |                      |                      |              |
| Are you able to:    |                        |                      |                      |              |
| Cut your meat?      |                        |                      |                      |              |
| Lift a full cup or  |                        |                      |                      |              |
| glass to your mouth?|                        |                      |                      |              |
| Open a new milk     |                        |                      |                      |              |
| carton?             |                        |                      |                      |              |

| WALKING             |                        |                      |                      |              |
| Are you able to:    |                        |                      |                      |              |
| Walk outdoors on    |                        |                      |                      |              |
| flat ground?        |                        |                      |                      |              |
| Climb up five steps?|                        |                      |                      |              |

Please check any AIDS OR DEVICES that you usually use for any of the above activities:

- Devices Used for Dressing (button hook, zipper pull, etc.)
- Built up or special utensils
- Crutches
- Walker
- Wheelchair
- Cane

Please check any categories for which you usually need HELP FROM ANOTHER PERSON:

- Dressing and Grooming
- Arising
- Eating
- Walking

385
Please place an "X" in the box □ which best describes your usual abilities OVER THE PAST WEEK:

<table>
<thead>
<tr>
<th>HYGIENE</th>
<th>WITHOUT ANY DIFFICULTY</th>
<th>WITH SOME DIFFICULTY</th>
<th>WITH MUCH DIFFICULTY</th>
<th>UNABLE TO DO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wash and dry your body?</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>Take a tub bath?</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>Get on and off the toilet?</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>REACH</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Reach and get down a 5 pound object (such as a bag of sugar) from above your head?</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>Bend down to pick up clothing from the floor?</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GRIP</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Open car doors?</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>Open previously opened jars?</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>Turn faucets on and off?</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ACTIVITIES</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Run errands and shop?</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>Get in and out of a car?</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>Do chores such as vacuuming or yard work?</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

Please check any AIDS OR DEVICES that you usually use for any of these activities:

- [ ] Raised toilet seat
- [ ] Bathtub bar
- [ ] Long-handled appliances for reach
- [ ] Bathtub seat
- [ ] Long-handled appliances in bathroom
- [ ] Jar opener (for jars previously opened)

Please check any categories for which you usually need HELP FROM ANOTHER PERSON:

- [ ] Hygiene
- [ ] Reach
- [ ] Gripping and opening things
- [ ] Errands and chores
APPENDIX V

RHEUMATOID ARTHRITIS ASSESSMENT QUESTIONNAIRE:
3 MONTHS (SELF-ADMINISTERED)

The content of the 3 month questionnaire was the same as the baseline questionnaire (from Section III onwards) and only the first page (see below) was added.
1. Overall, how would you describe changes in your rheumatoid arthritis since answering our LAST questionnaire (about 3 months ago)?

[Please check one]

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Much Worse</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Somewhat Worse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Same</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Somewhat Better</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Much Better</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX VI

RHEUMATOID ARTHRITIS ASSESSMENT QUESTIONNAIRE:
6 MONTHS (SELF-ADMINISTERED)

The content of the 6 month questionnaire was the same as the baseline questionnaire (from Section III onwards) and only the first page (see below) was added.
Please print within the specified boxes (if possible) and mark all checked boxes with an "X"

1. Overall, how would you describe changes in your rheumatoid arthritis since answering our LAST questionnaire (about 3 months ago)?

   [Please check one]

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Much Worse</td>
<td>Somewhat Worse</td>
<td>The Same</td>
<td>Somewhat Better</td>
<td>Much Better</td>
</tr>
</tbody>
</table>

2. Overall, how would you describe changes in your rheumatoid arthritis since answering our FIRST questionnaire (about 6 months ago)?

   [Please check one]

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Much Worse</td>
<td>Somewhat Worse</td>
<td>The Same</td>
<td>Somewhat Better</td>
<td>Much Better</td>
</tr>
</tbody>
</table>