Exploring the applicability of the minimally important difference concept in interpreting generic indirect preference-based health-related quality of life outcomes

by

Nathan S. McClure

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Public Health

School of Public Health
University of Alberta

Abstract

Generic indirect preference-based measures of health-related quality of life (HRQL), such as the EQ-5D-5L index score, were developed to assign health preference weights to time lived in different health states in order to calculate quality-adjusted life years (QALYs), a common outcome measure for economic evaluations. A key feature of the index score like the EQ-5D-5L is the anchoring of the scale at 1.0 for full health and 0.0 for dead. Since everyone is a potential patient whose condition and treatment needs are uncertain, it is recommended that preference weights reflect the average of the general population's health preferences. However, this masks underlying preference heterogeneity, which presents a challenge for end-users of generic HRQL instruments who seek to understand whether patients' HRQL has improved or worsened over time. Thus, the concept of the minimally important difference (MID), defined as the smallest change in EQ-5D-5L index score that can be expected to reflect minimally important improvement or deterioration in patients' HRQL, may be a useful way of interpreting observed index score changes.

Plausible MID estimates for EQ-5D-5L index scores range from 0.037 to 0.069 depending on the country-specific scoring algorithm and baseline score. For patients with type 2 diabetes, as an example of a chronic condition in the general population, MID estimates for index scores based on multiple approaches ranged from 0.03 to 0.05, and were further found to vary by direction of change. Secondary analysis of responses from the Canadian Valuation Study showed how the variability in individual-level interpretation of (small) index score differences resulted in a health state transition having to be 'large enough' (i.e., > +/- 0.05 change in index score) to be meaningfully interpretable as an improvement or deterioration in HRQL. Based on evidence of MIDs, a new method for calculating QALYs was proposed adjusting for meaningful within-patient change in HRQL. Comparing incremental QALY estimates using different methods in a case study for depression treatment in patients with type 2 diabetes showed how the method of QALY calculation and adjustment for between-group differences yields different results (ranging from -0.028 to 0.031). The uncertainty in incremental QALY estimates reflects uncertainty in regards to the value of small EQ-5D-5L index score changes.

This research found MID estimates that reflect greater than zero change in EQ-5D-5L index score, suggesting that observed index score changes smaller than the MID do not adequately represent HRQL improvement or deterioration from the patients' perspective. Therefore, the MID may be useful in determining whether or not the observed index score change is expected to represent meaningful change in patients' HRQL. In doing so, there is an explicit incorporation of patients' HRQL in the interpretation of HRQL outcomes from generic indirect preference-based measures.

Preface

I (NSM) hold primary responsibility for all aspects of the research included in this thesis and should be considered lead author for all chapters. I was responsible for the conceptualization of the studies, conducted all analyses and prepared all initial drafts.

This thesis involves secondary-use of data collected in the Alberta's Caring for Diabetes Cohort Study (Chapter 3), Canadian Valuation Study (Chapter 4), and TeamCare-PCN Trial (Chapter 5). The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board, Project Name "ABCD Team Care Depression Intervention Re-Analysis of Cost-Effectiveness Study", No. Pro00087637, January 7, 2020.

Chapter 2 of this thesis has been published as McClure NS, Al Sayah F, Xie F, Luo N, Johnson JA. Instrument-defined estimates of the minimally important difference for EQ-5D-5L index scores. Value Health. 2017;20:644-650.

Chapter 3 of this thesis has been published as McClure NS, Al Sayah F, Ohinmaa A, Johnson JA. Minimally Important Difference of the EQ-5D-5L Index Score in Adults with Type 2 Diabetes. Value Health. 2018;21:1090-1097.

JAJ is the supervisory author for Chapters 2, 3, 4 and 5. JAJ was the investigator for the Alberta's Caring for Diabetes Cohort Study and TeamCare-PCN trial. He provided guidance on concept formation and interpretation of results; he critically reviewed the content of this thesis and contributed to edits.

AO is co-author for Chapters 3, 4 and 5. AO was a member of NSM's supervisory committee. He provided guidance and critically reviewed all Chapters, contributing to edits therein.

MP is co-author for Chapters 4 and 5. MP was a member of NSM's supervisory committee. He provided guidance and critically reviewed all Chapters, contributing to edits therein.

FAS is secondary author for Chapters 2 and 3. FAS is the custodian of the Alberta's Caring for Diabetes Cohort Study. She provided guidance and critically reviewed the content of Chapters 2 and 3, contributing to edits therein.

FX is co-author for Chapters 2 and 4. FX was the investigator for the Canadian Valuation Study. He critically reviewed the content of Chapters 2 and 4, and contributed to edits therein.

NL is co-author for Chapter 2. He critically reviewed the content of Chapter 2, and contributed to edits therein.

Acknowledgements

I would like to thank the members of my supervisory committee, Drs. Jeffrey A. Johnson, Arto Ohinmaa and Mike Paulden for their mentorship throughout my studies. I greatly appreciate their patience, promptness, and perceptiveness. Jeff, I will always remember taking your Health-Related Quality of Life class. The perspective that you and Dr. Fatima Al Sayah shared helped lay the foundation to this research.

To my peers and colleagues at Alberta PROMs and EQ-5D Research and Support Unit (APERSU), the Alliance for Canadian Health Outcomes Research in Diabetes (ACHORD), the EuroQol Group, the Institute for Health Economics and the School of Public Health, you have been an important influence, and I am grateful for the inspiration and support you have provided me. I will sincerely miss learning from such a diverse group of students, staff, researchers and faculty. May we continue to champion the interdisciplinary nature of Public Health, and find strength in our collective diversity.

To my parents, Dianne and Larry McClure, your unconditional love and dedication to me is immeasurable. You instilled in me an inquisitiveness and passion for learning that has been integral to my accomplishments.

Lastly, I wish to dedicate this work to my loving partner, Lindsay McInnes. Despite the circumstances that have, at times, created physical distance between us, your love and consideration for a life-long learner is always felt close at hand. Your astute ability to balance pragmatism with a keenness for exploring meaning is truly motivational. I am so thankful to have shared this journey with you, and look forward to the many more adventures that follow.

-Nathan S. McClure

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AD	Anxiety/Depression
AUC	Area Under Curve
CADTH	Canadian Agency for Drugs and Technologies in Health
CFB	Change From Baseline
CI	Confidence Interval
cTTO	Composite Time Trade-Off
DCE	Discrete-Choice Experiment
EQ-5D (-3L or -5L)	EuroQol Five Dimensional (Three or Five level) Questionnaire
ES	Effect Size
HRQL	Health-Related Quality of Life
HUS	Health Utility Score
ICUR	Incremental Cost-Utility Ratio
idMID(*)	Instrument Defined Minimally Important Difference (*excluding maximum-valued scoring parameters)
IQR	Interquartile Range
MCID	Minimally Clinically Important Difference
MCS	Mental Health Component Score
MDC	Minimally Detectable Change
MID	Minimally Important Difference
MO	Mobility
oTTO	Ordinal Time Trade-Off
PAID5	Problem Areas in Diabetes Five Items Questionnaire

PAPRIKA	Potentially All Pairwise Rankings of All Pairwise Alternatives
PCS	Physical Health Component Score
PD	Pain/Discomfort
PHQ (8 or 9)	Patient Health Questionnaire (Eight or Nine Items)
PROM	Patient-Reported Outcome Measure
QALY	Quality-Adjusted Life Year
ROC	Receiver Operating Characteristic
SC	Self-Care
SD	Standard Deviation
SF-12	Short-Form Medical Survey Twelve Items
SRM	Standardized Response Mean
TTO	Time Trade-Off
tTTO	Traditional Time Trade-Off
UA	Usual Activities
UK	United Kingdom
VAS	Visual Analogue Scale

Summary

This thesis explored the concept of a minimally important difference (MID) as a method to support the interpretation of health-related quality of life (HRQL) index scores at the point of application. This exploration involved a combination of simulation-based and real-world case studies of HRQL measurement. To this end, the EQ-5D-5L index score based on the average of the Canadian population's preferences, was used as an archetypal HRQL measure due to its widespread adoption, multiple applications, and need for end-user support. In addition, EQ-5D-5L index scores collected from patients with type 2 diabetes are analyzed to support the interpretation of HRQL changes in this target population. Type 2 diabetes was chosen due to its prevalence in the general population, and impact on HRQL.

Chapter 1 provides background relevant to this thesis, placing the issue of HRQL interpretability within the wider objectives of collecting and assessing patient-reported outcomes. A conceptual framework is described, identifying the role of MIDs in HRQL score interpretation. Lastly, research questions and objectives are outlined for each study. **Chapter 2**, the first study of this thesis, presents plausible MID estimates for EQ-5D-5L index scores based on scoring algorithms from different countries. This was considered an 'instrument-defined' approach to MID estimation. **Chapter 3** details the second study, which used multiple methods and anchors to estimate the smallest change in EQ-5D-5L index score that is expected to represent minimally important change in HRQL for patients with type 2 diabetes. After finding evidence for MID estimates from Chapters 2 and 3, **Chapter 4** investigates the variability in interpretation of (small) EQ-5D-5L index score differences between and within individuals as well as for different preference

elicitation methods, in the general Canadian population. **Chapter 5** then applies concepts and evidence of MIDs in EQ-5D-5L index scores to propose a quality-adjusted life year calculation that adjusts for meaningful within-patient change in HRQL. This method is then used to support the interpretation of HRQL outcomes in a trial for depression treatment in patients with type 2 diabetes as a case-study. Lastly, **Chapter 6** provides an overview of the main findings from this thesis. The implications of these results in regards to end-user support are discussed, and recommendations for future research are given.

1. Introduction

1.1 Motivation for the Collection and Assessment of Patient-Reported Outcomes

There is interest in obtaining health information from the patient perspective, i.e., patient-reported outcomes, as it relates to the patient's function and health status, outcomes of medical care, and health-related quality of life (HRQL) [1,2]. The Canadian Institute for Health Information outlines how this type of health information can be used to address gaps and inform decision-making at the three levels of a health system: the clinical or patient care level, the administrative level, and the policy level [2]. Currently, there are gaps in the information system that limit evaluating and improving the quality of care delivered by the health system [2]. The increasing prevalence of chronic conditions creates a need for strengthening patient-centred management and shared decision-making that is more explicitly based in patients' preferences as well as their perspectives on health and trade-offs when considering options of care [1–3]. Changing healthcare costs and epidemiology of diseases emphasize the importance of evidence-informed decision-making that reflects societal values and cost-effective allocation of resources [2,4–6].

Traditional approaches to health and healthcare have focused on treating patient symptoms and risk factors for disease, gathering information from diagnostic measures that report on physiological and clinical outcomes; however, these approaches do not explicitly capture patients' perspectives of their own function or their preferences [1,2]. Similarly, at the administrative level, information on healthcare utilization and associated expenditures is in abundance, but this could be supplemented with aggregate information on patients' health status and HRQL to monitor and evaluate the effectiveness of care [2].

Finally, at the policy level, information on societal preferences and medical intervention effectiveness can be used to associate costs with trade-offs in health across the spectrum of services [2,3,5,7]. The integration of appropriate measures to capture patient-reported outcomes as complementary to current clinical and administrative metrics of disease and health system performance has the potential to offer a more holistic and patient-centred approach to care [2].

The EQ-5D is one of the most popular HRQL measures in the world [8,9]. Originally conceived of as a measure of HRQL to be used in the calculation of quality-adjusted life years (QALYs) for cost-utility analysis, the EQ-5D is also widely used in outcomes research to assess differences in health between populations, changes in health over time, and as a result of interventions [9–12]. Moreover, the EQ-5D is used in routine data collection (i.e., large scale applications) by several health systems (e.g., National Health Service, in the United Kingdom) [10,11]. Overall, these different uses of the EQ-5D have placed different demands on the instrument than originally conceived, including modes of administration as well as language and cultural validity [10,13]. Further, a general challenge for end-users of all patient-reported outcome measures (PROMs), including the EQ-5D, has been supporting the interpretation of scores for different purposes at multiple-levels (e.g., clinical, administrative, funding decision, etc.) to give meaningful and timely feedback in order to align resource allocation and clinical decisions with outcomes that matter to patients in a socially responsible manner [11,14–20].

1.2 Purpose Statement and Background

The purpose of this thesis is to examine the interpretability of the EQ-5D HRQL score, particularly the EQ-5D-5L index score based on a Canadian value set, through the estimation and application of the concept of minimally important differences (MIDs). The introduction will provide a background on types and measurement properties of PROMs. Next, we consider the normative (i.e., value-based) considerations involved in HRQL measurement and healthcare resource allocation decisions with particular attention to guidelines for Canada. A main tenet of this thesis is the notion that EQ-5D HRQL index scores, however limited, are useful in their representation of the average of the general population's health preferences. It is then reasoned that even though any one individual's preferences may differ from the EQ-5D HRQL index score, it is possible to support the interpretation of HRQL index scores (as measured by the EQ-5D) in terms of what change/difference in HRQL index score is meaningful to patients at the point of application.

1.2.1 Patient-Reported Outcome Measures

A patient-reported outcome is defined as "any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else" (p. 2) [21]. In this regard, an instrument that measures patient-reported outcomes is known as a PROM. The following provides background information on types of PROMs and their measurement properties, specifically highlighting differences between health status and preference-based measures, and generic and condition-specific measures, as well as responsiveness and interpretability.

1.2.2 Health-Status and Preference-Based Measures

Self-reported measures of health status are often referred to as measures of HRQL; however, for the purposes of this thesis, measures of HRQL will refer only to preference-based scores or utility values associated with health status [22]. This brings to attention the first level of categorization of self-reported measures of health status, namely the difference between measuring health profiles and health utilities. A health profile measure can be used to determine the respondent's functional or health status based on a descriptive system, in which the measure is constructed to capture particular aspects of health or attributes, typically including physical, mental and social abilities [3]. In this regard, the data obtained from a health profile measure is often reported for each attribute or concept in terms of level of functioning or impairment, and where multiple items that measure the same concept may be summarized by a single composite score [3].

Compared to a health profile measure, a preference based measure is formulated to determine the importance or value of one's overall health status (based on patient or societal values), and is summarized by a single score [3,6]. For example, the EQ-5D uses a multi-attribute health descriptive system encompassing five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression (Figure 1). Each dimension has 3 or 5 response options representing levels of impairment from "no problems" (level 1) to "extreme problems" or "unable to do" (level 3 or 5) [23]. In addition, the EQ-5D has a visual analogue scale (VAS) ranging from 0 "the worst health imaginable" to 100 "the best health imaginable" [24]. The self-reported scores from the health descriptive system can be used as health profile measures, while the VAS is a global self-rating of health status [24]. However, the EQ-5D can also be used to obtain an indirect measure of health utility (i.e., HRQL) through the application of a scoring algorithm that reflects the public's preference for the self-reported health state: anchored at 1.0 representing full health, and 0.0 representing a health state equivalent to being dead [25]. The scoring algorithm or value set is conventionally derived through valuation studies that involve interviews of a representative sample of the population [26,27]. According to the Canadian scoring algorithm, EQ-5D-5L index scores range between -0.1482 and 0.9489 for the worst (55555) and best health states (11111) defined by the descriptive system [7] (Figure 1). The use of a scoring algorithm is an indirect measure of health utility, whereas a direct measure involves eliciting the preferences of the respondent directly at the point of application.

1.2.3 Generic and Condition-Specific Measures

Both health profile as well as HRQL measures may be generic or condition-specific. Generic measures typically use broad or general aspects of health, while condition-specific measures are more tailored to the aspects of health considered relevant to a specific patient population [1]. The advantage of generic measures is that they can be applied across many types of health conditions giving a common measure of health to use in global assessments or to compare populations with multiple health problems [1,3]. However, there is a concern that due to their broad applicability, generic measures may not be able to detect meaningful change in specific populations or contexts [1,28]. This may be particularly true for more mild conditions or health states, where health can still be gained or lost as a result of disease progression or treatment, but such changes may not be adequately reflected in scores from generic measures [1,28].

1.2.4 Responsiveness and Interpretability of PROMs

The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for PROMs [29]. According to this consensus, responsiveness is a measurement property of a PROM that is defined as the "the ability of a PROM to detect change over time in the construct to be measured" (p. 743) [29]. To this end, measurement properties of an instrument are demonstrated by way of testing a priori hypotheses in a population of interest. Thus, it follows that responsiveness is a necessary pre-condition in order for an instrument's change in scores to be interpretable [30]. To evaluate responsiveness a user must know that meaningful change in health has occurred (e.g., after surgery). Similarly, if no change in health has occurred over time, an instrument ought to also show (little to) no change in scores. Based on classical test theory, poor test-retest reliability demonstrates that a score has much measurement error, thus a user cannot be confident that observed changes in scores over time reflect true change [31]. Therefore, adequate test-retest reliability is a necessary pre-condition for an instrument to be responsive.

Responsiveness and reliability differ from a score's interpretability, which is defined as "the degree to which one can assign qualitative meaning, that is, clinical or commonly understood connotations, to an instrument's quantitative scores or change in scores" (p. 743) [29]. In this regard, interpretability is considered a characteristic (not a property) of a PROM requiring investigation and development to be operationalized. To this end, the MID has been proposed as a useful metric to support the interpretability of a PROM [32,33]. While there is no clear general consensus on the definition of an MID, its origins are credited to Jaeschke et al. (1989) who referred to "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management" (p. 408) [34].

There are several other related metrics that also address interpretability and other (related) measurement properties or characteristics [35,36]. First, the minimally clinically important difference (MCID), or the smallest change that is considered clinically relevant (i.e., considered meaningful to the clinician), represents another type of interpretability. However, this may be less-aligned with the intent of a PROM as stated by Francis et al., "the term 'patient-important' is more appropriate than 'clinically important' to emphasize the patient-centrism of these outcomes and the goals of directed interventions" (p. 5) [31]. Another term, the minimally detectable change (MDC) is the smallest change that can be detected beyond random error, which may be most applicable to the reliability of the PROM. As emphasized by Turner et al., "the values of MID and MDC measure different concepts; the former measures important apparent change and the latter statistical distribution of margins of error" (p. 34) [37]. For these reasons, this thesis is most interested in the MID to support the interpretability of EQ-5D HRQL scores, representing the smallest change in score (over time) that patients would consider meaningful.

1.2.5 Normative Basis for Healthcare Allocation Decisions

It is generally acknowledged that poor health impacts one's quality of life. Thus maintaining, restoring or lessening the negative consequences from declining health are the outcomes of interest when seeking medical care [38]. It is also understood that there is uncertainty in when or to what extent a patient experiences illness, as well as uncertainty in the effectiveness of medical treatment [39]. These observations have been framed in a welfare economic perspective, which is based on ideas of utility and consequentialism [40]. In this regard, utility, specifically health-utility, is a preference for a health outcome or health state (and not the process or intention from which the outcome arose), which, under further assumptions, can be interpersonally compared, aggregated and/or traded-off [40]. Furthermore, uncertainty in illness and treatment effectiveness has created a market for medical-insurance [38,39], which in many developed economies has resulted in some form of publicly-funded health system (due to imperfect market characteristics and a societal value for delivering care based on need and not one's ability to pay) [38]. The major challenge of public health systems is how to appropriately allocate a budget (for healthcare) to serve the health needs and values of the public [41]. The following sections provide an overview of the normative (i.e., value-based) positions taken in the measurement of HRQL and the downstream effects in the calculation of QALYs, and ultimately the incremental cost-utility ratio (ICUR) used in cost-utility analysis.

1.2.6 Health-Utility Measurement

Utilities are used in the study of economics to represent an "individual's preference ordering over bundles of goods or states … [where] an individual moving to a preferred state of the world is an equivalent statement to an individual having a higher level of utility" (p. 328) [40]. While it is accepted that utilities cannot be directly measured or observed per se, researchers have attempted to estimate underlying utilities from people's choices, i.e., preference for one state over another [38,42]. In regards to health utilities, revealed preferences are not necessarily observable or informative as there are imperfect market characteristics and the patient often lacks knowledge of the outcomes of care [38,39,43]. Therefore, health utilities are typically estimated using stated preference methods [43,44]. To this end, the estimation of health utilities requires several components: a satisfactory description of health, a method of eliciting preferences, and a way of applying health utilities in decision-making. Taking the perspective of a 'socially legitimate decision-maker', the proceeding paragraphs address the first two components, while the last point is addressed in a later sub-section.

To derive a measure of health utility, a decision-maker has the following considerations. First, the aspects of health that are impacted by the health system (presently and in future planning) such that they are important to capture in measures of the health produced or forgone by different allocation decisions [45]. The literature provides multiple definitions for health; for example, the World Health Organization defines health as "a state of complete physical, mental and social well-being, and not merely the absence of disease and infirmity" [22]. However, the extent to which a decision-maker wishes to emphasize (to a greater or lesser extent) all or some aspects of health will ultimately be reflected by the choice of health status measure (namely, its content), which brings us to the second consideration. Again, there are numerous instruments of health status available to a decision-maker, which are broadly differentiated in terms of content and applicability. A decision-maker that seeks to make allocation-decisions across the spectrum of services will require a generic instrument, as opposed to a condition specific instrument [3]. The content of an instrument is primarily determined by its health descriptive system, which contains any number of attributes (i.e., aspects or dimensions of health) wherein each attribute may have multiple (mutually exclusive) response levels in terms of capacity, functioning (or impairment), behavior, performance, symptoms and/or consequences of health problems [1,9,12,44–46]. There are various practical and conceptual advantages and disadvantages to the number of attributes and levels in an instrument that need to be considered when seeking a sufficiently comprehensive health descriptive system. For example, the extent of ceiling (maximum value) and floor (minimum value) effects is important to consider as this limits the variation of a score affecting its responsiveness and divergent (or discriminatory) validity [1,23,31]. In addition, instruments' content may also differ in terms of their reference period (e.g., your health today versus in the last week, etc.). After choosing a measure of health status that adequately reflects the decision-maker's health objectives, we can now consider how to value the health states that it describes.

Welfarist theory may be a useful starting point when considering how to value health states [40,47,48]. In keeping with a welfare economics perspective, and its tenet of individual sovereignty, it follows that the patient is the best judge of his/her utility [40]. However, healthcare decision-making involves multiple patients, thus our method of valuing health states must allow for interpersonal comparisons, which presents the following challenges. First, different patients will have experienced different illnesses and treatments, and thus may introduce bias (e.g., self-interest, adaptation, etc.) in preference elicitation [44,49]. Second, according to Arrow's theorem it is impossible to make fair interpersonal comparisons based on ordinal preferences [50]. Third, since welfarist theory is a form of consequentialism, other aspects of health and care may not be given importance (e.g., processes of care, patient experiences and capabilities) [13,40,48]. Alternatively, provider or decision-maker preferences may also be biased by experience or conflict of interest, and would not align with welfarist decision-making

[44,49]. In light of these multiple challenges, it is important to recognize that the ultimate choice of how to value health states may reflect a compromise between individuals' preferences and some other criteria considered important (by the decision-maker) to the decision-problem, and in doing so, may not strictly conform to a welfarist framework [40,51–53].

Based on normative principles of justice and the 'fair' distribution of societal resources where everyone is a potential patient, it is generally accepted that societal preferences (i.e., involving a representative sample of the general population) are the appropriate choice when making allocation decisions for a publicly funded health system (i.e., shared societal resources), which is congruent with a shared decision-making perspective [27,52]. Moreover, preferences have been shown to differ between countries, thus country-specific preferences are recommended [26,27]. To ensure a 'veil of ignorance', typically the preference elicitation task involves evaluating different health states ex-ante (i.e., hypothetical as opposed to experienced health states) using various methods of determining preferences, including standard gamble, VAS, discrete-choice experiments (DCE) and time trade-off (TTO) tasks [38,42,54]. The various tasks have different advantages and disadvantages that differ in terms of simplicity (i.e., ease of understanding and respondent burden) and best representing one's preferences for health states (i.e., validity and responsiveness). For instance, a strict adherence to the axioms of von Neumann-Morgenstern utility theory would recommend standard gamble; however, this technique is considered to be cognitively demanding and studies have noted that probabilities are not well understood by respondents [44].

Observed preferences are transformed from their 'raw' value to a cardinal interval scale where 0 and 1 are the utility values for health states equivalent to being dead and in full health, respectively, and where health states worse that dead are represented as negative values (with some limitations) [44]. Importantly, cardinal and/or ordinal preferences are observed depending on the elicitation task [44]. Ordinal preferences describe a persons' preference ranking, and are based on choosing the best (i.e., preferred) health state in comparisons that include at least two health states that differ in attributes [44]. In addition to ranking information, cardinal preferences describe a person's strength of preference for health states (i.e., preference differences between health states) [44]. Observed cardinal preferences, such as those observed in a TTO task, are considered to be more directly related to a cardinal interval scale; whereas, additional assumptions or attributes are required when transforming observed ordinal preferences (e.g., from DCE) to a cardinal interval scale [44,54].

Due to the number of health states described by (sufficiently comprehensive) instruments of health status, additional methods (involving additional assumptions) must be used to take the values elicited for a sample of health states and build a formula (i.e., scoring algorithm) to generate values for all of the health sates. This involves multi-attribute utility theory (and assumptions of independence, etc.) and modeling the (population) mean health utilities to determine a best-fit model based on a priori specified criteria [44,46]. In summary, preferences used in generic indirect preference-based multi-attribute measures of HRQL are elicited from a representative sample of the general public to generate a value set, which functions as a scoring algorithm that maps a patient's self-reported health state to a utility score reflecting the average of the general population's preferences [42].

1.2.7 Quality-Adjusted Life Year

After deciding on a preference-based measure and a value set, a decision-maker can then conduct a cost-utility analysis to inform allocation decisions regarding different interventions. Again, we will adopt a publicly funded healthcare payer perspective, thus the relevant costs are those from the public payer (i.e., budget of healthcare system) [27]. As stated by the Canadian Agency for Drugs and Technologies in Health (CADTH), the purpose of a cost-utility analysis is "to estimate the cost and effect trade-off of two or more interventions" (p. 14) [27]. Thus, the health produced from one intervention is compared against a comparator intervention to determine its incremental benefit. This requires a common measure of health produced, which in a cost-utility analysis is the QALY [19,55].

After calculating the costs and QALYs produced by all of the relevant comparators, the ICUR (of any two comparators) can be compared by the difference in expected costs divided by the difference in expected QALYs, to give a cost per QALY [27,42]. However, in order for a decision-maker to understand if this ICUR will improve the efficiency of a healthcare budget, we need to know if the technology provides 'good value for money', which is commonly referred to as the cost-effectiveness (i.e., cost-utility) threshold [56]. Generally there are two approaches to determining a technology's cost-effectiveness threshold: demand and supply side approaches [56]. In the former case, willingness-to-pay methods are used to estimate its value; however, assuming that there are external budget constraints this approach may not be appropriate for publicly funded healthcare systems [27,56]. The supply side approach reflects allocation decisions under a constrained budget scenario wherein the cost-effectiveness threshold is the health (as measured by QALYs) that is displaced by paying for the new technology (i.e., the current cost of producing a QALY within the health system) [56]. Furthermore, the estimation of the health benefits forgone need to incorporate the same preference-based valuation of health as the health benefits produced [57]. As is often the case, a decision-maker may not know the identity of the patients who bear the opportunity cost, which, in addition to the fact that everyone is a potential patient, further supports the use of societal preferences in the calculation of health utility scores from generic indirect preference-based measures of HRQL.

1.3 Conceptual Foundation

This section outlines the conceptual foundation for this thesis. This includes: (1.3.1) an overview of the EQ-5D-5L as a measure of HRQL and the challenges in using generic indirect preference-based scores, (1.3.2) understanding sources of variation in HRQL scores, and (1.3.3) and the potential role for the MID concept to support the interpretability of HRQL scores.

1.3.1 Overview of EQ-5D-5L

The EQ-5D-5L is a generic indirect preference-based measure of HRQL composed of five health dimensions (mobility, self-care, usual activities, pain/discomfort, anxiety/depression) with five response levels (no problems, slight, moderate, severe, extreme/unable to function problems) per dimension (Figure 1). The Canadian Valuation

Study elicited TTO responses (using a 10-year time-horizon) for the EQ-5D-5L from a representative sample of the Canadian population to develop a scoring algorithm. The scoring algorithm is used to assign an index score representing the average of the general population's preferences for each of the 3,125 health states defined by the instrument. In this way, when a respondent is asked to complete the EQ-5D-5L questionnaire 'considering his/her health today' their response indicates their health status (Figure 1). As previously discussed, the scoring algorithm converts the respondent's health status to a HRQL score; however, this does not necessarily reflect the respondent's own preference for the self-reported health state [58].

The interpretation of changes in EQ-5D-5L HRQL scores may be complicated by the fact that the value set is based on population preferences, as such the change in HRQL score does not necessarily represent to what extent the patient values the change in health state [58–60]. Figure 2 outlines how the meaning of HRQL scores might change when considering outcomes for specific patient cohorts, subgroups, and individuals. Of course, changes in the VAS (of the EQ-5D-5L) can be used to determine improvement or deterioration as reported by the patient [24,61]. Furthermore, the MID of the VAS can be estimated to support its interpretation [62]. However, self-reported VAS scores do not capture (individuals') preferences for different health states, thus they are not typically used to evaluate HRQL trade-offs or to calculate QALYs [22,24,61]. Moreover, the measurement properties of the VAS (e.g., reliability and responsiveness) are different from the EQ-5D-5L health descriptive system and HRQL index score. For these reasons, VAS scores are not interchangeable with HRQL index scores, thus the interpretability of the HRQL index score is not necessarily improved through analysing VAS scores

[24,61].

The challenge is thus how (or perhaps if) changes in HRQL score can be interpreted in a way to reflect the value that patients place on the change at the point of application. The generic nature of the EQ-5D-5L (like other HRQL instruments) allows for generalizability of results so that comparisons can be made across conditions and treatments; however, this external validity may trade-off with its internal validity [1,13,28,46,63]. In this regard, assessments of HRQL changes that do not take into consideration the shortcomings of the instrument and the corresponding 'error' in scores may lead to sub-optimal decisions [12,46,64–69]. In other words, the health of patients may not be improved (and perhaps even harmed) through the allocation of healthcare resources that are based on fallible evaluations of HRQL scores [68,70].

1.3.2 Sources of Variation in HRQL Index Scores

The sources of variation in HRQL index scores can be categorized into two components: *uncertainty at the population-level* and *within population heterogeneity* (Figure 3). The population-level uncertainty can be further divided into random or sample error (i.e., stochastic uncertainty) and parameter uncertainty [71]. Since the objective of economic evaluations is to inform a decision based on existing information, population level uncertainty is addressed by way of probabilistic analysis using input values that are stochastically drawn from parameter distributions (e.g., Markov models) that yield expected or average values for the outputs of interest [71]. Thus inference, as conventionally understood by common statistical approaches (i.e., using a decision-rule based on a type 1 error rate), is considered irrelevant to the decision problem [72]. That said, greater certainty about a technology's cost-effectiveness may be desired, which can be operationalized through a ‘value of information’ analysis [72,73].

In contrast to population-level uncertainty, methods to address within population heterogeneity, are not as widely adopted [71,74–76]. It is well recognized that treatment effects can differ between members of the patient population (as defined by the decision-problem) based on differences in observable characteristics (e.g., age, gender, duration of condition, number of co-morbidities) [48,71,74,76]. Similarly, treatment effects as reflected through changes in health states (and the general public’s preference for the change) may also differ within the patient population (see Figure 2) [48,59,69,76]. Thus, heterogeneity in patient preferences involves partitioning the target population into subgroups to yield inter-subgroup differences, such that some subgroups may have higher reported changes and other subgroups lower reported changes than the aggregated (e.g., average) change of the target population [59,76,77]. This issue of inter-subgroup differences can be addressed by stratified analysis [74]. While the effects of subgroup heterogeneity could potentially result in different decisions with respect to cost-effectiveness among different subgroups, there is little guidance for conducting stratified analysis [75]. Furthermore, this form of preference heterogeneity that lends itself to stratified analysis does not necessarily provide insight in understanding the preferences of the target population (or subgroup) that is affected by the treatment decision [69]. This brings us to the second component of preference heterogeneity, namely that different people have different preferences [69,74]. While economic evaluation guidelines generally support the use of the average of the general population’s preferences (so as to be pursuant with a social decision-making perspective), CADTH provides the following caveat [27] (p. 46):

> *When there is, however, a concern that general population preferences may not fully represent the experiences or outcomes of those who are affected by the intervention (both in terms of new interventions to be funded, and those that would potentially be defunded), alternative sources of preferences may be considered.*

Personalized or individualized care and patient choice models have been proposed as a means of improving the efficiency of resource allocation compared to the conventional 'one-size-fits-all' approach to funding decisions [69,74,78]. In particular, this takes more of a patient-centric view [48,69]. However, this approach has numerous difficulties including eliciting and standardizing individual preferences of different patients, as well as understanding differences in cost-utility due to cost internalization [69,74]. To this end, we propose that MID estimates for HRQL index scores (measured by way of the average of the general population's preferences) may be useful in further informing assessments of HRQL changes (in terms of what is meaningful to patients in the decision-problem) in a way that complements (and does not compromise) the surrounding framework for using HRQL index scores [1,79,80].

1.3.3 The Relevance of the MID Concept

It is recognized that different value sets applied to the same sample of reported health states may affect the magnitude of change in HRQL index score, which has downstream implications on social decision-making and statistical properties [12,63,81]. In this way, the use of methods to support the interpretation of HRQL index score changes, specifically methods to estimate MIDs, may be useful to promote understanding of observed HRQL changes at the point of application. An MID is considered specific to an instrument (and thus its value set) as well as the patient population and clinical context

[30,36,80]. Similar to other quantitative data, such as the measurement of temperature, different instruments with different increments (i.e., units) produce different numbers, but the choice of any single instrument does not prevent an understanding of what change in temperature is meaningful [82]. In regards to generic HRQL index scores, different instruments and/or value sets have different increments on similar scales (i.e., 0 and 1) affecting the data's properties, and thus the threshold for an MID, but it is the use of an appropriate MID that might allow for a consistent interpretation of the study outcomes (all else equal).

Methods of estimating the MID are similar to testing for responsiveness, for instance, both use distribution and anchor-based approaches [30,31,35]. Importantly, there is no one-size-fits all method. Furthermore, it should be recognized that responsiveness is not conclusively 'proven', nor is '*the* MID' ever determined for an instrument [30]. Often, a combination of distribution-based and anchor-based methods are used to provide accumulated evidence of an instrument's responsiveness or the MID, wherein the quantity and strength of evidence provide a user with confidence in an instrument's responsiveness and MID in a particular patient population and/or clinical context [60].

The distribution-based approach is based on the properties of the change score's distribution as reported by patients included in a study sample [30,31,35]. This involves calculating other metrics from the sample such as effect size or standardized response means. The degree of change, (i.e., trivial, small, moderate or large change) is then determined using recognized benchmarks of interpretation (e.g., Cohen's guidelines for effect size) [83]. Of course, for the purposes of estimating an MID, a non-trivial yet small difference is desired. A systematic review of MID estimates concluded that many MIDs are approximately equal to a distribution-based estimate of one-half the standard deviation of the baseline score [84]. However, recommendations for MID estimation state that while distribution-based approaches, such as the one-half standard deviation, can inform MID estimates, they are generally considered as lower quality than anchor-based MID estimates [30,32,37].

Unlike the distribution-based approach, the anchor-based approach uses the instrument's association with other PROM scores (e.g., global rating of change) or external assessments, where the change in anchor score is independently interpretable from the patients' perspective [6,30,31,83]. Again, criteria for a trivial, small, medium or large change in anchors are then related to the change observed in the score of interest. Since the anchor-based approach depends on the choice of anchors, the responsiveness of an instrument will vary accordingly [30,46]. Furthermore, for the purposes of estimation, it is considered good practice to use multiple anchors (reported by the patient) to give a triangulated estimate or plausible range [30,85].

Methods to estimate the MID in EQ-5D-5L HRQL index score do not differ from the methods previously described. However, since the EQ-5D-5L HRQL index score is an indirect, multi-attribute preference-based measure, with a defined health descriptive system, an additional method, known as the instrument-defined approach, has also been proposed [86]. This is considered a variant of the anchor-based approach that is based on internally defined anchors, i.e., the difference in HRQL index scores between single-level transitions defined by the instrument [86].

There are a number of considerations to MID estimation that require further research, including how the MID may depend on direction of change and baseline scores [30,36,83]. Researchers have proposed criteria (e.g., magnitude of correlation between scores) to guide the selection of anchors and appraise the quality of MID estimates [30,32,36]. In effect, it is possible that smaller or larger MID estimates are produced for the same HRQL measure, for different contexts or applications. This may be particularly emphasized for generic preference-based measures such as the EQ-5D. For example, a large MID may suggest that the instrument is responsive to change in the population, but there may exist aspects of health that are either not captured by the questionnaire or are dominated by the value set resulting in a change that is large in magnitude but one that is considered by the patients to be minimally important [68,70].

1.4 Research Questions

The proceeding chapters will address the following research questions:

1. What are plausible MID estimates for the EQ-5D-5L index score?
2. What is the smallest change in EQ-5D-5L index score that is expected to represent minimally important HRQL change for adults with type 2 diabetes?
3. Does the concept of an MID have relevance in a preference-based measure of HRQL with a value set derived from the general population?
4. How can methods to calculate QALYs adjust for minimally important within-patient changes in a HRQL score?

1.5 Specific Aims

- **Study 1 – What are plausible MID estimates for the EQ-5D-5L index score?**
    - Aim 1.1: To apply the instrument-defined approach to the EQ-5D-5L for different country-specific scoring algorithms.
        - Rationale 1.1: The instrument-defined approach is considered a variant of the anchor-based approach that uses transitions in the dimensions and levels of the health-descriptive system to determine the smallest change in index score that is considered meaningful.
    - Aim 1.2: To compare the MIDs from different country-specific scoring algorithms.
        - Rationale 1.2: A country-specific scoring algorithm weights single-level transitions and the differences in index scores are averaged to produce an MID estimate that is specific to a country's value set.
    - Aim 1.3: To examine the magnitude of MID estimates across the range of baseline index scores.
        - Rationale 1.3: The magnitude of the MID estimate may vary across the range of baseline index scores according to the weights applied to single-level transitions defined by the scoring algorithm.
    - Aim 1.4: To determine the effect of removing maximum-valued scoring parameters for each dimension.
        - Rationale 1.4: Single-level transitions that invoke a maximum-valued scoring parameter may not be representative of a small difference. Therefore, removing these transitions may give a more representative

estimate of the smallest difference in index score that is considered meaningful.

- **Study 2 – What is the smallest change in EQ-5D-5L index score that is expected to represent minimally important HRQL change for adults with type 2 diabetes?**
    - Aim 2.1: To estimate what change in EQ-5D-5L index score represents the smallest meaningful change in HRQL for adults living with type 2 diabetes.
        - Rationale 2.1: The MID supports the interpretation of scores at the point of application, thus estimates need to be generated for different target populations. Hypothesis: The MID is equal to a non-zero change in index score.
    - Aim 2.2: To compare estimates from the anchor-based approach to estimates from the instrument-defined approach.
        - Rationale 2.2: Recommended methods for estimating MIDs include using multiple anchors and approaches to triangulate an estimate and/or suggest a plausible range of MID values.
    - Aim 2.3: To investigate if and how MID estimates depend on the starting baseline score.
        - Rationale 2.3: What patients consider to be a minimally important change in their HRQL may depend on their current HRQL.
    - Aim 2.4: To investigate if and how MID estimates depend on the direction of change.

        - Rationale 2.4: What patients consider to be a minimally important change in their HRQL may depend on whether their health is improving or worsening.
    - Aim 2.5: To obtain MID estimates for (clinically) relevant sub-sets of the population.
        - Rationale 2.5: The MID supports the interpretation of scores at the point of application, thus estimates need to be generated for different (clinically) relevant sub-sets of adults with type 2 diabetes.
- **Study 3 – Does the concept of an MID have relevance in a preference-based measure of HRQL with a value set derived from the general population?**
    - Aim 3.1: To determine how participants in the valuation study interpret differences in EQ-5D-5L index score.
        - Rationale 3.1: The EQ-5D-5L index score represents the average of respondents' preferences allowing for small differences to be calculated between health states; however, it is unclear to what extent small differences represent participants' ordinal preferences. Hypothesis: Large differences in EQ-5D-5L index score are uniformly representative of transitions to better or worse health states, while small differences are ambiguous, representing transitions to states that are perceived to be about the same, worse or better.
    - Aim 3.2: To examine how cardinal responses depend on the ordinal responses of participants in the TTO task.

  - Rationale 3.2: The smallest possible increment in the TTO task is 6 months on a 10-year time horizon. Therefore heterogeneity in ordinal preferences for small differences in index score may represent cardinal preferences that are small in magnitude. Hypothesis: For index score differences near 0, the cardinal responses of participants who perceived the transition to be better/worse reflect preferences that are small in magnitude (i.e., 6 month, minimum allowable increment).

- Aim 3.3: To compare ordinal responses from DCE with TTO.
  - Rationale 3.3: DCEs elicit respondents' ordinal preferences, while the cardinal preferences from TTO tasks can be converted to ordinal preferences. In addition, the Canadian EQ-5D-5L value set is based on TTO responses. Hypothesis: On average, if there is consistency in interpretation across respondents, the same difference in index score will have the same probability of representing a transition to a worse/better health state regardless of the preference-elicitation method.

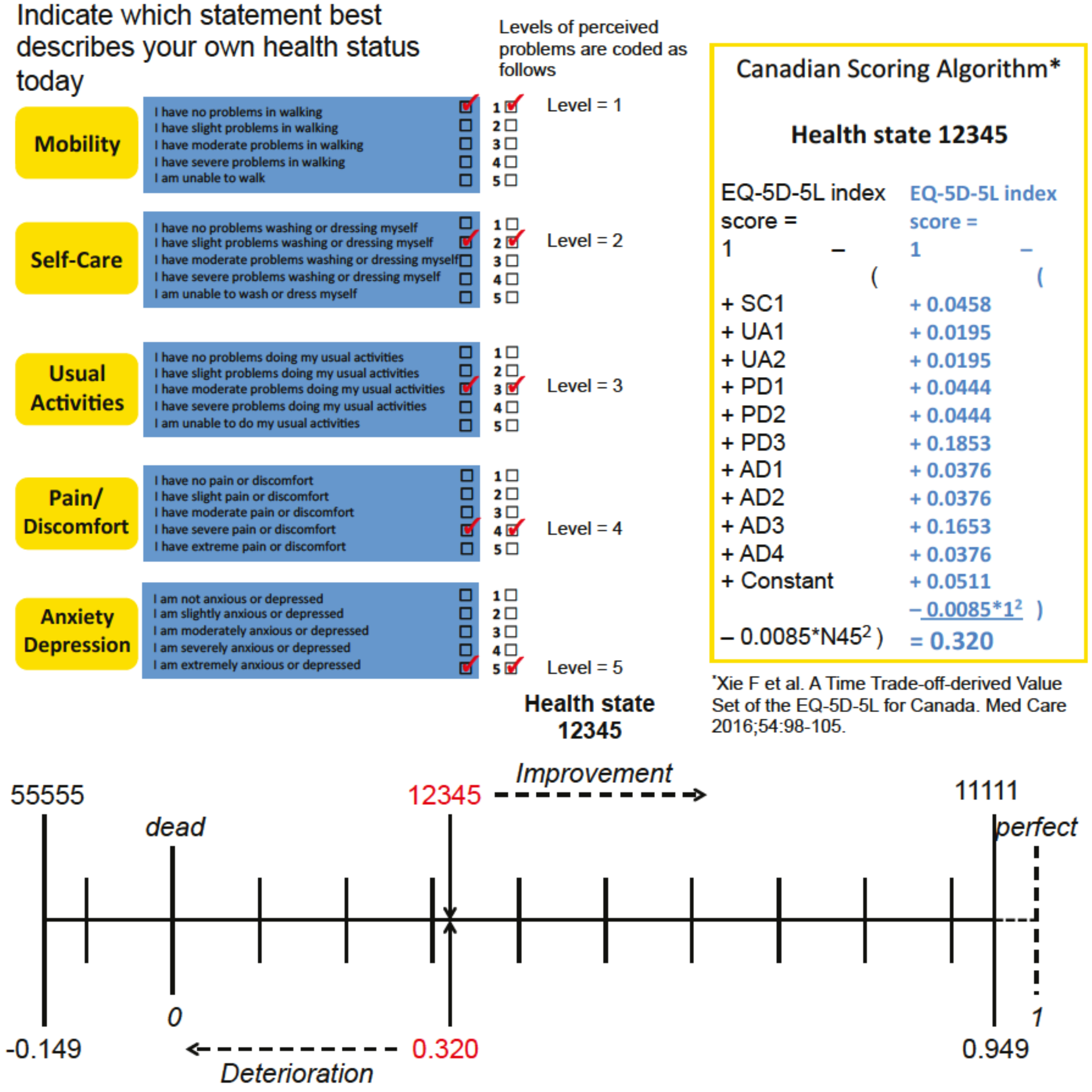
- Aim 3.4: To determine the extent of intra-individual heterogeneity in interpretation of index score differences.
  - Rationale 3.4: Population-level preference heterogeneity may result from heterogeneity between and/or within individuals. Hypothesis: If there is little intra-individual heterogeneity, the majority of respondents' ordinal preferences will be consistent with the direction of change represented by small index score differences.

  - Aim 3.5: To determine a difference in EQ-5D-5L index score that best represents participants' interpretation of the transition between health states.
    - Rationale 3.5: Due to preference heterogeneity, a difference in index score needs to be large enough to be meaningfully interpretable. Hypothesis: A value that is not different from zero would suggest that any non-zero difference in index score best represents participants' interpretation of the transition between health states.

- **Study 4 – How can methods to calculate QALYs adjust for minimally important within-patient changes in a HRQL score?**
  - Aim 4.1: To explore the challenges in assessing QALY outcomes.
    - Rationale 4.1: The QALY combines multiple HRQL scores over time into a single composite outcome that can be used to assess HRQL changes within and between groups. However, real-world assessments need to consider differences in baseline HRQL scores, measurement error, and the value of small HRQL changes as perceived by patients.
  - Aim 4.2: To re-analyse QALY outcomes in the Alberta TEAMCare-Primary Care Network trial in the treatment of depression for patients with type 2 diabetes, comparing results from different QALY calculation methods.
    - Rationale 4.2: The TEAMCare study collected EQ-5D-5L responses from patients to assess between-group QALY differences. Previous results found QALY benefits that do not agree with between-group changes in HRQL.

- Aim 4.3: To investigate the impact of adjusting for between-group differences in baseline HRQL scores.
    - Rationale 4.3: Between-group differences in baseline HRQL scores have been shown to affect assessments of QALY outcomes. Previous results did not adjust for differences in baseline HRQL scores.
- Aim 4.4: To investigate the impact of adjusting for ‘meaningful’ HRQL changes within-patients.
    - Rationale 4.4: Despite the attention given to ‘error’ in HRQL measurement and adjusting for differences in baseline HRQL scores, the importance of assessing the psychometric properties of an instrument, and the possibility of results leading to sub-optimal allocation decisions, no method currently exists to adjust for expected ‘meaningful’ HRQL changes within patients. We propose a new method based on the MID in EQ-5D-5L index score.
- Aim 4.5: To examine incremental QALY estimates across the range of baseline HRQL scores observed in the TEAMCare study.
    - Rationale 4.5: The left-skewed distribution of baseline HRQL scores may cause lower baseline HRQL scores to disproportionately affect incremental differences in QALYs.

Figure 1. The five-dimensional five-level EQ-5D-5L generic indirect preference-based health-related quality of life instrument, and the Canadian scoring algorithm for calculating an EQ-5D index score.



Indicate which statement best describes your own health status today

Levels of perceived problems are coded as follows

Mobility
- I have no problems in walking ☑ 1☑ Level = 1
- I have slight problems in walking □ 2□
- I have moderate problems in walking □ 3□
- I have severe problems in walking □ 4□
- I am unable to walk □ 5□

Self-Care
- I have no problems washing or dressing myself □ 1□
- I have slight problems washing or dressing myself ☑ 2☑ Level = 2
- I have moderate problems washing or dressing myself □ 3□
- I have severe problems washing or dressing myself □ 4□
- I am unable to wash or dress myself □ 5□

Usual Activities
- I have no problems doing my usual activities □ 1□
- I have slight problems doing my usual activities □ 2□
- I have moderate problems doing my usual activities ☑ 3☑ Level = 3
- I have severe problems doing my usual activities □ 4□
- I am unable to do my usual activities □ 5□

Pain/ Discomfort
- I have no pain or discomfort □ 1□
- I have slight pain or discomfort □ 2□
- I have moderate pain or discomfort □ 3□
- I have severe pain or discomfort ☑ 4☑ Level = 4
- I have extreme pain or discomfort □ 5□

Anxiety Depression
- I am not anxious or depressed □ 1□
- I am slightly anxious or depressed □ 2□
- I am moderately anxious or depressed □ 3□
- I am severely anxious or depressed □ 4□
- I am extremely anxious or depressed ☑ 5☑ Level = 5

Health state 12345

Canadian Scoring Algorithm*

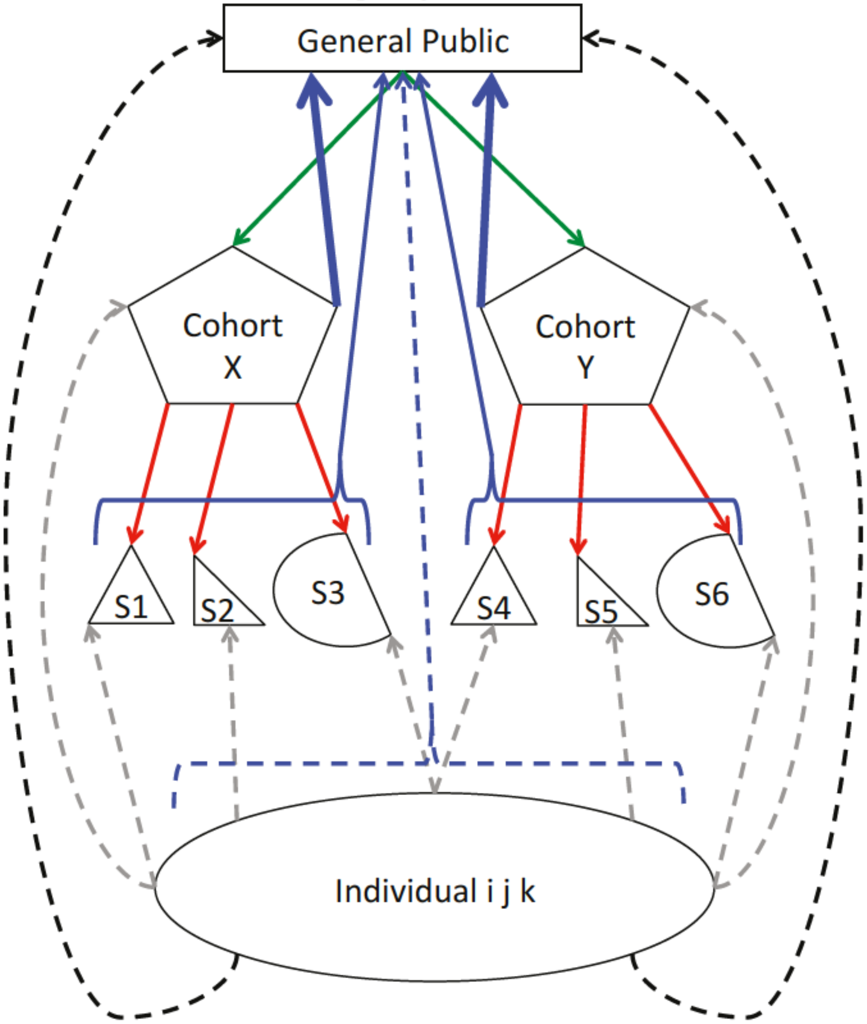
Health state 12345

EQ-5D-5L index score =	EQ-5D-5L index score =
1 − (	1 − (
+ SC1	+ 0.0458
+ UA1	+ 0.0195
+ UA2	+ 0.0195
+ PD1	+ 0.0444
+ PD2	+ 0.0444
+ PD3	+ 0.1853
+ AD1	+ 0.0376
+ AD2	+ 0.0376
+ AD3	+ 0.1653
+ AD4	+ 0.0376
+ Constant	+ 0.0511
	− 0.0085*1^2)
− 0.0085*N45^2)	= 0.320

*Xie F et al. A Time Trade-off-derived Value Set of the EQ-5D-5L for Canada. Med Care 2016;54:98-105.

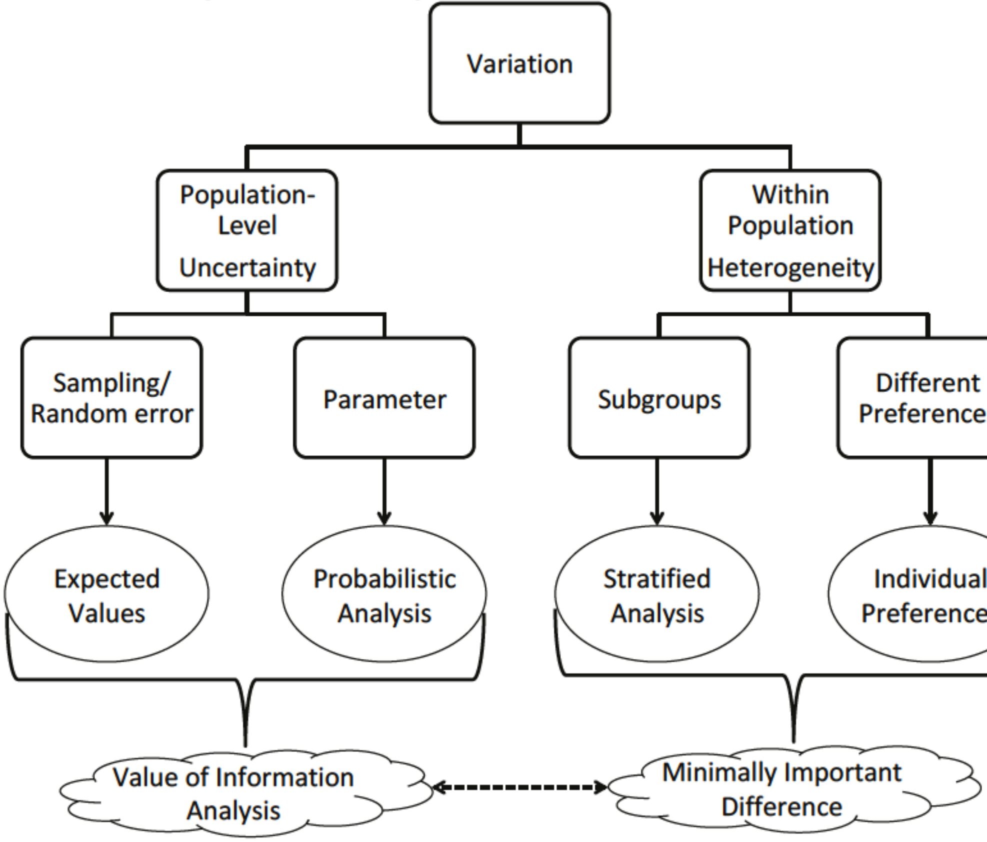
Note that the visual analogue scale, VAS, of the EQ-5D is not shown. Health state 12345 is an example response where the respondent has indicated the following: I have no problems in walking (Mobility, MO, dimension); I have slight problems washing or dressing myself (Self-Care, SC, dimension); I have moderate problems doing my usual activities (Usual Activities, UA, dimension); I have severe pain or discomfort (Pain/Discomfort, PD, dimension); I am extremely anxious or depressed (Anxiety Depression, AD, dimension). According to the Canadian scoring algorithm, the health state 12345 has an index score of 0.320, which is between the anchors of 0, dead, and 1, perfect or full health.

Figure 2. The interpretation of index scores from generic indirect preference-based measures of health-related quality of life.



From top to bottom: Index scores report the general public's preference (solid green arrows) for the changes in health reported by patients in each cohort. Cohort X and Y are groups of patients with a particular condition receiving different health technologies/interventions. Treatment effects differ between members of the cohort (solid red arrows) based on observable characteristics (e.g., sex, age, condition) to yield subgroups (S1 – S6). Individuals (i, j, k, etc.) with unique preferences compose the subgroups and cohorts (dashed grey arrows). The preferences of individuals from a representative sample (dashed black arrows) are used to derive the value set of the general public. Applied at a group-level, the minimally important difference (MID) defined as the smallest change in index score considered meaningful to patients provides feedback between the preferences of the cohort and the general public (solid blue arrows). The MID can also be applied at the level of the subgroup (blue-lines) and individual (dashed-blue lines). Note that each subgroup and individual may have a different MID.

Figure 3. Components of variation in health-related quality of life index scores and how they are addressed by economic analyses.



Boxes depict how components of variation are related, and circles describe the method that is used to address each component. From top to bottom: variation in index scores is separated into population-level uncertainty and within population heterogeneity. In turn, population-level uncertainty is composed of sampling/random error as well as uncertainty in the true value of the index score for the population (i.e., parameter). In contrast, within population heterogeneity is composed of differences in index score between subgroups as well as the fact that individuals' preferences can differ from the general public. Economic analyses address population-level uncertainty by using expected values and probabilistic analysis, while within population heterogeneity is addressed by stratified analysis, and if available, the use of individual preferences. In effect, the minimally important difference may be used as a threshold to interpret the meaning of scores that results from within population heterogeneity, while value of information analysis quantifies the value of the population-level uncertainty (i.e., cost of making the wrong decision). Finally, the minimally important difference may be used as the magnitude of desired effect to inform a value of information analysis. As well, the value of information analysis allows for the quantification of the value of determining meaningful change in patients' health-related quality of life.

2. Instrument-Defined Estimates of the Minimally Important Difference for EQ-5D-5L Index Scores

Authors: Nathan S. McClure, Fatima Al Sayah, Feng Xie[1,2], Nan Luo[3], Jeffrey A. Johnson

[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada; [2]Program for Health Economics and Outcome Measures (PHENOM), Father Sean O'Sullivan Research Centre, St. Joseph's Healthcare, Hamilton, Ontario, Canada; [3]Saw Swee Hock School of Public Health, National University of Singapore, Singapore

This has been published as McClure NS, Al Sayah F, Xie F, Luo N, Johnson JA. Instrument-defined estimates of the minimally important difference for EQ-5D-5L index scores. Value Health. 2017;20:644-650.

2.1 Abstract

Background: The five-level EuroQol five dimensions questionnaire (EQ-5D-5L) is a preference-based measure of health-related quality of life (HRQL), which yields an index score anchored at 0 (dead) and 1 (full health). We lack evidence on estimates for the minimally important difference (MID) of the EQ-5D-5L that will help in interpreting differences or changes in HRQL measured by this scale score. **Objective:** To estimate the MID of EQ-5D-5L index score for available scoring algorithms including Canada, China, Spain, Japan, England and Uruguay. **Methods:** A simulation-based approach based on instrument-defined single-level transitions was used to estimate the MID values of EQ-5D-5L for each country-specific scoring algorithm. **Results:** The simulation-based instrument-defined MID estimates for each country-specific scoring algorithm were as follows: Canada (mean = 0.056, standard deviation (SD) = 0.011), China (0.069, SD 0.007), Spain (0.061, SD 0.008), Japan (0.048, SD 0.004), England (0.063, SD 0.013) and Uruguay (0.063, SD 0.019). Differences in MID estimates reflect differences in population preferences, valuation techniques used, as well as differences in modeling strategies. After excluding the maximum-valued scoring parameters, the MID estimates were: Canada (0.037, SD 0.001), China (0.058, SD 0.005), Spain (0.045, SD 0.009), Japan (0.044, SD 0.004), England (0.037, SD 0.008) and Uruguay (0.040, SD 0.010).

Conclusions: Simulation-based estimates of the MID of EQ-5D-5L index score were generally between 0.037 and 0.069, which are similar to the MID estimates of other preference-based HRQL measures.

2.2 Introduction

The five-level EuroQol five dimensions questionnaire (EQ-5D-5L) is a generic preference-based measure of health-related quality of life (HRQL) developed by the EuroQol Group (http://www.eurqol.org) [23]. It is comprised of five dimensions: Mobility (MO), Self-Care (SC), Usual Activities (UA), Pain/Discomfort (PD), and Anxiety/Depression (AD), each with five levels [23]. A level 1 response represents "no problems", level 2 "slight problems", level 3 "moderate problems", level 4 "severe problems", and level 5 "extreme problems" or "unable to perform", the worst response in the dimension [23]. The EQ-5D-5L also asks respondents to rate their health on a visual analog scale ranging from 0 "the worst health imaginable" to 100 "the best health imaginable" [23].

The EQ-5D index score is calculated using a scoring algorithm, which is derived from preference data elicited through interviews using choice based techniques such as time trade-off (TTO) and discrete-choice experiment (DCE) tasks [25,87]. In this paper, a scoring algorithm refers to the calculation of EQ-5D-5L index score using the five responses to the questionnaire. The EQ-5D index score is anchored at 1.0 "full health" and 0 "dead", and allows for scores less than 0 representing health states that the population considers worse than being dead [25]. There are 3,125 (i.e., 5^5) possible unique health states defined by the EQ-5D-5L, with 11111 and 55555 representing the best and worst health states, respectively [87].

There is much interest in defining and estimating the minimally important difference (MID) of HRQL scores, which may serve as a useful guideline to inform the evaluation of patient reported outcomes [6,30,62,83,86,88–90]. For the purposes of this study, the MID is considered the smallest meaningful difference or change in the EQ-5D-5L index score [86,88]. While MIDs of other preference-based HRQL measures have been estimated such as the Health Utilities Index (HUI) 2 and 3, the SF-6D, and EQ-5D-3L [6,62,86,88,89], currently, there are few published MID estimates for the EQ-5D-5L scoring algorithms [91].

Similar to other HRQL instruments, the estimation of the MID for the EQ-5D index score might follow two general approaches: distribution- and anchor-based [30,83,88]. The distribution-based approach involves using statistical distributions and includes methods such as calculating an effect size or standardized response mean [30,83,88]. However, this approach has been criticized for its failure to determine whether the observed change in HRQL score is minimally important from the patient or clinical perspective [30,83,88]. In contrast, the anchor-based approach identifies an external anchor of difference or change to inform interpretation of the HRQL measure such as a measure of global rating of change or a clinical-based measure [30,83,88] . While there is no one-size fits all method to MID estimation, the general recommendation has been to use anchor-based methods employing multiple anchors with supportive information for MID estimates from distribution-based measures [30,83].

The collection of primary data for the MID estimation is time consuming and resource-intensive and may be limited by several challenges such as the choice of the anchor and its association with the HRQL score, agreement on what defines minimally important change, and the proper identification and analysis of participants who experience a meaningful yet minimal change [30]. In this regard, the estimation of MIDs for preference-based instruments using a simulated-approach based on instrument-defined health transitions has several advantages including the fact that it is based completely on the instrument's health classification system [86]. Using this method, MIDs were estimated for the EQ-5D-3L that were comparable to published estimates, demonstrating the validity of using instrument-defined health transitions to estimate MIDs [86]. The purpose of this study was to estimate the MID through simulation using instrument-defined single-level transitions for the EQ-5D-5L for scoring algorithms available from Canada, China, Spain, Japan, England and Uruguay [7,87,92–95].

2.3 Methods

2.3.1 Scoring Algorithms

For this study, we reviewed the available EQ-5D-5L scoring algorithms for each country, and selected the "best" algorithm, as identified by the respective authors [7,87,92–95]. For example, the Canadian scoring algorithm was selected based on having face validity (i.e., logically consistent index scores for all health states) as well as the lowest prediction errors and best goodness of fit as measured by mean absolute error, mean squared error and Akaike Information Criterion. The scoring algorithms of Canada, China, Japan and Uruguay were TTO derived, while the scoring algorithms for Spain and the England were based on a hybrid (TTO and DCE) approach [7,87,92–95]. Table 1 shows the country-specific algorithms, for which the scoring parameters are defined as the decrement in EQ-5D-5L index score resulting from a single-level transition to a worse health state at the specified dimension and level. This representation was used for the purposes of showcasing the difference in index score between any two adjacent levels within a dimension (not including the effects of interaction terms). Based on this representation, a reported health state of 12345 is scored as one minus the sum of the value of the parameters SC1, UA1, UA2, PD1, PD2, PD3, AD1, AD2, AD3, AD4, and the constant term (if applicable). This is sufficient to give the EQ-5D-5L index score based on the Chinese, Spanish, Japanese, England and Uruguayan scoring algorithms yielding scores of 0.336, 0.359, 0.477, 0.322, and 0.552, respectively [87,92–95]. However, since there are two dimensions at a level 4 or 5 ($N45^2 = 1$), the Canadian scoring algorithm then subtracts 1*(-0.009) to yield an index score of 0.320 [7].

2.3.2 The MID Estimation Procedure

The instrument-defined MID estimation procedure involves calculating the average absolute difference between the index score of the baseline health state and the index score of all single-level transitions from the baseline health state [86]. A single-level transition is defined as a change from the baseline health state to an adjacent (worse/better) level in a single dimension holding all other dimensions constant. For example, a baseline health state coded as 33333 has a single-level transition to a better or worse level in the first dimension resulting in health states 23333 and 43333, respectively. The set of health states defined by all possible single-level transitions from a baseline health state of 33333 is shown in vector form as {23333, 32333, 33233, 33323, 33332, 43333, 34333, 33433, 33343, 33334}. The index score of each of these health states is calculated based on the country-specific scoring algorithms [7,87,92–95]. The absolute difference in index scores between the baseline state and each single-level transition is first computed, and then averaged to yield a single MID estimate for the baseline health state. For example, using the Canadian scoring algorithm the vector of absolute differences in index score for all single-level transitions from a baseline health state of 33333 (with a baseline score of 0.577) is {0.039, 0.046, 0.020, 0.044, 0.038, 0.090, 0.104, 0.130, 0.185, 0.165}, which is then averaged to yield a single MID estimate of 0.086 (Figure 4; note, values have been rounded to three decimal places, thus the MID estimate differs as a result of rounding error). In this regard, the summarized mean MID estimate is the average of MID estimates for all 3,125 unique health states obtained through simulation. In total, there are 25,000 single-level transitions for the EQ-5D-5L.

It is however possible that certain single-level transitions result in changes or differences in EQ-5D-5L index score that may be considered larger than an MID, which are referred to as maximum-valued scoring parameters [86]. A maximum-valued scoring parameter is the largest difference in index score resulting from a change between any two adjacent levels within a single dimension (excluding the effects of interaction terms if applicable). For example, for the EQ-5D-3L, transitions between levels 2 (some/moderate problems) and 3 (extreme problems) were excluded from the MID estimation procedure because they were larger than transitions between levels 1 and 2 within a dimension [86]. Based on this reasoning we compared MID estimates that excluded transitions invoking the maximum-valued country-specific EQ-5D-5L scoring parameters within each dimension. For example, in the Canadian scoring algorithm, transitions between levels 3 (moderate problems) and 4 (severe problems) are larger than transitions between levels 1 and 2, 2 and 3, or 4 and 5 (see Table 1). Therefore, by excluding maximum-valued scoring parameters (in the Canadian scoring algorithm) the vector of possible single-level transitions from the baseline health state of 33333 is {23333, 32333, 33233, 33323, 33332} resulting in absolute differences in index score of {0.039, 0.046, 0.020, 0.044, 0.038}, which is then averaged to yield an MID estimate of

0.037 (Figure 4). The exclusion of five transitions between two adjacent levels (i.e., one scoring parameter for each dimension) reduces the total number of single-level transitions to 18,750.

Summary statistics for the MID estimates of country-specific scoring algorithms were calculated. The average absolute difference in EQ-5D-5L index score resulting from single-level transitions for every one of the 3,125 simulated baseline health states was plotted for each country-specific scoring algorithm, with a loess curve showing how the MID estimate changed as a function of baseline index score. All analysis of MID estimates was conducted using R statistical software (https://www.r-project.org/).

2.4 Results

2.4.1 Canada

The scoring parameters for the Canadian EQ-5D-5L scoring algorithm provide some information on instrument-defined estimation of the MID (Table 1). Within dimensions, with the exception of scoring parameters representing transitions between levels 3 and 4, there is a constant change in EQ-5D-5L index score resulting from single-level transitions. The UA dimension contains the minimum-scoring parameter of 0.020 representing transitions between levels 1 and 2, 2 and 3, and 4 and 5. The maximum-scoring parameter is 0.185, in the PD dimension representing the transition between levels 3 and 4. Furthermore, within all dimensions, the maximum-valued scoring parameter is for transitions between levels 3 and 4.

Figure 5a shows how the distribution of instrument-defined MID estimates changes as a function of baseline EQ-5D-5L index score. The MID estimates have a summarized mean (standard deviation [SD]) of 0.056 (0.011), and summarized median (interquartile range [IQR]) of 0.056 (0.049-0.063) (Table 2). The loess curve suggests that larger MID estimates are possible from states with middle range baseline scores (Figure 5a). Excluding single-level transitions between levels 3 and 4 (i.e., transitions with maximum-valued scoring parameters) resulted in a noticeably different distribution causing the MID estimate to become constant across the range of baseline scores with very little variation (Figure 5a), and decreased the summarized mean (SD) and median (IQR) of MID estimates to 0.037 (0.001) and 0.037 (0.037-0.038), respectively (Table 2).

2.4.2 China

The scoring parameters for the Chinese scoring algorithm suggest that there is heterogeneity among single-level transitions (Table 1). The UA dimension has the minimum-scoring parameter of 0.039, which represents transitions between levels 4 and 5. The MO dimension has the maximum-valued scoring parameter of 0.129, which represents transitions between levels 3 and 4. Within each dimension, the maximum-valued scoring parameter is also for transitions between levels 3 and 4.

The Chinese scoring algorithm had the largest MID estimate of all studied scoring algorithms. The summarized mean and median of MID estimates were approximately the same at 0.069 with a SD of 0.007 and IQR of 0.064 to 0.074 (Table 2). Figure 5b shows that smaller MID estimates are possible for states at the upper and lower limits of possible baseline scores. Excluding transitions with maximum-valued scoring parameters within each dimension shifted the distribution to lower values of MID estimates, while also causing higher baseline scores to have larger MID estimates. The summarized mean (SD) and median (IQR) of MID estimates decreased to 0.058 (0.005) and 0.058 (0.054-

0.061), respectively (Table 2).

2.4.3 Spain

The Spanish scoring algorithm uses only main effects (Table 1). The minimum-valued scoring parameter is 0 in the SC dimension representing no change in EQ-5D-5L index score resulting from the transition between levels 2 and 3. The maximum-valued scoring parameter is 0.130 in the MO dimension for transitions between levels 3 and 4. All transitions between levels 3 and 4 represent the maximum-valued scoring parameter within each dimension.

When all 25,000 single-level transitions were included, the summarized mean (SD) and median (IQR) of MID estimates were 0.061 (0.008) and 0.060 (0.055-0.066), respectively (Table 2). The flat lined loess curve suggests that MID estimates were equally spread above and below the summarized mean across the entire range of baseline scores (Figure 5c). However, when transitions between levels 3 and 4 were excluded, the distribution of MID estimates was less equally distributed over the range of baseline scores, giving rise to larger estimates for baseline EQ-5D-5L index scores above 0.6 (Figure 5c), and the summarized mean (SD) and median (IQR) of MID estimates decreased to 0.045 (0.009) and 0.046 (0.039-0.051), respectively (Table 2).

2.4.4 Japan

The Japanese scoring algorithm is solely based on main effects with little inter- and intra-dimensional variation (Table 1). The minimum-valued scoring parameter (0.024) is for transitions between levels 2 and 3 in the PD dimension. The maximum-valued scoring parameter (0.072) is for transitions between levels 1 and 2 in the AD dimension; for all other dimensions, the maximum-valued scoring parameter is for transitions between levels 3 and 4.

When including all 25,000 single-level transitions, the MID estimate for the Japanese scoring algorithm had the lowest summarized mean (SD) and median (IQR) values of 0.048 (0.004) and 0.048 (0.046-0.051), respectively (Table 2). There was also a narrow variation in MID estimates as a function of baseline-index score (Figure 5d). The summarized mean (SD) and median (IQR) decreased to 0.044 (0.004) and 0.044 (0.041-0.047), respectively, after excluding transitions with maximum-valued scoring parameters (Table 2). Generally, the MID estimate is constant across the range of baseline index scores, in which excluding transitions with maximum-valued scoring parameters shifts the distribution of MID estimates to smaller values (Figure 5d). However, there is some indication that the MID is larger for higher baseline index scores, specifically, a baseline index score of 1.0 resulted in a larger MID due to the influence of the constant term (Figure 5d).

2.4.5 England

The England scoring algorithm only includes main effects and no constant term, in addition to inter and intra dimensional variation in the magnitude of the scoring parameters (Table 1). The minimum-valued scoring parameter (0.005) is for transitions between levels 4 and 5 in the AD dimension, while the maximum-valued scoring parameter (0.194) is for transitions between levels 3 and 4 in the PD dimension. Similarly, for all other dimensions, the maximum-valued scoring parameter within each dimension is for transitions between levels 3 and 4.

The plot of MID estimates shows that the loess curve approximately follows the summarized mean MID estimate (Figure 5e). The mean (SD) and median (IQR) of MID estimates decreased from 0.063 (0.013) and 0.064 (0.055-0.073) to 0.037 (0.008) and 0.037 (0.031-0.042) respectively after excluding transitions with maximum-valued scoring parameters.

2.4.6 Uruguay

The Uruguayan scoring algorithm only has main effects with large variation between levels within a dimension and among different dimensions (Table 1). The scoring parameters range from 0.003 for transitions between levels 2 and 3 in UA to 0.191 for transitions between levels 4 and 5 in the MO dimension. Transitions between levels 4 and 5 represent maximum-valued scoring parameters for all dimensions, except PD in which the maximum-valued scoring parameter is for transitions between levels 3 and 4.

The MID estimate is negatively associated with the baseline EQ-5D-5L index score such that higher baseline index scores had even smaller MID estimates as indicated by the negatively sloped loess curve (Figure 5f). Excluding the maximum-valued scoring parameters within each dimension moderated this relationship; however, baseline index scores above 0.6 resulted in smaller MID estimates. In addition, the summarized mean (SD) and median (IQR) of MID estimates decreased from 0.063 (0.019) and 0.062 (0.050-0.076) to 0.040 (0.010) and 0.039 (0.033-0.046), respectively (Table 2).

2.5 Discussion

Our results provide estimates of the instrument-defined MID in EQ-5D-5L index scores between 0.037 and 0.069 based on scoring algorithms for Canada, China, Spain, Japan, England and Uruguay [7,87,92–95]. Excluding transitions with maximum-valued scoring parameters within each dimension decreased the summarized mean and median of MID estimates, and also affected the shape of the distribution as a function of baseline index score. Differences in MID estimates between country-specific scoring algorithms reflect differences in population preferences obtained through interview, valuation techniques used, as well as differences in modeling strategies among others.

The instrument-defined EQ-5D-5L MID estimation procedure is based on a methodology first proposed for the EQ-5D-3L [86]. According to this application, the instrument-defined MID estimates for the EQ-5D-3L were reported to be 0.040 and 0.082 for the US and England scoring algorithms, respectively [86]. Additionally, the mean empirical MID estimate from anchor-based studies was 0.075 for the US EQ-5D-3L and 0.079 for the UK EQ-5D-3L [6,62]. It was noted that the difference between these two approaches to estimate MID for the US EQ-5D-3L might result from the fact that the mean baseline index score of the study subjects was 0.8 [86]. In this case, the instrument-defined MID estimate for the US scoring algorithm was also found to vary according to baseline index score yielding a larger group-level MID estimate for a mean baseline index score of 0.8 [86]. Other reviews of empirical estimates of the EQ-5D-3L MID have found that the MID value varies according to the definition of meaningful change or difference used, the computational method applied, whether the estimate is based on the patient's judgment of improvement/change, as well as the condition and population under study [6,88]. A review of the minimal clinically important difference for the EQ-5D-3L using the UK scoring algorithm found that estimates ranged from 0.03 to 0.52; however, only six of the eighteen estimates examined were based on the patient's judgment of meaningful change [88]. While the MID estimates of EQ-5D-5L and 3L are comparable, we do not necessarily expect them to be identical. In fact, it is possible that the 5L's MID estimates may be smaller than the 3L estimates considering the ten-fold difference in the number of unique health states between the two instruments.

We found two published studies of the EQ-5D-5L index score MID based on the scoring algorithms of Japan and England [91,96]. The Japanese MID estimate of 0.061 was based on a group-level comparison of survey respondents with and without reported disease, in which the difference in index score between the two groups was reported as the MID [91]. Despite the limitations of using population norms and a cross-sectional survey to estimate the MID, the 0.01 to 0.02 difference between the Japanese estimate and our instrument-defined estimate is small. In contrast, the MID estimates reported for the scoring algorithm of England were based on multiple anchors and used a longitudinal cohort study design [96]. Moreover, the MID estimates of the EQ-5D-5L index score ranged from 0.037 to 0.063, values that are identical to the instrument-defined MID estimates [96]. In this regard, it is our hope that MIDs generated from empirical studies of the EQ-5D-5L index score will be compared to the instrument-defined MID estimates.

The results of this study need to be interpreted in light of a few considerations. Luo and colleagues suggested that the instrument-defined MID estimation procedure represents an anchor-based approach wherein each instrument-defined single-level transition acts as a reference point or criterion of minimally important change yielding an MID estimate that is based on multiple internal anchors [86]. In this regard, the scoring algorithm used in the instrument-defined MID estimation procedure provides the relative assessment that uses the population's preferences to determine the difference in scale score. Furthermore, the procedure is only relevant to multi-attribute, preference-based measures such as the EQ-5D, and not psychometrically scored HRQL measures [86]. However, some of the instrument-defined single-state health transitions assumed by the estimation procedure may not occur in reality [86]. For example, all single-state transitions from a baseline state of 55555 would involve improvements to a level 4 in each dimension, in which the assumption is made that it is equally likely to improve in the MO dimension, as it is to improve in the SC, UA, PD or AD dimensions from a level 5. Further work in the development of the instrument-defined MID estimation procedure might involve empirically estimating the likelihood or frequency of different single-state transitions from a reported EQ-5D-5L health state, which may also differ by condition, population, or intervention under study. While maximum-valued scoring parameters within each dimension were excluded based on the assumption that these represent differences/changes in index score larger than an MID, it is also possible that some transitions may represent trivial differences/changes in index score, that are less than an MID [86].

The objective of the instrument-defined MID estimate is to quantify the smallest difference in index score that might be meaningful to the patient [6,30,62,83,86,89]. In this regard, the MID may be considered relevant to clinicians, by suggesting a value that patients may place on an observed difference, whereby a clinician can better interpret the significance of an observed difference in index score of a patient or panel of patients [30,83]. The estimates obtained from the instrument-defined single-level transitions provide an MID for the EQ-5D-5L index score that does not require information from another anchor, but is instead completely specified by the instrument and its scoring algorithm [86]. Furthermore, the relationship between the instrument-defined MID of the EQ-5D-5L index score and the baseline index score gives some indication of how the instrument-defined MID estimate might vary across different populations or groups within a population. In this case, it may be useful to consider the distribution of baseline EQ-5D-5L index scores obtained in applications; this could involve estimating the instrument-defined MID from a representative sample of respondents from the general population and various patient groups of interest.

2.6 Conclusion

The simulation-based MID values of EQ-5D-5L index scores estimated using the instrument-defined single-level transitions are in general agreement with estimates of MIDs for similar preference-based measures of HRQL. The reported methodology and results may be useful in analyzing and comparing candidate scoring algorithms in combination with other anchor based methods, and in informing interpretation of EQ-5D-5L health state index scores using MID estimates.

Table 1. Country-specific scoring algorithms for the EQ-5D-5L index score showing the dimension and level specific scoring parameters [7,87,92–95].

Scoring	Country-specific scoring algorithm					
parameter	Canada	China	Spain	Japan	England	Uruguay
MO1	0.039	0.066	0.084	0.064	0.049	0.014
MO2	0.039	0.092	0.014	0.049	0.012	0.018
MO3	**0.090**	**0.129**	**0.130**	**0.066**	**0.144**	0.076
MO4	0.039	0.058	0.060	0.064	0.061	**0.191**
SC1	0.046	0.048	0.056	0.044	0.055	0.026
SC2	0.046	0.068	0.000	0.033	0.018	0.035
SC3	**0.104**	**0.095**	**0.097**	**0.048**	**0.102**	0.056
SC4	0.046	0.043	0.016	0.035	0.035	**0.157**
UA1	0.020	0.045	0.053	0.050	0.049	0.042
UA2	0.020	0.062	0.005	0.041	0.015	0.003
UA3	**0.130**	**0.087**	**0.072**	**0.057**	**0.104**	0.073
UA4	0.020	0.039	0.004	0.027	0.015	**0.113**
PD1	0.044	0.058	0.078	0.045	0.058	0.017
PD2	0.044	0.081	0.024	0.024	0.015	0.044
PD3	**0.185**	**0.113**	**0.115**	**0.063**	**0.194**	**0.126**
PD4	0.044	0.051	0.105	0.060	0.063	0.084
AD1	0.038	0.049	0.085	**0.072**	0.076	0.010
AD2	0.038	0.069	0.044	0.039	0.024	0.034
AD3	**0.165**	**0.096**	**0.121**	0.058	**0.186**	0.061
AD4	0.038	0.043	0.053	0.028	0.005	**0.073**
Constant	0.051		0.007	0.061		0.013
N45^2	-0.009					
55555	-0.148	-0.391	-0.223	-0.025	-0.281	-0.264
11111	0.949	1.0	0.993	1.0	1.0	1.0

Note. Scoring parameters are defined as the decrement in EQ-5D-5L index score resulting from a single transition to a worse health state at the specified dimension and level. Values have been rounded to three decimal places, which may differ from the number of decimal places reported in the reference. Numbers in bold indicate maximum-valued scoring parameters for each dimension.

MO, Mobility; SC, Self Care; UA, Usual Activities; PD, Pain/Discomfort; AD, Anxiety/Depression; Constant, value deducted from 1.0 for all health states except for Japan and Uruguay where it is applied when there is at least one problem in any dimension; N45^2, multiplicative variable that equals the number of 4s or 5s in any dimension past the first and squares this number; 55555, worst possible health state; 11111, best possible health state.

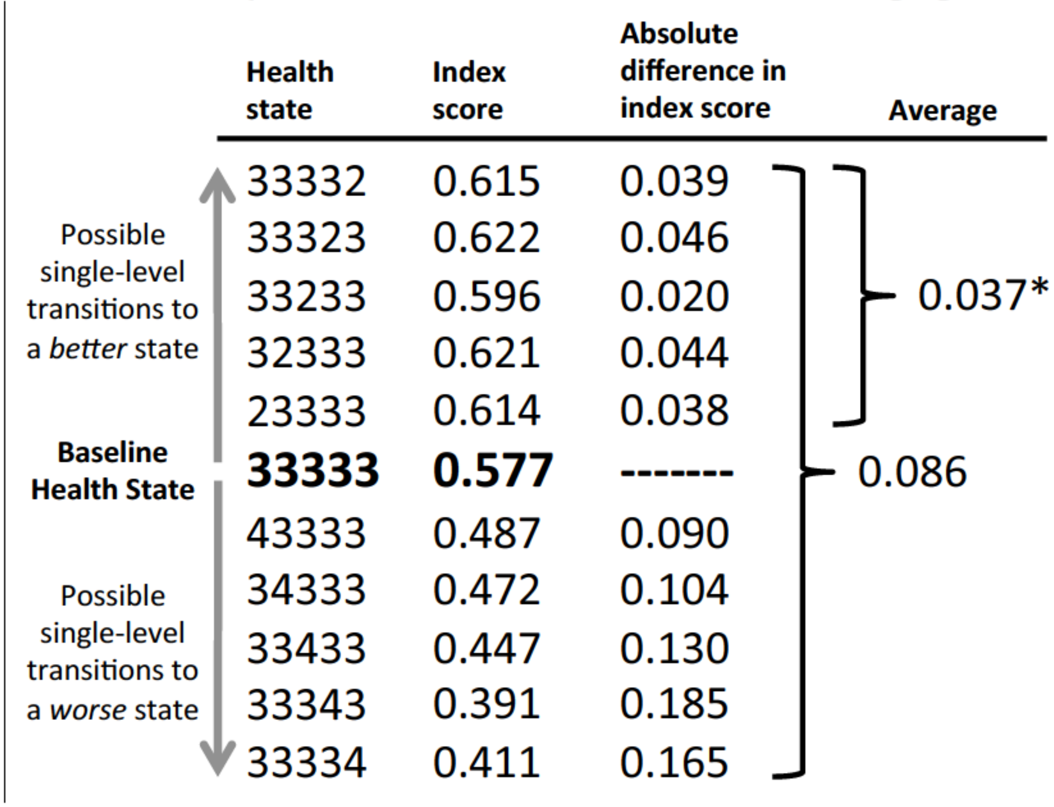
Table 2. Summary statistics of instrument-defined minimally important difference (MID) estimates for EQ-5D-5L country-specific scoring algorithms.

	Mean	SD	Median	Q1	Q3
Canada	0.056	0.011	0.056	0.049	0.063
Canada*	0.037	0.001	0.037	0.037	0.038
China	0.069	0.007	0.069	0.064	0.074
China*	0.058	0.005	0.058	0.054	0.061
Spain	0.061	0.008	0.060	0.055	0.066
Spain*	0.045	0.009	0.046	0.039	0.051
Japan	0.048	0.004	0.048	0.046	0.051
Japan*	0.044	0.004	0.044	0.041	0.047
England	0.063	0.013	0.064	0.055	0.073
England*	0.037	0.008	0.037	0.031	0.042
Uruguay	0.063	0.019	0.062	0.050	0.076
Uruguay*	0.040	0.010	0.039	0.033	0.046

Note. * denotes values after excluding maximum-valued transitions for each dimension within a country-specific scoring algorithm.

SD, standard deviation; Q1, first quartile; Q3, third quartile.

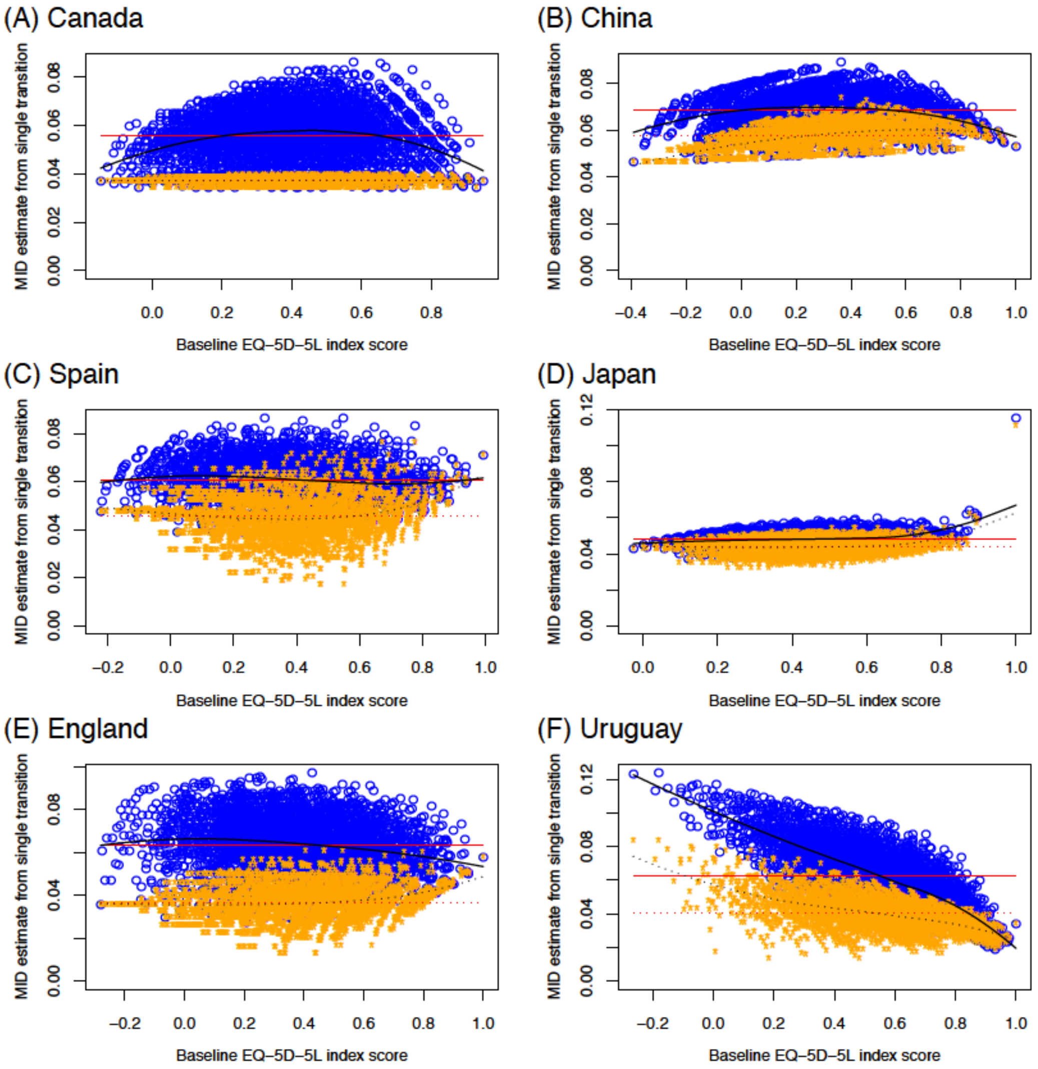
Figure 4. Example calculation of the instrument-defined minimally important difference estimate for the EQ-5D-5L index score based on the Canadian scoring algorithm.



	Health state	Index score	Absolute difference in index score	Average
Possible single-level transitions to a *better* state	33332	0.615	0.039	0.037*
	33323	0.622	0.046	
	33233	0.596	0.020	
	32333	0.621	0.044	
	23333	0.614	0.038	
Baseline Health State	**33333**	**0.577**	-------	0.086
Possible single-level transitions to a *worse* state	43333	0.487	0.090	
	34333	0.472	0.104	
	33433	0.447	0.130	
	33343	0.391	0.185	
	33334	0.411	0.165	

33333 represents a baseline health state with a level 3 response for every dimension: mobility, self care, usual activities, pan/discomfort, anxiety/depression. The asterisk (*) denotes the estimate after excluding maximum-valued transitions for each dimension. Values have been rounded to three decimal places for display purposes, thus the MID estimate differs as a result of rounding error.

Figure 5. Instrument-defined estimates of minimally important difference (MID) for the country-specific EQ-5D-5L scoring algorithm as a function of baseline index score.



Blue open circles and solid lines are for MID estimates obtained from all 25,000 single-level transitions, while orange stars (*) and dashed lines are for MID estimates with maximum-valued transitions within a dimension excluded; red lines represent the summarized mean MID estimate, and black lines are generated from a loess curve.

3. Minimally Important Difference of the EQ-5D-5L Index Score in Adults with Type 2 Diabetes

Authors: Nathan S. McClure, Fatima Al Sayah, Arto Ohinmaa, Jeffrey A. Johnson

This has been published as McClure NS, Al Sayah F, Ohinmaa A, Johnson JA. Minimally Important Difference of the EQ-5D-5L Index Score in Adults with Type 2 Diabetes. Value Health. 2018;21:1090-1097.

3.1 Abstract

Objectives: The EQ-5D is a generic preference-based measure of health-related quality of life and several studies have made attempts to estimate the minimally important difference (MID) for the EQ-5D index score. The objectives of the study are: 1) to estimate the MID of the EQ-5D-5L index score in a population-based sample of adults with type 2 diabetes; and 2) to explore whether the MID estimate varies by baseline index score and the direction of change in health status. **Methods:** We used longitudinal survey data of adults with type 2 diabetes in Alberta, Canada. The EQ-5D-5L MID was estimated first by the instrument-defined approach, which used the difference between the baseline index scores and the index scores of simulated single-level transitions. Then, by the anchor-based approach, which categorized one-year changes in depressive symptoms, diabetes-related distress as well as physical and mental health functioning into: no, small and large change groups, wherein the MID was estimated as the average change in index score of the small change group. **Results:** Based on the instrument-defined approach, MID estimates were 0.043, 0.040 and 0.045 while anchor-based MID estimates were 0.042, 0.034, and 0.049 for all change, improvement, and deterioration, respectively. Larger MID estimates were observed for lower baseline index scores and for deterioration in health status. **Conclusion:** MID estimates of the EQ-5D-5L index score were consistent between instrument-defined and anchor-based approaches, and ranged between 0.03 and 0.05. Estimates varied by baseline index score and the direction of change with similar results for patient subgroups.

3.2 Introduction

Patient-reported outcome measures (PROMs), particularly measures of health-related quality of life (HRQL), are increasingly common in routine measurement of health outcomes as a means of capturing the patients' perspective of their own health and the valuation of health services in terms of the health produced, which can be used to support patient-centred decision-making [11]. However, there are many unanswered questions that need to be addressed to facilitate the application of PROMs in health systems and to realize the full value of the collected data [11].

Diabetes is a prevalent chronic condition that can have adverse effects on a patient's health and quality of life [97]. According to the World Health Organization's 2016 global report on diabetes, it is estimated that 422 million people or 8.5% of the adult population was living with diabetes in 2014, with approximately 6 new cases per one-thousand individuals diagnosed annually in Canada [97,98]. Measuring the HRQL of patients with diabetes may be useful to understanding changes in health status, improving care, and informing healthcare investment decisions [99,100].

The EQ-5D is a generic preference-based measure of HRQL developed by the EuroQol Group (http://www.eurqol.org) [23]. The EQ-5D was developed as a brief measure of health status, and can be used for calculating quality-adjusted life years (QALYs) for use in economic studies. There is good evidence supporting its validity, reliability and responsiveness in type 2 diabetes [10,100]. However, the routine collection of EQ-5D data in health systems for quality evaluation and improvement has created a demand from end-users to interpret changes or differences in index score over time or between groups [6,11]. Commonly, statistical significance is used as a decision-criterion to interpret what may be considered meaningful change or difference in a score. While statistical significance is useful for quantifying the role of random variation giving rise to the observed change, it is not necessarily reflective of the value that the patient places on the change [11]. For example, the large sample sizes obtained by routine outcome measurement may allow for the detection of statistically significant change or differences among patients or as a result of an intervention, but this may not necessarily equate to a meaningful change or difference [101]. To this end, estimating the minimally important difference (MID) of the EQ-5D-5L index score, defined as the smallest change in index score that would be considered meaningful to the patient, may be a useful method to support the interpretability of the EQ-5D-5L [6,11,30,32,35,88,101].

Since the MID purports to capture the value that patients place on change, it may be considered specific to a patient population and/or clinical context of interest [35]. Similarly, it can be important to consider whether patients place different value on health improvement (or gain) versus health deterioration (or loss) as well as in regards to his/her baseline (or current) health status as measured by the instrument, and how these factors may affect the MID. Previous studies have suggested that the MID differs for health improvement compared to health deterioration, while patients in better health may perceive a different threshold of change as minimally important compared to those in worse health [35,83].

The primary objective of this study is to estimate the MID of the EQ-5D-5L index score in a representative sample of adults with type 2 diabetes in Alberta, Canada. The secondary objectives are (1) to explore whether the MID estimate varies by baseline index score, (2) to examine whether it varies by the direction of change in health status (improvement versus deterioration), and (3) to determine MID estimates of defined patient subgroups.

3.3 Methods

3.3.1 Data Source

Data (N=1927) were from baseline and one year follow-up of an ongoing cohort of adults with type 2 diabetes in Alberta, Canada (Alberta's Caring for Diabetes cohort study). Details of the study have been reported elsewhere [99]. Briefly, the study aims to research the various factors associated with the development of disease complications and other health outcomes in patients with type 2 diabetes by gathering information on individual medical, behavioral and psychosocial factors [99]. Participants completed a mailed self-reported survey, which included questions on sociodemographic factors (i.e., age and sex), diabetes-history, comorbidities, diabetes complications, self-care and diabetes-specific management, well being, and HRQL measures including the EQ-5D-5L [99]. The items wording of selected measures is presented in the Appendix.

3.3.2 Measures

EuroQol 5-Dimensional 5-Level Questionnaire (EQ-5D-5L)

The EQ-5D-5L is based on a multi-attribute health classification system that includes five response levels for five dimensions: mobility (MO), self-care (SC), usual activities (UA), pain/discomfort (PD), and anxiety/depression (AD) [23]. According to the Canadian scoring algorithm, EQ-5D-5L scores range between -0.148 for the worst health state (55555) and 0.949 for the best health state (11111) [7].

Patient Health Questionnaire 8-items (PHQ8)

The patient-health questionnaire is an eight-item questionnaire measuring depression in which each item is scored on a 4-point Likert Scale from 0 (not at all) to 3 (nearly everyday), and with a reference period of the last two weeks. The sum of all eight items yields a total score ranging from 0 (no depressive symptoms) to 24 (severe depressive symptoms).

Problem areas in Diabetes 5-items Questionnaire (PAID5)

The problem areas in diabetes questionnaire measures emotional functioning with 5-items by way of a 5 point Likert Scale from 0 (not a problem) to 4 (serious problem) at the present time. The average of all five items gives a final score ranging between 0 and 4 wherein higher scores indicate more problems.

SF-12

The short form health survey (version 2) is a twelve-item questionnaire with eight subscales: physical functioning, energy and vitality, role limitations due to physical problems, role limitations due to emotional problems, bodily pain, general health perceptions, social functioning, and mental health [102]. The reference period of the items varies from the present moment to as long as the past 4 weeks. The subscale scores are used to calculate two summary scores: the physical component score (PCS), and mental component score (MCS). The summary scores are transformed to a scale score ranging from 0 (worst score) to 100 (best score) with a mean of 50 and a standard deviation of 10.

3.3.3 MID Estimation

We used the instrument-defined and the anchor-based approaches to estimate the MID of the EQ-5D-5L index score in the overall sample and in patient subgroups based on sex, and splits at approximately the median value for age, duration of diabetes, and number of comorbidities. In addition, changes in EQ-5D-5L health states were analysed using a Paretian classification method [81].

Instrument-defined approach: This approach is based on the average of index score differences between the baseline health state and single-level transitions to other health states. Further details of the instrument-defined approach have been published elsewhere [86,103]. Instrument-defined MID estimates were calculated for baseline EQ-5D-5L profiles using all single-level transitions (*all*), only transitions to a better state (*improve*), or only transitions to a worse state (*deteriorate*). Because a proportion of the sample had baseline EQ-5D-5L profiles of 11111 (or the maximum index score), these individuals had no transitions to a better state and were therefore excluded from the *improve* instrument-defined MID estimate. In the Canadian scoring algorithm, transitions between level 3 (i.e., moderate problems) and level 4 (i.e., extreme problems) in the MO, SC, UA, PD, and AD dimensions represent as much as 2.31, 2.28, 6.66, 4.17, and 4.40 times the value of other scoring parameters within each respective dimension. Based on the assumption that transitions involving a maximum-valued scoring parameter within a dimension constitute a difference or change in index score that is larger than an MID, we excluded the maximum-valued scoring parameters within each dimension. Therefore, instrument-defined MID estimates were calculated using all single-level transitions (idMID) and excluding maximum-valued scoring parameters (idMID*).

Anchor-based approach: As there is no "one-size-fits-all" method to estimating the MID of HRQL scores, it is generally considered good practice to estimate the MID using multiple approaches and within each approach to use different methods (i.e., anchors and distribution parameters) to yield a pooled or triangulated MID estimate and/or a plausible range [30]. In this way, the anchor-based approach used multiple anchors with distribution-based cut-offs. Specifically, one-half standard deviation of the baseline anchor score was used as the lower cut-off of small change for each anchor with an upper cut-off of two times the lower cut-off [84,104,105]. The anchor-based approach involved first categorizing the change scores of the anchors (PHQ8, PAID5, PCS and MCS) at follow-up into three groups: no change (< ½ standard deviation, SD, at baseline), small change (≥ ½ and ≤ 1 SD), and large change (> 1 SD). MID estimates were calculated as the average change in EQ-5D-5L index score of the small change group. The *all* MID estimate included minimal important change for both improving and worsening anchor change scores, in which worsening scores were multiplied by negative one (-1). Otherwise, MID estimates were categorized as *improve* or *deteriorate* if they included only a subset of the data representing improving or worsening anchor change scores respectively. A pooled MID estimate was then calculated as the average across all anchor-based MID estimates.

MID estimate as a function of baseline index score: Anchor-based MID estimates and mean baseline index scores were calculated using a combined moving average and loess (i.e., local regression) smoothing approach [106]. This involved ordering the dataset based on baseline index score (from lowest to highest score) and taking multiple sequential sub-samples that included at least 20 percent of the total baseline dataset. The group of individuals with the lowest baseline index score in each sub-sample was then not used to calculate the next sub-sample (i.e., the next 20% of the baseline data). For each sub-sample, the baseline index score was calculated as the mean of the sub-sample, and the MID was estimated for each anchor using the same methodology as previously described. The instrument-defined MID estimates were similarly represented as the MID estimate of each sub-sample. Scatter-plots were generated to show the loess curve line of the MID estimate as a function of baseline index score.

Effect Size and Standardized Response Mean: The effect size (ES) and standardized response mean (SRM) were calculated by dividing the MID estimate (numerator) by the standard deviation of the EQ-5D-5L index score at baseline (ES) and, for anchor-based MID estimates, the standard deviation of the EQ-5D-5L change score (SRM). With respect to the ES estimates, we are mainly interested in small effect sizes (i.e., between 0.2 and 0.5) to show that the MID is the smallest meaningful change or difference in index score.

Depending on the response rate to each anchor questionnaire, between 19% and 33% of patients had incomplete follow-up information at one year (Table 3). The subset with missing follow-up information had worse baseline anchor and EQ-5D-5L index scores compared to the entire dataset at baseline (Table 3). Results reported for anchor-based MID estimates were based on complete-case (i.e., complete EQ-5D and anchor score information). All analyses were conducted using R statistical software (The R Foundation, Vienna, Austria).

3.4 Results

The median age of participants at baseline (N=1927) was 64.7 years (interquartile range [IQR] 57.2–72.2 years), and 45% were female (Table 3). The median duration that participants had lived with diabetes was 10.7 years (IQR 5.3–16.8 years), and reported a median of 4.0 (IQR 3.0 – 6.0) comorbidities. At baseline there were 281 unique EQ-5D-5L health states reported with an average EQ-5D-5L index score of 0.79 (SD 0.17) decreasing to 239 unique health states at one-year follow-up but with a similar average index score and SD (Table 3). The mean (and SD) of the anchor scores at baseline were 5.3 (5.4), 0.87 (0.88), 47.9 (9.8), and 46.0 (10.8) for PHQ8, PAID5, MCS and PCS respectively. The strength of association between change in anchor score and change in EQ-5D-5L index score varied across the anchor measures (Figure 6), with correlations of 0.41 (95% CI: 0.35–0.48) for PHQ8, 0.27 (0.20–0.33) for PAID5, 0.45 (0.39–0.50) for MCS, and 0.51 (0.45–0.55) for PCS.

The Paretian classification of health change shows a large proportion of changes in EQ-5D-5L health states are ambiguous (i.e., mixed) within the small and large anchor change groups (Table A1 in Appendix). Furthermore, the distribution of EQ-5D-5L health states by level and dimension (Figures A1 and A2 in Appendix) as well as the density of index scores (Figure A3 in Appendix) change from baseline to one-year follow-up, and vary among anchors and by the direction of change.

3.4.1 MID Estimate for "All" Changes

Based on the instrument-defined estimation method, the overall MID estimate ranged between 0.037 and 0.049, with an ES of 0.22 to 0.28 in the overall sample (Figure 7, Table 4). MID estimates (0.037–0.053) and ES (0.20–0.40) were consistent across examined sub-groups (Table A2 in Appendix). Based on the anchor-based approach, the MID estimates ranged from 0.031 to 0.057 (pooled estimate 0.042), with an ES of 0.18 to 0.33 (pooled ES 0.25) in the overall sample (Figure 7, Table 4), and pooled MID estimates ranged from 0.033 to 0.047 (pooled ES: 0.22–0.30) across sub-groups (Table A2 in Appendix).

3.4.2 MID Estimate for "Improve" Changes

The MID estimate for improvement in the overall sample ranged between 0.038 and 0.043 with an ES between 0.22 and 0.25 based on the instrument-defined approach (Figure 7, Table 4). Using the anchor-based approach, the MID estimates for improvement were between 0.021 and 0.054 (pooled MID estimate of 0.035) with an ES from 0.12 to 0.31 (pooled ES 0.20). In contrast, the examined subgroups had an instrument-defined MID ranging between 0.037 and 0.045 (ES 0.19–0.38), and pooled anchor-based MID estimates from 0.027 to 0.040 (pooled ES 0.16–0.26) (Table A2 in Appendix).

3.4.3 MID Estimate for "Deteriorate" Changes

For the overall sample, MID estimates for worsening health ranged from 0.038 to 0.053 with an ES of 0.22 to 0.31 using the instrument-defined approach (Figure 7, Table 4). The anchor based approach gave MID estimates from 0.029 to 0.060 (pooled MID estimate 0.049) with an ES of 0.17 to 0.35 (pooled ES 0.29). For the examined subgroups, similar MID estimates and ES were observed using the instrument-defined approach (MID: 0.037–0.059; ES: 0.20–0.42), while the anchor-based approach gave pooled MID estimates ranging from 0.037 to 0.056 (pooled ES: 0.26–0.33) (Table A2 in

Appendix).

3.4.4 MID Estimate as a Function of Baseline Index Score

The relationship between the instrument-defined MID estimate and the baseline index score depended on whether or not maximum-valued scoring parameters were excluded. When excluded, the MID estimates remained at a constant value of approximately 0.037 across the range of baseline index scores (Figure 8). Conversely, when all single-level transitions were included, the MID estimate started at a larger value for lower baseline index scores and decreased to a minimum of 0.037 as the baseline score approached its upper limit. Specifically, for improvement in health, the maximum MID value of 0.058 was at the minimum baseline index score of 0.54, and declined to the minimum MID value of 0.036 at a baseline score of 0.76 (Figure 8). This differs from the deterioration case, where the MID estimates started at approximately 0.074 at a baseline index score of 0.54, before peaking at approximately 0.084 at a baseline score of 0.67, then gradually declined to 0.036 at a baseline score of 0.93.

The anchor-based *improve* MID estimates start at larger values than the instrument-defined MID estimate (0.071 to 0.113; pooled estimate of 0.092 at a baseline index score of 0.55) before quickly descending (with some indication of a levelling off of estimates before dropping off again) to negative minimum values (-0.006 to -0.041; pooled estimate of -0.020) at the upper-limit of baseline-index scores (0.941; see Figure 8). For *deteriorate* MID estimates, there appears to be some consistency among the PAID5 and SF-12 PCS anchors and the instrument-defined estimates; however, there is an observable divergence at higher baseline index scores (Figure 8). In contrast, the PHQ8 and SF-12 MCS anchors display a relationship in which the *deteriorate* MID estimate increases for increasing baseline index score (Figure 8).

3.5 Discussion

This study provides evidence that the MID of the EQ-5D-5L index score in adults with type 2 diabetes is in the range of 0.03 to 0.05. The instrument-defined and anchor-based approaches represent two distinct methods of MID estimation for the EQ-5D-5L index score. Anchor-based MID estimates were generally consistent with instrument-defined MID estimates, for which differences in MID estimates were observed according to baseline index score and direction of change.

When including all single-level transitions in the instrument-defined approach, the shape of the curve relating the MID estimate to the baseline index score may be interpreted as follows: at the extremes of the baseline index score (i.e., near -0.148 and 0.949), the MID is more representative of change in a single direction, in which it is important to interpret a small change in the possible direction as meaningful, yielding a small MID estimate. As we move away from the extremes of the baseline index score range, we expect the overall MID to reflect an increasing mixture of transitions to worse and better health states. Consequently, the overall MID estimate becomes larger as the baseline index score moves toward intermediate values (i.e., near 0.5). In summary, by including all possible instrument-defined single state transitions, we are relying completely on the instrument (and its scoring algorithm) to determine the MID estimate, which in this case, suggests that larger MID estimates are associated with intermediate baseline index scores. The larger MID estimates for worsening health compared to MID estimates for improving health suggest that for the same baseline index score a patient may consider a smaller change to a better health state as important while the same magnitude of change to a worse health state may be unimportant or trivial.

The anchor-based MID estimates also show differences based on baseline index score and direction of health state change; however, there is greater variability among anchors likely as a result of how changes in an anchor are reflected in the EQ-5D descriptive system (e.g., changes in the PHQ8 may be reflected more in the AD dimension than in other dimensions). Comparison of anchor-based MID estimates to the instrument-defined MID estimates for *all*, *improve* and *deteriorate* (as well as across subgroups) shows that there is reasonable agreement. Based on the 95% bootstrap confidence limits, there is evidence that the PAID5 anchor-based MID estimates have the greatest uncertainty, which is likely attributable to the low correlation between change in anchor score and change in EQ-5D-5L index score. There is an observable ceiling effect in which 14.3% to 17.3% of respondents self-reported “no problems” in all dimensions (i.e., 11111), thus further improvement in health for these individuals cannot be reflected by the EQ-5D descriptive system. The ceiling-effect of the EQ-5D-5L is noticeable in the anchor-based *improve* MID estimates, which show a sharp decline for increasing baseline index score becoming negative at the upper limit (i.e., reflecting a decrease in index score despite improvement in health as measured by changes in anchor scores). In contrast, the instrument-defined approach excluded health states 11111 in *improve* MID estimates.

A number of limitations warrant consideration. First, the results are based on complete case analysis representing 67% to 81% of the original cohort depending on response and completion rates of each anchor questionnaire. There is evidence that those individuals who were lost to follow-up were in a poorer health state at baseline, suggesting that an assumption of missing-at-random is unlikely. However, since anchor-based MID estimates are based on a priori-defined small change groups, it may be reasonable to assume that the small change group had less loss to follow-up (than the large change group) so as to not adversely impact the MID estimates. Furthermore, the instrument-defined MID estimates use only the baseline information and are therefore unaffected by attrition. Second, it is important to consider that while we treat the MID estimate as a "threshold" of meaningful change, this threshold cannot suggest or indicate how "difficult" it is to achieve a change score at least as large as the MID estimate. Finally, while this study has used several methods to quantify what a patient may consider as the smallest meaningful change in index score, it is important to further test and validate estimates using other methods and in other clinical contexts.

3.6 Conclusion

The instrument-defined approach can be a useful method of MID estimation in a specific patient population. This provides a plausible range of smallest change in index score that may be considered meaningful to the patient. Furthermore, the results suggest that the MID for health improvement is less than that for health deterioration as well as decreasing for higher baseline index scores, which proposes that researchers ought to consider these issues when interpreting study results. However, it is unknown if these phenomena are unique to patients with diabetes and/or the Canadian value set. Further research that seeks patient input directly is needed to determine what patients consider the smallest meaningful change in index score.

Table 3. Descriptive statistics of sample at baseline and follow-up.

	Baseline									Follow-up
		Data specific to anchor (complete-case | _LTFU_)								
	All	PHQ8		PAID5		SF-12 MCS		SF-12 PCS		
Sample size	1927	1428	_499_	1432	_495_	1288	_639_	1288	_639_	1560
Age median IQR	64.7 57.2–72.2	64.9 58.1–72.0	_63.0_ _55.1–72.7_	64.9 58.0–72.1	_63.4_ _55.1–72.4_	64.6 57.7–71.2	_65.2_ _56.2–74.3_	64.6 57.7–71.2	_65.2_ _56.2–74.3_	66.2 59.4–73.4
%Female	44.9%	45.9%	_42.3%_	46.0%	_41.8%_	45.8%	_43.2%_	45.8%	_43.2%_	43.2%
Diabetes duration median IQR	10.7 5.3–16.8	10.8 5.6–16.7	_9.0_ _4.8–18.5_	10.8 5.6–16.7	_9.2_ _4.8–18.1_	10.8 5.5–16.6	_10.1_ _5.0–18.5_	10.8 5.5–16.6	_10.1_ _5.0–18.5_	10.8 5.6–16.8
No. comorbidities median IQR	4 3–6	4 2–6	_4_ _3–6_	4 2–6	_4_ _3–6_	4 3–6	_4_ _2–6_	4 3–6	_4_ _2–6_	4 2–6
EQ-5D-5L No. Health states 11111 55555 Index score mean SD	281 15.9% 0.1% 0.790 0.171	220 16.9% 0.0% 0.802 0.161	_159_ _13.0%_ _0.2%_ _0.758*_ _0.194_	219 16.8% 0.0% 0.802 0.160	_157_ _13.5%_ _0.2%_ _0.758*_ _0.196_	209 17.3% 0.0% 0.804* 0.160	_178_ _13.1%_ _0.2%_ _0.762*_ _0.189_	209 17.3% 0.0% 0.804* 0.160	_178_ _13.1%_ _0.2%_ _0.762*_ _0.189_	239 14.3% 0.0% 0.792 0.170
PHQ8 mean SD	5.3 5.4	4.9* 5.1	_6.6*_ _6.0_	4.9* 5.1	_6.6*_ _6.0_	4.9* 5.1	_6.3*_ _5.9_	4.9* 5.1	_6.3*_ _5.9_	5.1 5.1
PAID5 mean SD	0.867 0.880	0.813 0.831	_1.024*_ _0.994_	0.817 0.833	_1.018*_ _0.996_	0.817 0.831	_0.968*_ _0.966_	0.817 0.831	_0.968*_ _0.966_	0.796 0.821
SF-12 MCS mean SD	47.9 9.8	48.6* 9.6	_45.5*_ _10.0_	48.6* 9.6	_45.6*_ _10.0_	48.7* 9.7	_45.7*_ _9.7_	48.7* 9.7	_45.7*_ _9.7_	48.1 9.8
SF-12 PCS	46.0	47.2*	_41.6*_	47.2*	_41.6*_	47.3*	_41.6*_	47.3*	_41.6*_	46.5

mean SD	10.8	10.6	*11.1*	10.5	*11.2*	10.4	*11.3*	10.4	*11.3*	10.5

All, entire baseline dataset; LTFU, lost to follow-up; EQ-5D-5L, EuroQol five-dimensional five-level questionnaire; PHQ8, patient health questionnaire 8 items; PAID5, problem areas in diabetes 5 items; SF-12, short-form medical survey 12 item; MCS, mental health component score; PCS, physical component score; No., number of. Note that values shown in italics are specific to the data subset that was lost to follow-up for each anchor; asterisk [*] denotes a statistically significant difference (p-value < 0.05) based on a two sample t-test compared to the entire baseline dataset.

Table 4. Minimally important difference estimates of the EQ-5D-5L index score by estimation method and direction of change.

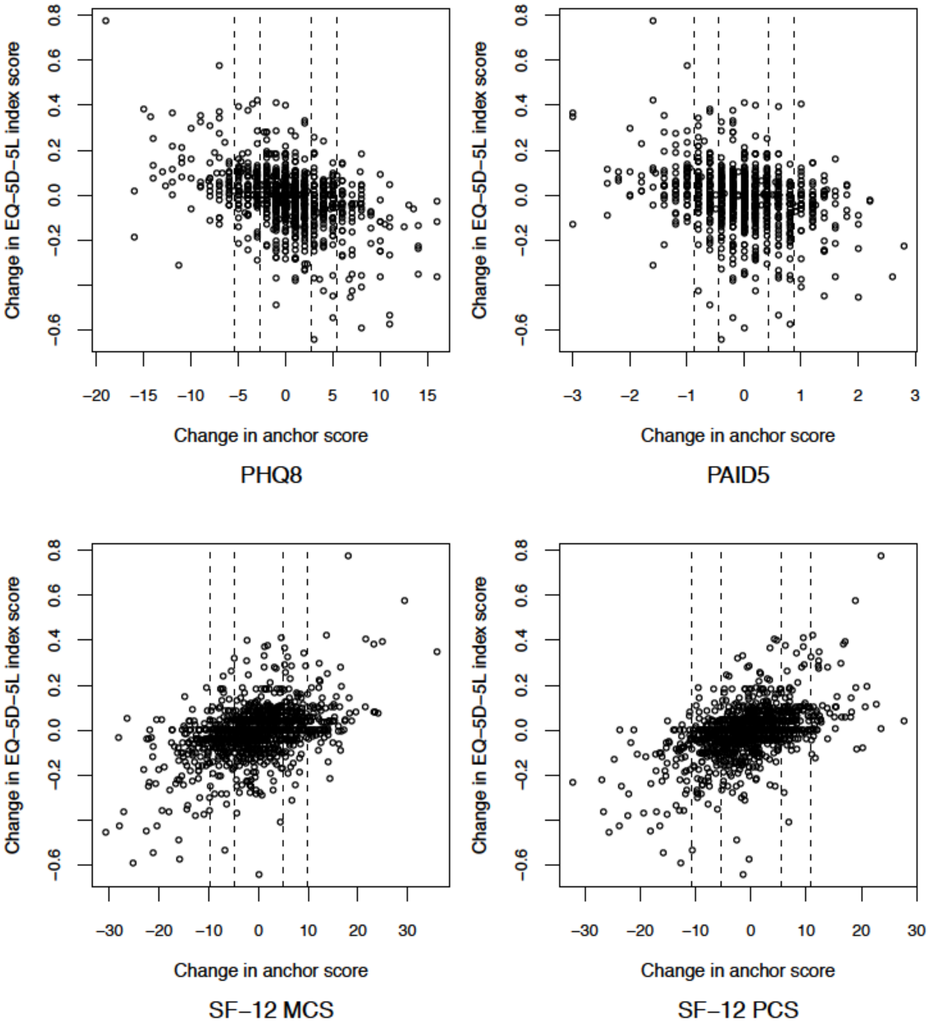
Direction of change	Method	Sample size	Mean±SD	MID	95% CI	ES	SRM
All	Instrument-defined:						
	idMID	1927	----	0.049	0.048–0.049	0.285	0.434
	idMID*	1927	----	0.037	0.037–0.037	0.217	0.331
	Anchor-based:						
	PHQ8	301	0.012±0.086	0.043	0.030– 0.056	0.251	0.382
	PAID5	266	-0.001±0.096	0.037	0.022–0.053	0.219	0.333
	SF-12 MCS	318	0.015±0.087	0.031	0.021–0.042	0.184	0.281
	SF-12 PCS	294	0.016±0.085	0.057	0.046–0.069	0.334	0.509
	Pooled	----	0.010±0.089	0.042	0.030–0.055	0.247	0.376
Improve	Instrument-defined:						
	idMID	1927	----	0.043	0.043–0.044	0.254	0.386
	idMID*	1927	----	0.038	0.037–0.038	0.220	0.335
	Anchor-based:						
	PHQ8	137	0.004±0.081	0.031	0.012– 0.048	0.183	0.279
	PAID5	136	-0.009±0.094	0.021	0.000–0.042	0.122	0.186
	SF-12 MCS	148	0.012±0.091	0.034	0.018–0.050	0.198	0.301
	SF-12 PCS	129	0.014±0.075	0.054	0.036–0.071	0.314	0.478
	Pooled	----	0.005±0.085	0.035	0.017–0.053	0.204	0.311
Deteriorate	Instrument-defined:						
	idMID	1927	----	0.053	0.052–0.054	0.312	0.475
	idMID*	1927	----	0.038	0.037–0.038	0.220	0.335
	Anchor-based:						
	PHQ8	164	0.012±0.083	0.053	0.034–0.071	0.308	0.469
	PAID5	130	0.012±0.089	0.055	0.034–0.077	0.320	0.487
	SF-12 MCS	170	0.019±0.083	0.029	0.014–0.045	0.173	0.263

	SF-12 PCS	165	0.018±0.093	0.060	0.046–0.076	0.350	0.533
	Pooled	----	0.015±0.087	0.049	0.032–0.067	0.287	0.438

Note. Values for direction of change *deteriorate* have been multiplied by −1; sample size is the number of respondent scores used in the calculation of the MID; mean ± SD represents a statistic for the no change group (anchor-based); the no change group by direction of change includes responses with anchor change scores between 0 and the corresponding limit of change (i.e., trivial improvement/deterioration).

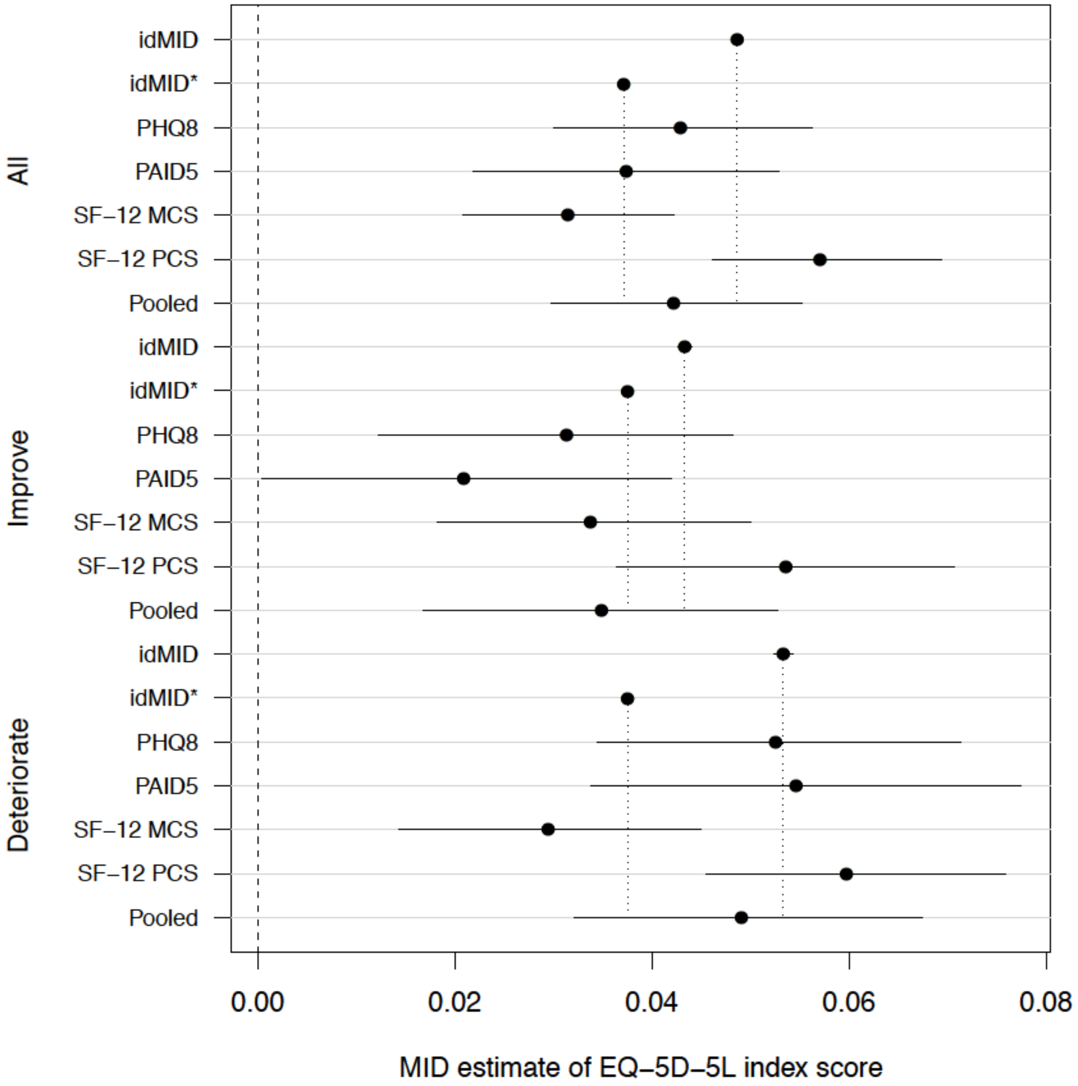
CI, confidence interval (based on 1000 bootstrap replicates); EQ-5D-5L, five-level EuroQol five-dimensional questionnaire; ES, effect size; idMID, instrument-defined minimally important difference (*excluding maximum-valued scoring parameters); MCS, mental component summary; MID, minimally important difference; PAID5, problem areas in diabetes 5-item; PCS, physical component summary; PHQ8, patient health questionnaire 8-item; Pooled, average of anchor-based estimates; SD, standard deviation; SF-12, 12-item short form health survey; SRM, standardized response mean.

Figure 6. The association between change in anchor score and change in EQ-5D-5L index score.



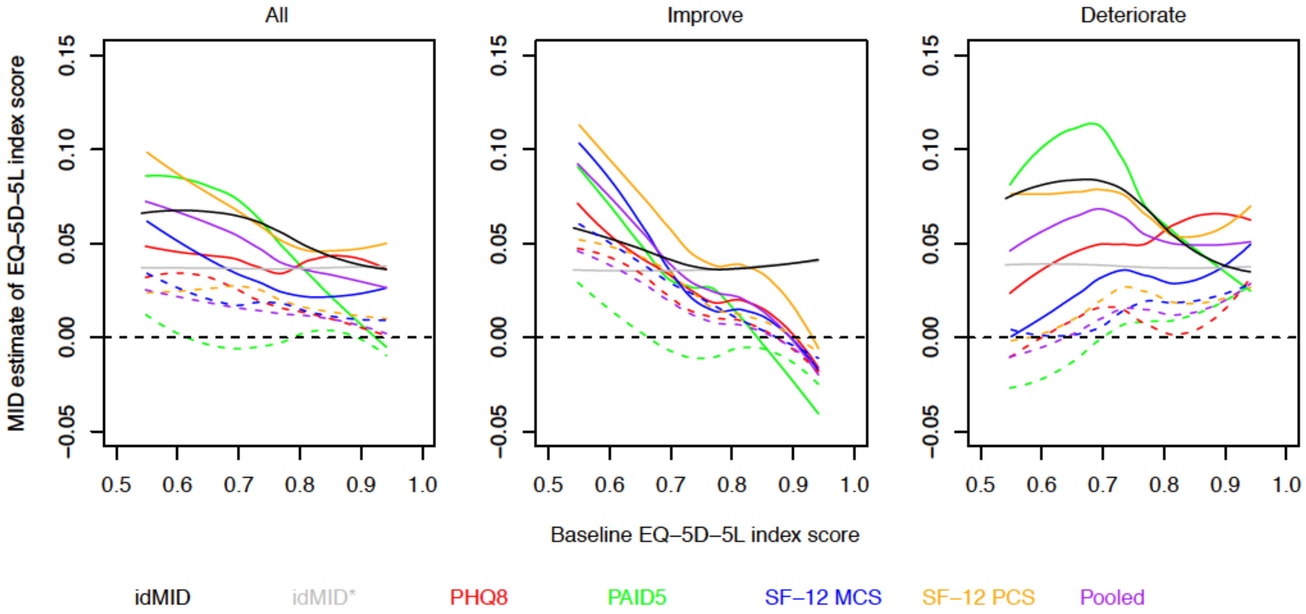
The dotted lines represent the limits of the small change group based on the standard deviation of the baseline anchor score. EQ-5D-5L, EuroQol five-dimensional five-level questionnaire; PHQ8, patient health questionnaire 8 items; PAID5, problem areas in diabetes 5 items; SF-12, short-form medical survey 12 item; MCS, mental health component score; PCS, physical component score.

Figure 7. Minimally important difference estimates of the EQ-5D-5L index score by estimation method and direction of change.



Point estimates (solid dots) and 95 percent confidence intervals based on 1000 bootstrap replicates (solid lines) for all change, and the direction of change (improve versus deteriorate). MID, minimally important difference; EQ-5D-5L, EuroQol five-dimensional five-level questionnaire; idMID, instrument-defined minimally important difference (*excluding maximum-valued scoring parameters); PHQ8, patient health questionnaire 8 items; PAID5, problem areas in diabetes 5 items; SF-12, short-form medical survey 12 item; MCS, mental health component score; PCS, physical component score; Pooled, average of anchor-based estimates. Note that values for direction of change "deteriorate" have been multiplied by negative one (-1).

Figure 8. Minimally important difference estimates of the EQ-5D-5L index score, and average change in index score of the no change group as a function of baseline index score, by estimation method and direction of change.



Lines are based on local regression (loess) curves of estimates from ordered subsets comprising at least 20% of the baseline data; solid lines represent MID estimates; dashed lines represent average change in index score of the “no change” group. MID, minimally important difference; EQ-5D-5L, EuroQol five-dimensional five-level questionnaire; idMID, instrument-defined minimally important difference (*excluding maximum-valued scoring parameters); PHQ8, patient health questionnaire 8 items; PAID5, problem areas in diabetes 5 items; SF-12, short-form medical survey 12 item; MCS, mental health component score; PCS, physical component score; Pooled, average of anchor-based estimates. Note that values for direction of change “deteriorate” have been multiplied by negative one (-1); the “no change” group by direction of change includes responses with anchor change scores between 0 and the corresponding limit of change (i.e., trivial improvement/deterioration).

4. Unpacking Small Differences in EQ-5D-5L Health Utility Scores: Are You Better, Worse or the Same?

Authors: Nathan S. McClure, Feng Xie[1,2], Mike Paulden, Arto Ohinmaa, Jeffrey A. Johnson

[1]Department of Health Research Methods, Evidence, and Impact, McMaster University, Canada; [2]Centre for Health Economics and Policy Analysis, McMaster University, Canada

4.1 Abstract

Background: At a macro-level there is support for using the average of the general public's preferences to measure health-related quality of life (HRQL) trade-offs. However, since the general public is composed of different individuals, it is unclear to what extent population-based health utility scores (HUS) reflect individuals' stated health preferences. This study seeks to determine the level of agreement between EQ-5D-5L HUS differences and individuals' (ex ante) health preferences using responses from time trade-off (TTO) and discrete-choice experiment (DCE) tasks (n=1073).

Methods: First, participants' TTO responses were transformed into pairwise comparisons to yield ordinal TTO (oTTO) responses. Then, three mixed-effect logistic regression models were used to construct curves representing the average probability that participants consider a pairwise comparison between health states to be the same, worse or better. Second, the average pairwise differences in participants' observed TTO values were stratified according to their oTTO responses to determine how (small) EQ-5D-5L HUS differences differ from participants' stated cardinal preferences. Thirdly, the predicted probabilities from the oTTO responses were compared to predictions from DCE responses using a linear model. Lastly, the probability that HUS differences represent participants' majority ordinal preferences was investigated.

Results: Probability curves show that HUS differences near 0 have as much as 30.6% (95% confidence interval: 29.1-31.9%) probability of representing a tie in individuals' TTO values. Differences in EQ-5D-5L HUS of -0.054 (-0.071 to -0.029) and 0.047 (0.026 to 0.076) maximized the sensitivity and specificity of discriminating a transition to a worse and better health state, respectively. Small differences in HUS of +/-0.03 to +/-0.07 had average TTO differences of -/+0.17, and -/+0.35 whether ties were included or excluded, respectively. When ties were included in the oTTO responses, the slope coefficient of predicted probabilities regressed on DCE probabilities were 0.81 (0.79-0.82); however, if ties were excluded, the slope coefficient increased to 0.99 (0.97-1.00). Absolute HUS differences in the range of 0.043 to 0.064 had a 50% probability of representing respondents' majority ordinal preferences.

Conclusions: Differences in population-based HUS for the EQ-5D-5L, particularly small differences, are not uniformly representative of individuals' stated preferences. This suggests that a difference in health utility score needs to be large enough (e.g., ≥ +/-0.05) to overcome heterogeneity in preferences, lending support to the application of a minimally detectable or important difference for decision-making.

4.2 Introduction

Generic indirect preference-based measures of health-related quality of life (HRQL) are used to inform macro-level healthcare decision-making in regards to the allocation of scarce societal resources [20]. Since everyone is a potential patient whose condition and treatment are uncertain, use of health utility scores (HUS) that reflect the average of the general population's preferences for health states is recommended [27,38]. However, at a micro-level, such as in a patient-clinician encounter, shared decision-making promotes understanding individuals' HRQL so to inform on the most appropriate care [18,48,107]. The EQ-5D is one of the most commonly used generic measures of HRQL in the world [9]. The EQ-5D HUS is increasingly reported and assessed in a variety of contexts, which in turn, has brought challenges to its interpretation [10,11].

Methods that can bridge the gap between observed changes in HUS and expected individual-level changes may be useful to end-users of the EQ-5D (or other generic HRQL instruments), especially in non-economic assessments of HRQL changes [20]. Various methods have been proposed to assist in interpretation of generic HRQL scores at the point of application including the concept of the minimally important difference (MID) [35]. Defined as the smallest change in score that is meaningful to patients, the MID is arguably the most patient-centred approach [31,33,37]. Overall, MID estimates for generic HRQL scores suggest that very small differences/changes in scores (i.e., < MID) are not expected to represent minimally important improvement or deterioration in HRQL from the patients' perspective [32,80]; however, the applicability of the MID is not without controversy [108–110]. The present study presents a novel investigation of individuals' stated preferences to determine if and how the concept of a minimally important difference may be relevant to a generic preference-based measure of HRQL.

In a homogenous population, where everyone has the same preferences for health, we would expect there to be no difference between the general population's preferences and any one individual's preferences. However, in a heterogeneous population there is likely to be variation among individuals' preferences, and thus differences between individuals' preferences and the average of the population's preferences [59,111]. When there is a lack of congruency between the average of the population's preferences and individuals' preferences, it may be useful to consider the heterogeneity or uncertainty in HUS differences in terms of how the observed score difference reflects individuals' preferences. Specifically, it may be useful to quantify the smallest difference in HUS that can be expected to represent what individuals consider to be an improvement or deterioration in health status (on average).

The normative approach to elicitation of health preferences requires a choice-based task [38]. Recent valuation studies for the EQ-5D-5L [112], such as that conducted in Canada [7], included two different choice-based tasks: 1. time trade-off (TTO, i.e., cardinal task), and 2. discrete-choice experiment (DCE, i.e., ordinal task). The TTO determines the duration of life in perfect or full health that a participant is willing to forgo to avoid living in a health state that is worse than perfect/full health (hereafter referred to as cardinal preferences). The DCE task asks respondents to choose their preferred health in a pairwise comparison involving two different health states (hereafter, a response that identifies a respondent's preferred health state in a pairwise comparison is referred to as an ordinal preference). Respondents' responses are aggregated to develop a value set that generates an index score, or HUS. If the sample of respondents in the valuation study is representative of the general population, the HUS then represents the average of the general population's preferences for a health state described by the generic instrument's descriptive system.

This study makes use of health preferences elicited from a representative sample of the Canadian population [7]. The primary objective of this study was to explore to what extent the HUS represents individuals' stated ordinal preferences. This will be based on an analysis of increments in HUS, which define transitions between two health states. The main study results focus on the TTO responses of individuals, with supplementary analysis of DCE responses.

4.3 Methods

This study uses HRQL index scores from the EQ-5D-5L, which is a popular generic indirect preference-based measure of HRQL [8,9]. The EQ-5D-5L uses a health descriptive system with 5 dimensions of health, each with 5 response options describing levels of impairment from 'no problems' (level 1) in the dimension to 'extreme problems' or 'unable to do' (level 5). In total the descriptive system defines 3,125 unique health states of which 86 health states (~3%) were evaluated by participants in the Canadian Valuation Study. Respondents' TTO responses were aggregated to develop a value set, which produces an index, or HUS, representing the population's preference for every health state out of a possible 3,125 health states defined by the EQ-5D-5L descriptive system. The details of the Canadian EQ-5D-5L Valuation Study have been published elsewhere [7].

4.3.1 Canadian EQ-5D-5L Valuation Study

In brief, multicentre quota sampling was used to recruit participants representative of the Canadian general population [7]. The typical valuation task for the EQ-5D-5L uses a time horizon of 10 years, where participants can forgo time in perfect health in 6-month increments until an indifference point is reached [112]. As per the EQ-5D-5L valuation protocol, each participant evaluated 10 out of a total of 86 unique health states selected, with a composite TTO (cTTO) task [7]. For the TTO task in the Canadian Valuation Study, two 'severe' health states were re-evaluated by each respondent using the traditional TTO (tTTO) task [7]. Ultimately, the HUS was based on all tTTO values and positive cTTO values, with negative and zero cTTO values censored at zero [7]. In the DCE task, participants were presented with two health state descriptions and asked to choose the 'better' health state (i.e., the one he/she preferred). A total of 196 pairwise comparisons were used, wherein each participant evaluated 7 pairs of health states with a total of 14 different EQ-5D-5L health state descriptions. Time lived in a health state was not considered by participants in the DCE task. The type and order of health states evaluated by each participant were chosen by block randomization [7]. In this study, a participant evaluated different EQ-5D-5L health states for the TTO and DCE tasks. The DCE responses were not, however, used in the development of the Canadian EQ-5D-5L value set [7].

4.3.2 Data Transformation

The HUS is based on TTO data on the [-1,1] interval, which involves a linear transformation of negative tTTO values (i.e., health states considered 'worse than dead') by dividing by 19 [7]. For the purpose of this analysis, these data were transformed into pairwise comparisons of health states based on differences in observed TTO values. A pairwise comparison is calculated as the TTO value for the comparator health state subtracting the TTO value of the reference health state. There are 90 possible pairwise comparisons per participant (i.e., 10 health states choose 2 to make pairwise comparisons, multiplied by two directions of change: better and worse); however, to accommodate any effect of order, a decision was made to only include 45 pairwise comparisons based on the order in which health states were evaluated. In this way, the first health state will have 9 pairwise comparisons as the reference state, while the last health state evaluated by each participant will have 0 pairwise comparisons as the reference state.

The pairwise comparison of two health states and their TTO values yields a difference that is positive, negative or equal to 0. A positive difference indicates that the participant considers the comparator health state as ‘better’ than or preferred to the reference health state (i.e., a transition to a better health state or improvement), while a negative difference indicates that the comparator health state is ‘worse’ than the reference health state (i.e., a transition to a worse health state or deterioration). A difference equal to 0 implies that the participant is indifferent between the reference and comparator health states (i.e., the health states are ‘(about) the same’), indicating a ‘tie’. We refer to this transformed TTO response indicating a tie/improvement/deterioration as an ordinal TTO response (oTTO).

Unlike the oTTO responses, in the DCE a participant must identify a preference for a health state (i.e., there is no response option for ‘(about) the same’ or ‘equal’) such that ties cannot occur. We did not consider any ordering effect on the DCE responses, and so included both directions in the analysis (i.e., health state A subtract health state B, and vice versa). Since there are no DCE comparisons that are identical to the TTO pairwise comparisons to allow for direct comparison, we applied the TTO-based Canadian valuation set to calculate the HUS difference between the reference and comparator health states. In effect, the HUS difference (ΔHUS) acted as a common explanatory variable to compare oTTO data to the DCE data. In addition, the HUS difference represents the average of the population's preference for the transition between two health states, thus the extent of agreement between an outcome variable and the explanatory variable is used to demonstrate differences between individuals' preferences and the HUS. Importantly, the following methods allow for the investigation of heterogeneity among individuals' preferences *on average* (i.e., at a mean parameter level), and as such do not necessarily reflect the uncertainty of any one individual's preferences (i.e., at an individual-level).

4.3.3 Analysis

The following describes analyses of increments in health utility, wherein a HUS increment represents a transition between two health states. All statistical models presented in the following used a random-intercept by respondent, which is consistent with the approach used in value set development [7]. The analyses compare differences in HUS (i.e., population's preference) to respondents' ordinal preferences. All analyses were preformed using R statistical software (version 3.5.1) [113] and the 'lme4' package [114].

Logistic Regression Models

Three logistic regression models were used to construct logistic regression curves representing the probability that participants consider a difference in EQ-5D-5L HUS between health states to represent a tie or transition to a worse/better health state (oTTO):

Equation (1) $P(\text{oTTO}=y \mid \Delta\text{HUS}=x) = (1 + e^{-(\beta_0+\beta_1 x+\beta_2 x^2)})^{-1}$

where y = tie, worse or better difference between health states A and B, and x = HUS of the comparator health state (e.g., B) substract the HUS of the reference health state (e.g., A). Average marginal probabilities were then predicted from each regression model, and scaled to sum to one. Since there are no ties in the DCE data, we also conducted analyses using oTTO responses with ties excluded.

Receiver Operating Characteristic Curve

Receiver operating characteristic (ROC) curve analysis was used to determine optimal thresholds for the HUS difference that discriminates a transition to a worse/better health state as indicated by respondents' oTTO responses. The HUS thresholds that maximize the sensitivity and specificity of discriminating transitions to worse/better health states were calculated, along with the area under curve (AUC) values.

Average Differences in TTO Values

Average differences in observed TTO values (ΔTTO, i.e., cardinal preferences) were stratified according to respondents' oTTO response and analyzed based on HUS differences (ΔHUS) using a linear random-effects model:

Equation (2) $E[\Delta\text{TTO} \mid (\text{oTTO}=y, \Delta\text{HUS}=x)]$.

To account for non-linearity we considered models up to and including 5th order effects by HUS difference. The effect of excluding ties from oTTO responses was considered.

For comparison, we also gave first-order models without stratification by oTTO. It was expected that there is a one-to-one relationship between average differences in observed TTO values and the HUS difference when ties were included.

Comparison to DCE Data

The oTTO responses were compared to DCE responses by predicting the probability that a HUS difference represents a transition to a worse/better health state. A linear model that included predicted probabilities from oTTO responses as the outcome variable regressed on predicted probabilities from DCE responses was used to determine the level of agreement:

Equation (3) $E[\ P(\text{oTTO=y} \mid \Delta\text{HUS=x}) \mid P(\text{DCE=y} \mid \Delta\text{HUS=x})\ ]$

$= 0+(\beta)\text{P}(\text{DCE}=y \mid \Delta\text{HUS}=x)$

The effect of excluding ties from oTTO responses on the slope coefficient (β) was investigated.

Probability that HUS Difference Represents Respondents' Ordinal Preferences

We analysed the probability of respondents' having the majority (i.e., >50%) of their own ordinal preferences represented by the HUS difference. We determined the HUS difference at which there was a 0.5 probability of respondents having their ordinal preferences represented (response *j* of individual *i*):

Equation (4) $\text{P}(\text{Q}_i(x)=1 \mid \Delta\text{HUS}=x) = (1 + e^{-(\beta_0+\beta_1 x+\beta_2 x^2)})^{-1}$

where $\text{Q}_i(x) = 1$ if $\sum_{j=1}^{j=r}((\text{z}_{i,j} \times \theta_{i,j}(x) / \sum_{j=1}^{j=r}\theta_{i,j}(x)) > 0.5$; 0 otherwise

$\theta_{i,j}(x) = 1$ if $|\Delta\text{HUS}| \leq x$; 0 otherwise

$z_{i,j} = 1$ if $\text{oTTO}_{i,j}$ or $\text{DCE}_{i,j}$ = better, and $\Delta\text{HUS}_{i,j}>0$

$z_{i,j} = 1$ if $\text{oTTO}_{i,j}$ or $\text{DCE}_{i,j}$ = worse, and $\Delta\text{HUS}_{i,j}<0$

$z_{i,j} = 0$ otherwise.

In summary, the binary outcome variable, $Q_i(x)$, is one for the *i*th respondent when more than 50% of its ordinal responses, $z_{i,j}$, are in agreement with the direction of the HUS difference (note, $\theta_{i,j}(x)$ is a dummy variable). Finding the value of x when Equation (4) equals 0.5 gave the HUS difference of interest. We analyzed oTTO responses (with ties excluded) and DCE responses separately, as well as combining the data sets together.

95% Bootstrap confidence intervals

Due to the nested structure of the data (i.e., multiple observations per respondent), bootstrapping involved sampling respondents with replacement as well as observations within each respondent [115]. The 95% bootstrap confidence intervals are based on percentiles from 1,000 bootstrap replicates [115].

4.4 Results

Data were from 1,073 participants from 4 Canadian cities (Vancouver, Edmonton, Hamilton and Montreal), after excluding 136 participants based on a priori criteria. Between the TTO and DCE tasks, there were six pairwise comparisons with the same HUS difference (+/-0.014, 0.2945, 0.349); however, the respondents differed for the two tasks. Among the 48,285 TTO pairwise comparisons, there were 440 unique absolute HUS differences that ranged from as small as 0.0003 (between states 11425 and 42115) to as large as 1.0776 (between states 11211 and 55555). For the 15,022 DCE responses, there were 207 unique absolute HUS differences ranging from as small as 0.0045 (between states 34333 and 33142) to as large as 0.6607 (between states 33111 and 32545). Figure 9a shows the cumulative number of pairwise TTO comparisons and DCE responses up to and including a specific HUS difference, while Figure 9b shows the cumulative number of respondents categorized by TTO and DCE response. Per respondent, the mean absolute HUS difference was 0.3748 on average (IQR: 0.1587 to 0.5499) for pairwise TTO comparisons, and 0.2000 on average (IQR: 0.0938 to 0.3022) for DCE responses (Table 5).

The oTTO response distribution consisted of 20.2% ties, 40.1% worse, and 39.8% better (Table 5). Ignoring the nested structure of the data, the mean HUS difference for TTO ties was 0.0034 (IQR: -0.2155 to 0.2281); however, the mean absolute HUS difference for ties was 0.2562 (IQR: 0.0885 to 0.3783). Per respondent, on average, 20.2% (IQR: 8.9% to 26.7%) of the 45 pairwise comparisons were ties, wherein ties per respondent had a mean HUS difference of -0.0009 on average (IQR: -0.1650 to 0.1615), and a mean absolute HUS difference of 0.2386 on average (IQR: 0.1288 to 0.3350).

4.4.1 Logistic Regression Models

Probability curves show that HUS differences slightly above 0 (or slightly below 0) have as little as 34.8% (95% CI: 34.1-35.5%) probability of representing a transition to a better health state (or a 34.8% probability to a worse health state), and as much as 30.6% (95% CI: 29.1-31.9%) probability of representing a tie (Figure 10). When ties were excluded from the oTTO responses, a near 0 HUS difference was equally likely to represent a transition to a worse or better health state (i.e., probability of 50%) (Figure 10).

4.4.2 Receiver Operating Characteristic Curve

EQ-5D-5L HUS differences of -0.054 (95% CI: -0.071 to -0.029) and 0.047 (95% CI: 0.026 to 0.076) maximized the sensitivity (0.79, 95% CI: 0.77 to 0.81; 0.80, 95% CI: 0.77 to 0.82) and specificity (0.79, 95% CI: 0.77 to 0.81; 0.78, 95% CI: 0.76 to 0.80) of discriminating transitions to a worse and better health state, respectively (Figure 10) with AUC values of 0.87 (95% CI: 0.86-0.88). However, when ties were excluded, the ROC curves identified near 0 (0.003) differences in HUS maximizing the sensitivity (0.84) and specificity (0.84) of discriminating transitions (to worse and better health states, respectively) with AUC values of 0.92.

4.4.3 Average Differences in TTO Values

When stratified according to oTTO responses, differences in EQ-5D-5L scores of +/-0.03 to +/-0.07 were associated with average TTO differences of approximately -/+ 0.17 (95% CI: 0.16-0.19), -/+ 0.35 (95% CI: 0.33-0.37) whether ties were included or excluded, respectively (Figure 11a). The average TTO difference displayed near perfect agreement with the HUS difference when ties were excluded (slope coefficient of 0.99, 95% CI: 0.96-1.02), whereas a shallower slope was observed when ties were included (slope coefficient of 0.90, 95% CI: 0.87-0.93).

4.4.4 Comparison to DCE Data

When ties were included in the oTTO responses, the slope coefficient of predicted probabilities were 0.81 (95% CI: 0.79-0.82); however, if ties were excluded, the slope coefficient increased to 0.99 (95% CI: 0.97-1.00) (Figure 12).

4.4.5 Probability that HUS Difference Represents Respondents' Ordinal Preferences

The probability of respondents' majority ordinal preferences being represented by the HUS increased for increasing HUS differences (Figure 13). However, for HUS difference near 0, the probability was less than 0.5, increasing to 0.5 at HUS differences of 0.056 (95% CI: 0.014-0.084), 0.076 (95% CI: 0.062-0.087), and 0.054 (95% CI: 0.043-0.064) for DCE, oTTO (with ties excluded) and combined responses, respectively.

4.5 Discussion

This analysis suggests that small differences in HUS may misrepresent individuals' ordinal preferences for health states. By taking the average of participants' preferences to develop a value set, small differences in HUS may mask the underlying heterogeneity and uncertainty in individuals' preferences. The value of measuring health utilities cannot be overstated; as such these results should not be interpreted to suggest that there is too much uncertainty in the measurement of health utilities. However, it is important to recognize the limitations of HUS, and to develop methods that appropriately address the shortcomings of any measure. In this regard, when small differences in HUS are observed, researchers should not conclude that there is uniformity in ordinal preferences and/or differences in cardinal preferences that are small in magnitude. Instead, this analysis suggests that small HUS differences are not rigorously interpretable from the individuals' perspective, and, in consequence, may lack robust meaning.

Thus, in the face of preference heterogeneity, a HUS difference may have to be 'large enough' to be expected to reflect what individuals consider to be an improvement or deterioration in HRQL. In this analysis, HUS differences near 0.05 appear as a type of threshold, wherein HUS differences below 0.05 do not maximize the sensitivity/specificity of truly reflecting a transition to a better/worse health state, and have less than 50% of reflecting within-individual ordinal preferences. This value of 0.05 is similar to estimates of minimally important differences for the EQ-5D-5L (see [103,116] for estimates based on the Canadian value set), which may lend support to the application of a minimally detectable or important difference in the interpretation of small HUS differences.

The extent of the heterogeneity in ordinal preferences is not consistent across the spectrum of HUS differences. When the transition between health states is large in magnitude, as represented by a large HUS difference (e.g., transitions from near perfect health to a health state equivalent to being dead, i.e., ~1 to ~0 in HUS), there is greater consensus among individuals in the transition's meaning, which results in less heterogeneity in ordinal preferences (i.e., it is almost certainly a transition to a worse or better health state). But, as the HUS difference shrinks, there is less uniformity among individuals in terms of whether the difference represents a transition to a worse or better health state, or if the two health states are 'about the same' (i.e., a 'tie' in TTO pairwise comparison). In this regard, quantifying the smallest HUS difference that is expected to represent HRQL improvement or deterioration according to individuals' preferences is potentially a useful way to promote the relevance of HUS at various levels of decision-making (including at the individual-level with caveats).

Of course, it is also important to consider by how much individuals' cardinal preferences depart from the HUS (i.e., the magnitude of difference). Even though there is greater heterogeneity in ordinal preferences for near 0 differences in HUS, it could be that the magnitude by which cardinal preferences depart from the HUS is so small that it may be considered inconsequential. However, our analyses suggests otherwise: individuals who differ from the direction of the HUS do so, on average, by a considerable amount. For example, a HUS difference of +/-0.03 to +/-0.07 is considered by those who perceive the difference as a transition in the opposite direction as a -/+0.17 (or -/+0.35 difference when ties are excluded) in TTO score (on average), which, on a 10-year time-horizon, is equivalent to trading-off more than 20 months in perfect health. Furthermore, while it was expected that there is a one-to-one relationship between the average observed TTO difference and HUS difference when ignoring oTTO stratification, it was surprising to find that this only occurred when ties were excluded. In contrast, including ties in TTO differences resulted in a shallower slope, indicating that respondents were, on average, less willing to trade-off time in perfect health for a unit HUS difference.

Finally, while the aforementioned results represent inter-individual heterogeneity, when analysing disparities between population and individual-level preferences, there is also intra-individual heterogeneity. For example, a respondent may consider a difference of +0.01 HUS to be an improvement, but a -0.03 HUS difference as an improvement as well. In this case, the respondent’s ordinal preferences are not well represented by the (population-based) HUS difference. Again, this intra-individual disparity is most likely to occur for small HUS differences; therefore there is likely a point at which respondents’ ordinal preferences are (mostly) represented by the HUS difference. In this analysis, we considered any one individual’s ordinal preferences to be represented by the HUS difference when greater than 50% (i.e., the majority) of their own ordinal preferences were in agreement. We then calculated the HUS difference at which the probability of respondents having their own ordinal preferences (mostly) represented reached 50%. This occurred at a HUS difference in the range of 0.043 to 0.064 (based on combined data set with DCE responses and ties excluded from oTTO), which is similar to the value obtained from the ROC analysis of oTTO data (with ties included).

Previous research has identified the gap between individual preferences (i.e., doing what is 'best' for the patient), and allocation decisions based on the average of the population's preferences [51,59,69,107]. Importantly, differences between the average of the population's preferences and the cardinal preferences of patients with specific characteristics have been documented [117–120]. This has stimulated further study such as the analysis of preference heterogeneity, the value of individualized care, the development of individual value sets, and decentralized decision-making [59,69,107,121]. However, to our knowledge, this is the first study to explicitly demonstrate the limitations of HUS (based on the average of the population's preferences) in terms of how well incremental HUS differences represent respondents' stated ordinal preferences.

The challenges and limitations of health preference elicitation tasks have been well documented [43,44]. Previous studies have shown that the TTO task produces more 'ties' than DCE or visual analogue tasks [122,123]. Indeed, the DCE task used by the EQ-5D Canadian valuation study did not include a response option for '(about) the same' [7]. This differs from the PAPRIKA (potentially all pairwise rankings of all pairwise alternatives) method that includes the option 'they are equal' when a respondent is asked to make a choice between two EQ-5D health states [121]. However, results from our study showed that there is good agreement between DCE responses and oTTO data when ties are excluded. Furthermore, Purba et al. (2018) analysed the test-retest reliability of TTO and DCE in the EQ-5D-5L, finding evidence to suggest that TTO is more reliable [124]. In addition, Purba et al. (2018) reported an average mean absolute difference of 0.079 and an overall mean increase of 0.042 for TTO values from first test to re-test [124]. Taylor et al. (2017) also compared increments in health utility to test the intra-person interval property of the EQ-5D HUS [125]. This study was similar in that it also measured the agreement between HUS differences and individuals' preferences, finding that individuals may have different preferences for equivalent gains in health utility increments depending on the severity of the baseline health state [125]. However, the smallest HUS difference evaluated by Taylor et al. (2017) was 0.25 based on the EQ-5D-5L United Kingdom official cross-walk [125].

The development of EQ-5D value sets has seen greater use of 'hybrid' (i.e., TTO and DCE) data as well as protocols that include 'feedback' to respondents on the rank order of evaluated health states [112]. It is possible that different value sets may have different levels of preference heterogeneity due to methodological aspects of the study design and value set development, or cultural aspects of the population. Future studies might consider replicating this analysis for other EQ-5D valuation studies, such as those that use 3-month smallest intervals in TTO values, comparing the 3L version to the 5L version, as well as the inclusion of DCE responses in value set development. In addition, this methodology may also be useful in the analysis of ordinal responses from patient-based HUS.

This study has a number of limitations that warrant consideration. Firstly, the TTO task is a cardinal task, converting the values into ordinal information may result in a loss of data integrity. Most importantly, the ‘ties’ that result may not reflect a ‘true’ indifference between health states since the smallest allowable interval is 6-months (i.e., +/- 0.05 on the 10-year time horizon). Secondly, this study represents a secondary use of data. This provides advantages in terms of the strength and standardization of the data collection protocol; however, the Valuation Study was not explicitly designed to investigate preference heterogeneity in small HUS differences. Furthermore, the fact that the HUS is based on the same TTO data may result in ‘over-fitting’ of estimates obtained from the current analysis. However, DCE responses were not used in developing the HUS, and results appear to be consistent with oTTO. In addition, a non-parametric multi-level bootstrapping approach with resampling at the individual and response levels was used to compute 95% confidence intervals for estimates. It is debatable whether it is more appropriate to resample from only a single-level (i.e., individuals or responses) [115]. We investigated other resampling approaches (unpublished), and only noticed an impact on the range of estimates from analysis of DCE responses. This may be due to the fact that there are only 14 (or 7 unique) DCE responses per individual (compared to 45 oTTO responses). We therefore opted to be conservative, choosing the approach that produced the largest 95% confidence interval (i.e., most uncertainty). Future studies may be improved through purposeful sampling of health states with small HUS differences.

4.6 Conclusion

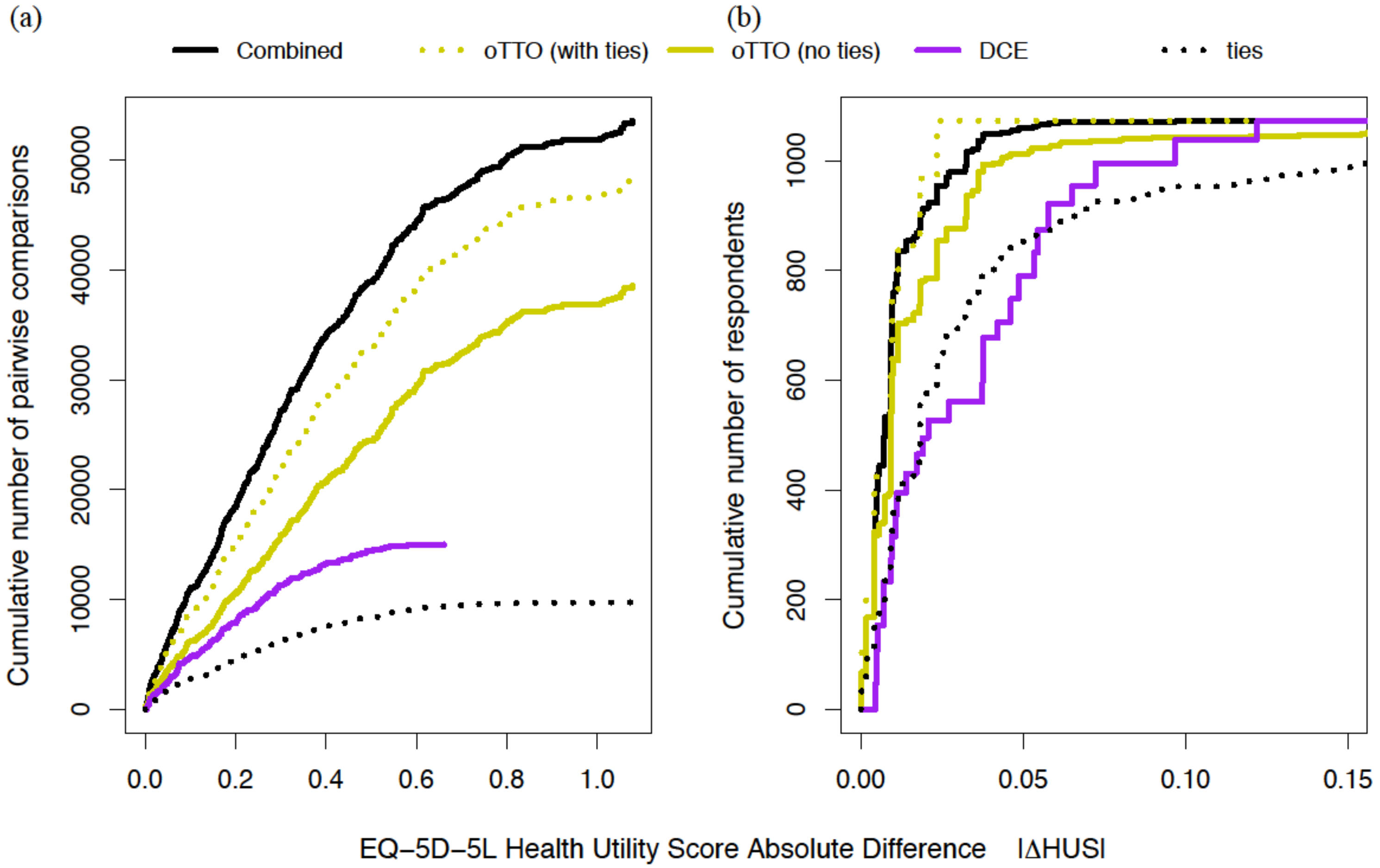
Differences in population-based EQ-5D-5L health utility scores, particularly small differences (<0.05), are not uniformly representative of individuals' stated ordinal preferences. This suggests that small differences in HUS are ambiguous, lending support to minimally important difference methods for the identification of HUS differences that are large enough to be considered meaningful (e.g., ≥ +/-0.05).

Table 5. Descriptive statistics of EQ-5D-5L health utility score for time trade-off and discrete-choice experiment pairwise comparisons.

Description	**TTO pairwise comparison**						**DCE pairwise comparison**	
	Per Respondent			Total			Per Respondent	Total
	All	Ties	No Ties	All	Ties	No Ties		
Number of pairwise comparisons (%)	45	9.1 (IQR: 8.9-26.7%)	35.9 (IQR: 73.3-91.1%)	48 285	9 736 (20.2%)	38 549 (worse, better: 39.8%, 40.1%)	14*	15 022
Unique absolute HUS difference comparisons	45	9.1	35.9	440	437	440	7*	207
Mean HUS difference (SD)	-0.0006 (0.4426)	-0.0009 (0.2763)	-0.0025 (0.4686)	-0.0006 (0.4601)	0.0034 (0.3245)	-0.0017 (0.4884)	0 (0.2546)	0 (0.2542)
Median HUS difference (IQR)	0.0004 (-0.3121, 0.3131)	-0.0004 (-0.1650, 0.1615)	-0.0026 (-0.3392, 0.3348)	0.0017 (-0.3388, 0.3376)	0.0017 (-0.2155, 0.2281)	0.0017 (-0.3699, 0.3669)	0 (-0.1841, 0.1841)	0 (-0.1851, 0.1851)
Range of HUS difference	-0.9293, 0.9268	-0.3773, 0.3764	-0.9243, 0.9202	-1.0776, 1.0776	-1.0776, 1.0776	-1.0776, 1.0776	-0.4192, 0.4192	-0.6607, 0.6607
Mean absolute HUS difference (SD)	0.3748 (0.2648)	0.2386 (0.1731)	0.4102 (0.2686)	0.3748 (0.2668)	0.2562 (0.1992)	0.4048 (0.2733)	0.2061 (0.1359)	0.2061 (0.1488)
Median absolute HUS difference (IQR)	0.3257 (0.1587, 0.5499)	0.2231 (0.1288, 0.3350)	0.3784 (0.1996, 0.5829)	0.3376 (0.1591, 0.5482)	0.2189 (0.0885, 0.3783)	0.3696 (0.1754, 0.5949)	0.2000 (0.0938, 0.3022)	0.1851 (0.0735, 0.2968)
Range of absolute HUS difference	0.0093, 1.0600	0.0411, 0.4924	0.0210, 1.0587	0.0003, 1.0776	0.0003, 1.0776	0.0003, 1.0776	0.034, 0.4192	0.0045, 0.6607

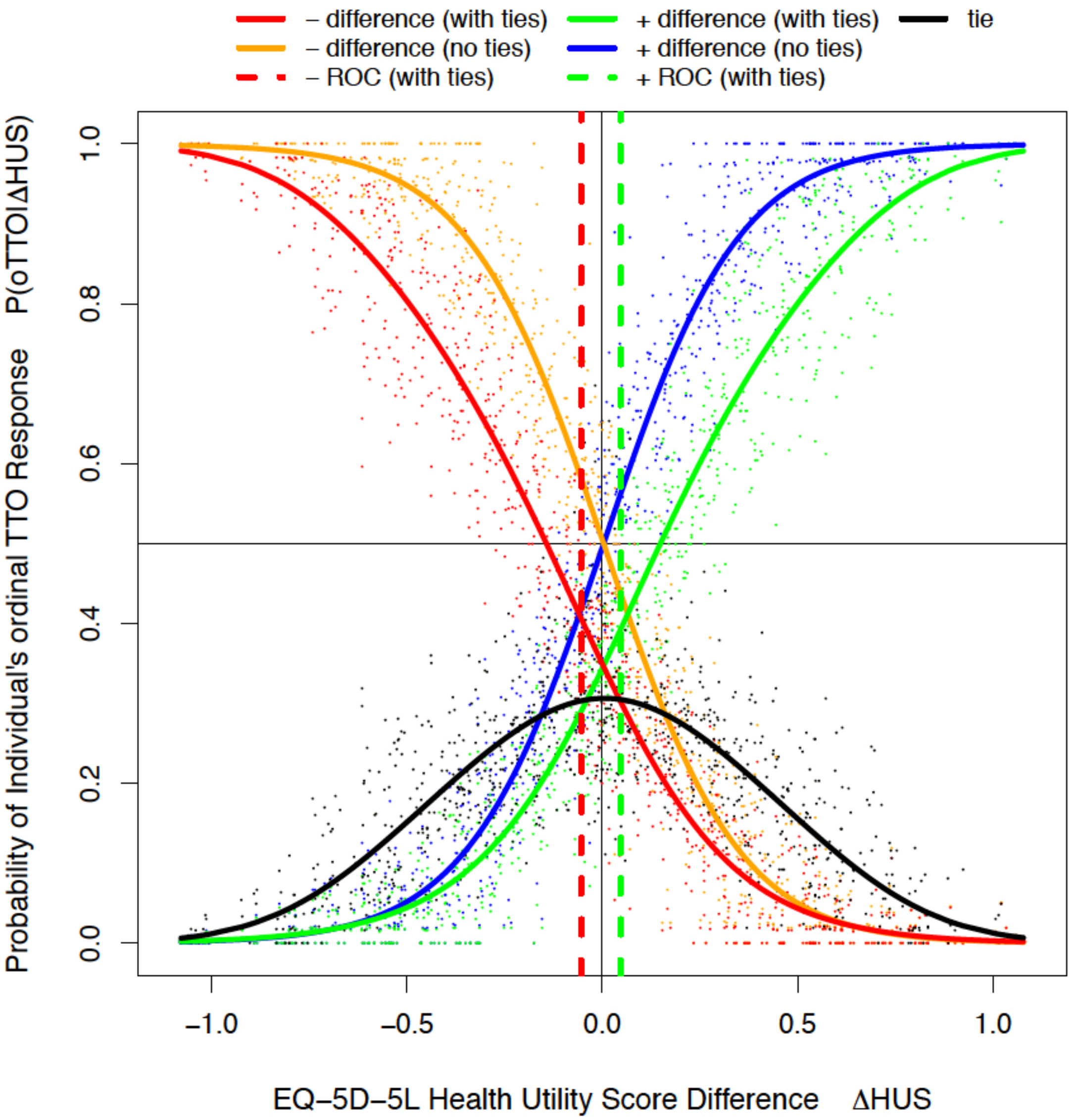
*The number of DCE tasks per respondent was originally 7, however both directions were considered (i.e., health state A subtract health state B, and vice versa) doubling the number of pairwise comparisons. Per respondent statistics are reported as averages of statistics. TTO, time trade-off; DCE, discrete-choice experiment; Ties, pairwise comparison with no difference in time trade-off value; HUS, health utility score.

Figure 9. The cumulative number of pairwise comparisons (a), and respondents (b), for ordinal time trade-off (yellow), discrete-choice experiment (purple), combined responses (black-solid), and ties (black-dashed), up to and including an EQ-5D-5L health utility score absolute difference.



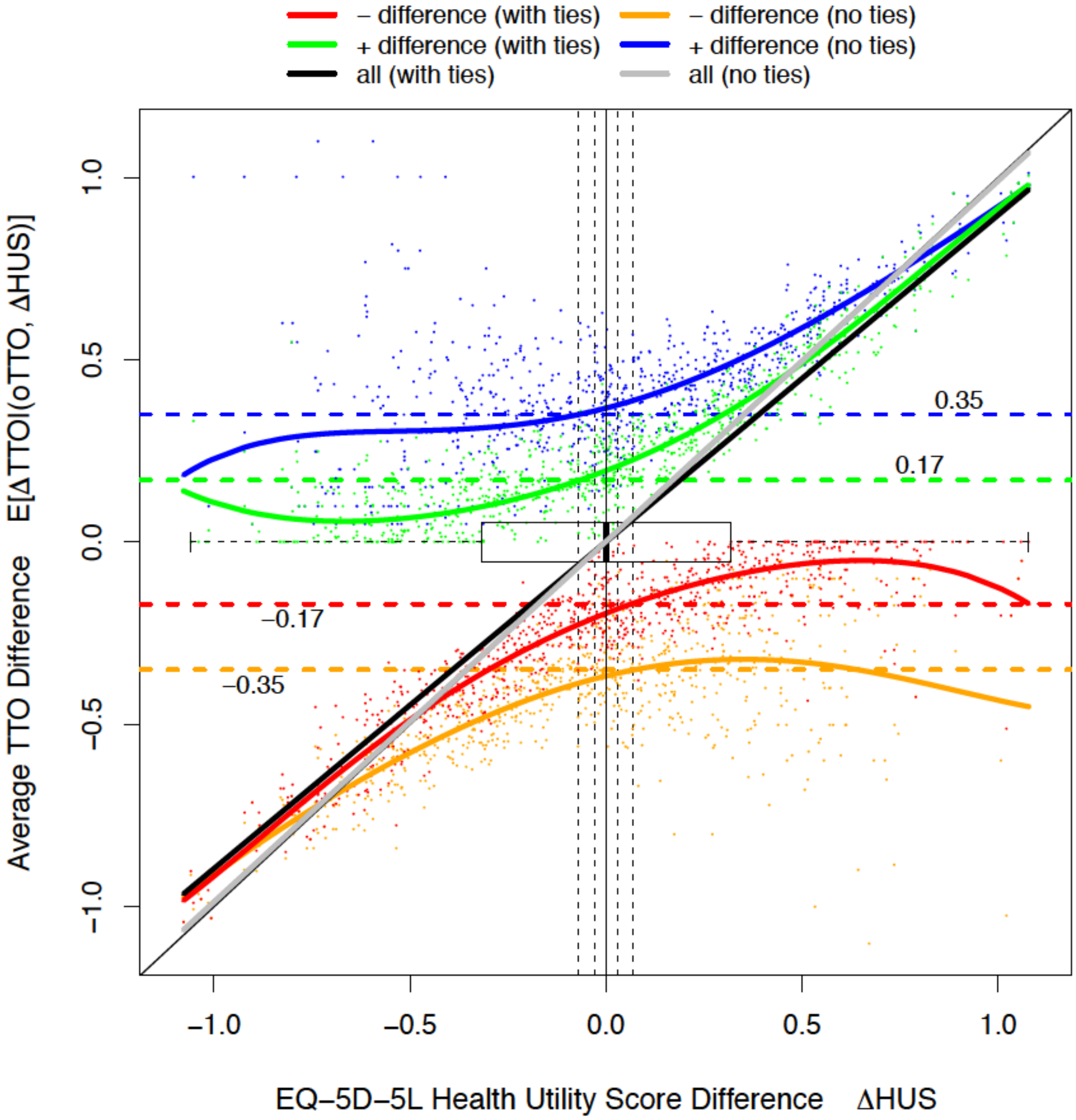
oTTO, ordinal time trade-off; DCE, discrete-choice experiment; ties, pairwise comparison with no difference in time trade-off value; HUS, health utility score.

Figure 10. Probability that a difference in EQ-5D-5L health utility score represents a tie, worse, or better difference between health states.



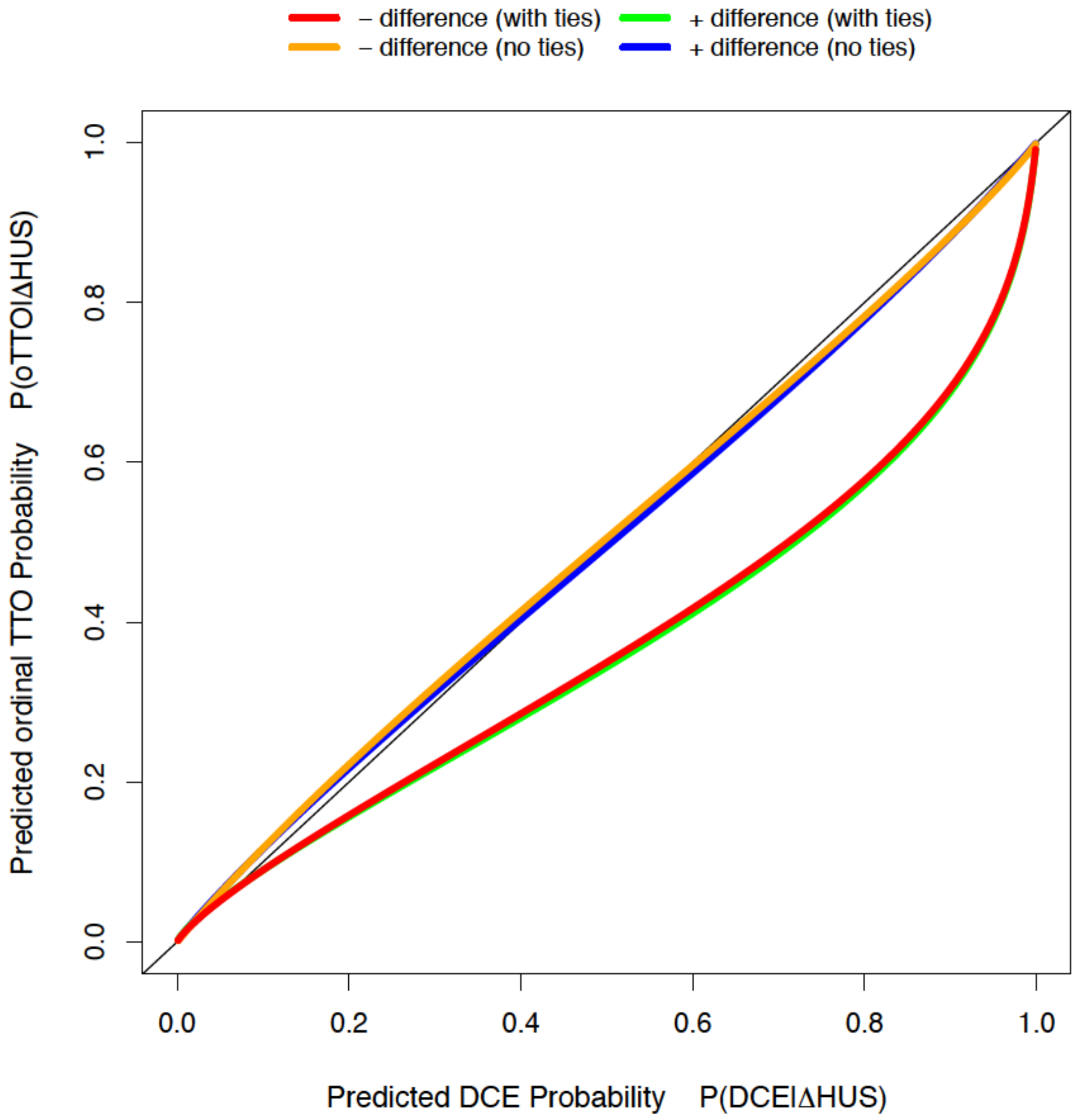
Solid lines are predicted average marginal probability from random-intercept logistic regression model; dashed lines from ROC analysis maximizing sensitivity/specificity; dots represent proportion of oTTO responses; ROC, receiver operating characteristic; oTTO, ordinal time trade-off; –/+, transition to worse/better health state; ties, pairwise comparison with no difference in time trade-off value; HUS, health utility score.

Figure 11. The average difference in observed time trade-off values based on ordinal time trade-off response and difference in EQ-5D-5L health utility scores.



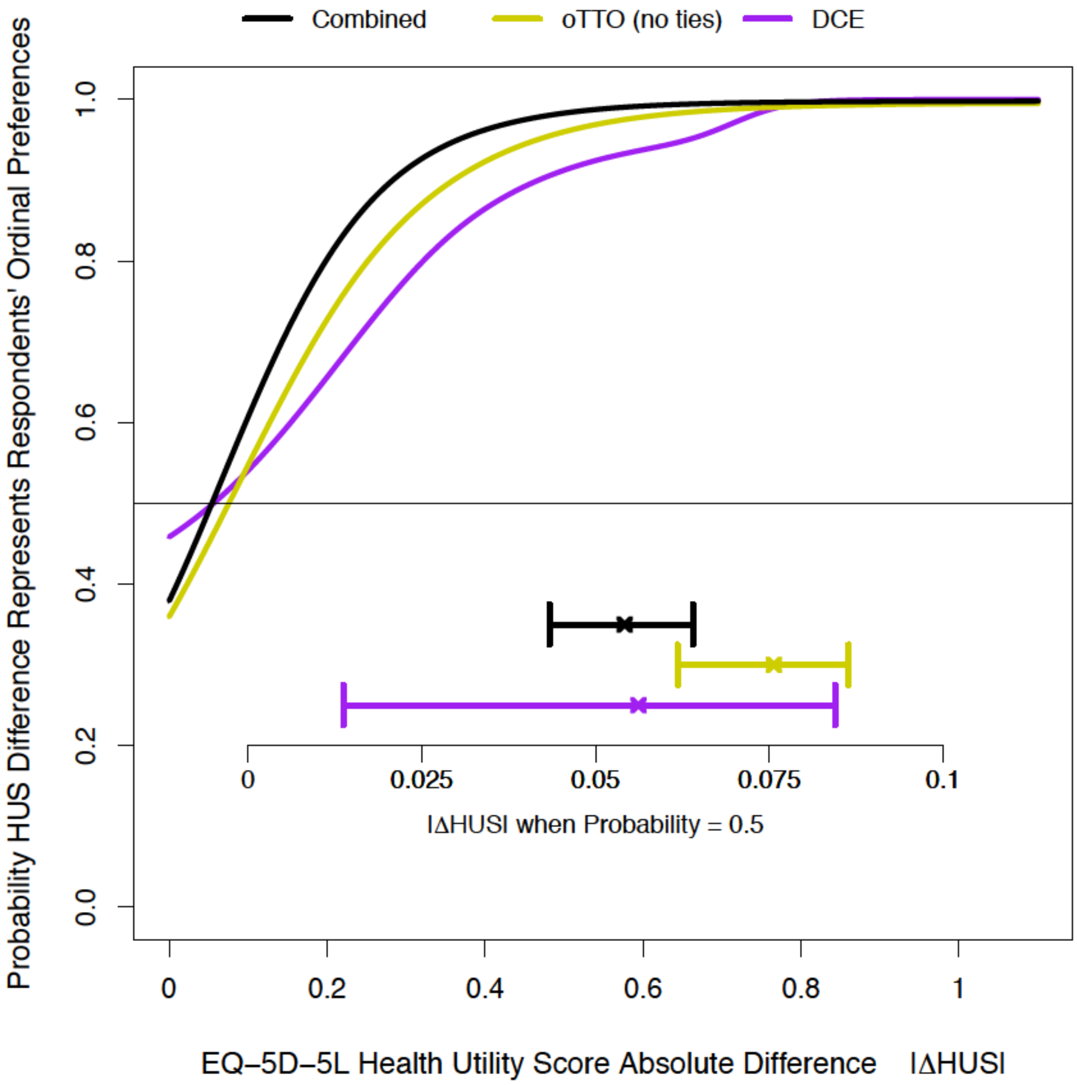
Solid lines from random-intercept linear regression model; dots represent TTO differences averaged by health utility score difference; oTTO, ordinal time trade-off; –/+, transition to worse/better health state; ties, pairwise comparison with no difference in time trade-off value; HUS, health utility score.

Figure 12. Probabilities of ordinal time trade-off responses compared to discrete-choice experiment responses predicted by differences in EQ-5D-5L health utility scores.



Solid lines are predicted average marginal probability from random-intercept logistic regression model; oTTO, ordinal time trade-off; DCE, discrete-choice experiment; –/+, transition to worse/better health state; ties, pairwise comparison with no difference in time trade-off value; HUS, health utility score.

Figure 13. The probability that the absolute difference in EQ-5D-5L health utility score represents the majority ordinal preferences of respondents for ordinal time trade-off with ties excluded (yellow), discrete-choice experiment (purple), and combined responses (black).



Solid lines are predicted average marginal probability from random-intercept logistic regression model; oTTO, ordinal time trade-off; DCE, discrete-choice experiment; ties, pairwise comparison with no difference in time trade-off value; HUS, health utility score.

5. Modifying the Quality-Adjusted Life Year Calculation to Account for Meaningful Change in Health-Related Quality of Life: Insights from a Pragmatic Clinical Trial

Authors: Nathan S. McClure, Mike Paulden, Arto Ohimnaa, Jeffrey A. Johnson

5.1 Abstract

Background: Health-related quality of life (HRQL) scores are used to calculate quality-adjusted life years (QALYs), a period measure of HRQL gains/losses that is subsequently used in cost-utility analyses. However, end-users of generic HRQL instruments may seek non-economic assessments of HRQL changes over time. To this end, this study proposes a modified QALY calculation that may be useful to end-users, particularly in non-economic applications. This QALY calculation incorporates the minimally important difference (MID) in generic HRQL change scores to reflect meaningful HRQL improvement or deterioration from the patients' perspective. In doing so, this study also reviews common issues in QALY calculations such as adjustment for baseline scores and standardizing for between-group differences. **Methods:** Using EQ-5D-5L outcome data from the Alberta TEAMCare-Primary Care Network trial in the treatment of depression for patients with type 2 diabetes, this study compared results from different QALY calculation methods to investigate the impact of (i) adjusting for baseline HRQL score, (ii) standardizing between-group differences in baseline HRQL scores, and (iii) adjusting for 'meaningful' HRQL changes within-patients. The following QALY calculation methods are examined: area under curve (QALY-AUC), change from baseline (QALY-CFB), regression modeling (QALY-R), and incorporating a minimally important change from baseline (QALY-MID). **Results:** The average between-group differences in HRQL scores at baseline was 0.06 (p-value>0.05) with larger scores observed in the Collaborative Care group (n = 55) compared to the Enhanced Care group (n=43). The incremental QALY-AUC estimate favoured the Collaborative Care group (0.031) while the incremental QALY-CFB (-0.028) estimate favoured Enhanced Care. Adjusting for meaningful HRQL changes resulted in a crude incremental QALY-MID of -0.023; however, after adjusting for between-group differences in baseline scores, sex and age, QALY-R and adjusted incremental QALY-MID estimates were -0.007 and -0.001, respectively. In addition, results from recursive regression analyses showed that incremental QALY estimates are impacted by very low baseline HRQL scores, and failing to adjust for baseline scores may lead to incorrect conclusions in treatment benefits. **Conclusions:** Insights from this study can support end-users of generic HRQL instruments in the assessment of QALY outcomes for non-economic applications.

5.2 Introduction

The adoption of generic health-related quality of life (HRQL) measures in observational and clinical studies suggests end-users recognize the value in outcomes that capture people's preferences for different health states. HRQL scores from patients' self-reported health status can be combined with duration (i.e., time lived in the health state) to calculate quality-adjusted life years (QALYs), a period measure of HRQL gains/losses over time [38,55]. End-users of generic HRQL instruments may wish to compare QALYs between groups to inform care and resource allocation decisions. However, assessing trade-offs in HRQL has many challenges, which may include adjustment for baseline scores, standardizing differences between groups, accounting for 'measurement error', and interpreting HRQL changes based on community preferences from the patients' perspective [60,67,85].

The use of QALYs in the denominator of cost-utility analyses was championed by health-economists, and has been the impetus for creating various generic HRQL instruments [12,19]. As stated by Kind et al., "the QALY was developed to inform top-down decision-making process through the medium of cost-utility analyses" [20]. However, the popularization of generic HRQL measures may be the result of a broader appreciation by end-users for the role of a common end-point to support decision-making at multiple-levels. Despite a well-established history and research community in health economic applications, there is less research and evidence to support the assessment of QALY outcomes (from patient-level data) in non-economic applications [20]. Arguably, such assessments are needed if QALYs, and generic HRQL measures alike, are to have relevance to the broader community of end-users including patients and clinicians [20].

Indeed, failing to bridge this gap may lead to unintended divisions between economists, medicine, and patients [58,126]. To this end, the main objective of this study is to propose a new method of calculating QALYs that incorporates the minimally important difference (MID) in HRQL change scores. The MID represents the smallest change in HRQL score that is expected to be minimally important to patients at the point of application, thus it is considered a method to support the interpretability of HRQL scores [31,33,116]. This study extends the concept of the MID to generate QALY results that are relevant at the point of application.

This study presents an analysis of HRQL outcomes from the Alberta TEAMCare-Primary Care Network trial in the treatment of depression for patients with type 2 diabetes [127]. The overall objective of the trial was to determine the best mode of depression care by comparing changes between two groups: enhanced (usual) care, and collaborative care [127]. To evaluate HRQL, the trial measured HRQL using the EQ-5D-5L instrument to produce index scores that were subsequently used to calculate QALYs [128,129]. The current study has the following objectives: (i.) Examine current QALY calculation methods and their rationale; (ii.) Propose a new method adjusting for 'meaningful' HRQL score changes from the patients' perspective; (iii.) Investigate differences in QALY gains/losses between groups in the trial; and (iv.) Examine incremental QALY estimates across the range of baseline scores observed in the trial.

Type 2 diabetes is a chronic condition that may impact many aspects of health, including an increased risk for adverse health events such as cardiovascular diseases, foot disease, and depression [127]. Depression can further exacerbate the effects of type 2 diabetes through poor management and self-efficacy [127]. In itself, depression is recognized as an important aspect of HRQL [45]. To this end, one of the health dimensions of the EQ-5D health classification system is anxiety/depression [112].

Trial results have been previously published with evidence to suggest that patients in the Collaborative Care group experienced larger reductions in depressive symptoms [128,129]. In addition, a cost-utility analysis suggested incremental QALY benefits in favour of Collaborative Care compared to Enhanced Care [129]. However, a difference between groups in baseline EQ-5D-5L index scores was evident [128]. This difference was not adjusted for in the QALY calculation, therefore the result may not reflect 'true' changes in HRQL that are due to the type of care patients received (i.e., the difference in HRQL that is 'expected' from switching care). In fact, assuming there is no treatment effect, incremental QALYs that do not adjust for baseline index score will favour the group that starts at a higher baseline score [67]. Therefore, the purpose of this study is to re-analyze the incremental QALY gains/losses using methods to adjust for baseline index scores as well as meaningful within-patient change in HRQL.

5.3 Methods

The trial rationale and design have been previously published in detail [6]. Briefly, patients were invited to participate in the trial if they screened 'positive' for depressive symptoms based on the PHQ-9 at baseline (i.e., score ≥ 10) [127]. Eligible and consenting participants were then allocated to treatment arms by an 'on-off' group assignment method based on the month that the patient scheduled their baseline assessment [127]. Evidence suggests that this pragmatic assignment method, while non-random, reliably results in balanced patient groups [127]. All participants completed the

EQ-5D-5L questionnaire at baseline, six, and twelve months. Responses were then scored using the Canadian scoring algorithm to yield an index score [7]. Index scores were used to calculate QALYs based on the following methods:

5.3.1 QALY Calculation

Conventionally, the QALY is calculated as the area under the curve (QALY-AUC) for the combined assessments of time and HRQL [3]. As shown in Equation (5), the QALY-AUC over time-period *T* represents the equivalent length of time lived in perfect health. Based on multiple HRQL measurements at discrete time-points, $Y_i(t)$, this is expressed as:

Equation (5) $$\text{QALY-AUC}_i(T) = \sum_{t=0}^{t+d=T} (Y_i(t) + Y_i(t+d))d/2$$

where *d* is the increment between measurements, and where the *i*th patient has a baseline HRQL score of $Y_i(0)$. The incremental QALY-AUC is then calculated as the difference in QALY-AUC group-level averages.

Two methods have previously been proposed to adjust for baseline HRQL scores including a change from baseline approach (QALY-CFB) and a regression model (QALY-R) [67]. The QALY-R equation can be expressed as:

Equation (6) $$\text{QALY-R}_i(T) = \text{E}[\text{QALY-AUC}_i(T) \mid G_i, Y_i(0)] = \beta_0 + \beta_1 G_i + \beta_2 Y_i(0)$$

where the *i*th patient has a baseline HRQL score of $Y_i(0)$ and a group assignment G_i. The model coefficients β_0 and β_2 respectively represent the intercept and QALY gain per unit increase in baseline HRQL score. Of interest is the β_1 coefficient, which represents the 'treatment-effect', i.e., incremental QALY-R estimate adjusted for baseline HRQL scores.

In this study ($d = 0.5$; $t = 0, 0.5, 1$ years), the QALY-CFB can be expressed as:

Equation (7) $\text{QALY-CFB}_i(T) = (Y_i(0.5) - Y_i(0)) / 2 + (Y_i(1) - Y_i(0)) / 4$

where an overall weighted summary change score (replacing $Y_i(t) - Y_i(0)$) of

Equation (8) $(2(Y_i(0.5) - Y_i(0)) + Y_i(1) - Y_i(0)) / 3$

satisfies Equation 7. As before, the incremental QALY-CFB is calculated as the difference in group-level averages.

The QALY-R method predicts expected QALYs from a linear model for every patient based on their baseline index score. Furthermore, expected QALYs can be calculated for unobserved baseline scores using a regression model. In this way, comparisons between groups are standardized so that both groups have identical baseline scores. In this regard, the QALY-R and incremental QALY-CFB estimates may be considered 'adjusted' and 'crude' incremental QALY estimates, respectively [130].

An implied assumption of regression is that model coefficients are constant over the explanatory variable's range [131]. Recursive regression was applied to investigate the stability of incremental QALY estimates across the range of observed baseline index scores [131]. This involved plotting the estimates of linear regression models (e.g., β_1) from subsets of data that included ever-smaller (i.e., decreasing) baseline index scores.

5.3.2 QALY-MID

In addition to the QALY-R and QALY-CFB methods, we propose a new method for calculating QALYs that considers meaningful within-patient HRQL change from baseline (QALY-MID). This method applies 'responder-criteria' for within-patient HRQL change scores to discriminate expected improvement and deterioration in HRQL, while small score changes that fall within the range are assumed to have no value (as the direction of change is not meaningfully representative of the change experienced by patients). In this regard, the QALY-MID generates QALY estimates that are relevant to end-users assessing HRQL changes at the point of application.

In the current study, responder-criteria were informed from a previous observational study estimating the smallest change in EQ-5D index score that patients with type 2 diabetes consider minimally important [116]. Specifically, instrument-defined MID values were generated for observed baseline scores [116]. The responder-criteria were applied to the overall weighted summary change score shown in Equation (8). The difference in group-level QALY-MID estimates were used to calculate a *crude* incremental QALY-MID, while a regression model adjusting for differences in baseline scores, sex and age, is also used to yield an *adjusted* incremental QALY-MID estimate.

We used a complete-case analysis approach. Incremental differences represent the mean difference between groups. Regression methods use ordinary least squares fit. All analyses were conducted in R version 3.6.3 [113].

5.4 Results

In total, 43 patients in the Enhanced group and 55 patients in the Collaborative Care group completed the study. At baseline, the mean index scores were 0.66 and 0.72 for the Enhanced and Collaborative Care groups respectively, wherein the 0.06 difference between groups was not statistically significant (p-value = 0.16 from 2-sided t-test). The age of participants were similar (59 versus 58 years, p-value = 0.46); while Enhanced Care had fewer females (47%) than Collaborative Care (64%), the difference was not statistically significant (p-value = 0.14 from chi-squared test).

5.4.1 Distribution of Baseline EQ-5D Scores

Figure 14 shows the left-skewed distributions of baseline EQ-5D scores for the Enhanced Care and Collaborative Care groups. Higher baseline EQ-5D scores were more common in the Collaborative Care group compared to the Enhanced Care group. Moreover, the box and whisker plots suggest that low observed baseline index scores (<0.5) are outliers, particularly the minimum value observed in the Collaborative Care group (Figure 14).

5.4.2 Distribution of QALY Outcomes

The QALY-AUC distributions show how the Collaborative Care group is favoured, while the Enhanced Care group is favoured for incremental QALY-CFB or (crude) QALY-MID estimates (Figure 15). The empirical complementary cumulative distributions of the weighted summary score change illustrate that while larger changes are more common in the Enhanced Care group, a greater proportion of changes fall within a range that is not meaningfully representative of improvement or deterioration in HRQL (Figure 16).

5.4.3 Association between Change in HRQL and Baseline EQ-5D Score

A large negative correlation (-0.63) between weighted summary score change and baseline index score suggests that lower baseline scores are associated with larger gains in HRQL. Figure 17 plots the weighted summary score change as a function of baseline score for the different treatment groups. This suggests that deterioration in HRQL is more common among patients with high baseline scores, while improvement in HRQL is more common among lower baseline scores.

5.4.4 Incremental QALY Estimates

The mean incremental QALY estimates (standard error) for the QALY-AUC, QALY-CFB and crude QALY-MID methods were 0.031 (0.033), -0.028 (0.022) and -0.023 (0.022) respectively. After adjusting for differences in baseline index scores, age and sex, the QALY-R and incremental QALY-MID estimates are -0.007 (0.018) and -0.001 (0.018) respectively.

5.4.5 Stability of Incremental QALY Estimates over the Range of Baseline EQ-5D Scores

The recursive regression plot shows how including an increasing range of baseline index scores results in coefficient stability (Figure 18a). However, with the exception of the QALY-AUC method, the inclusion of extremely low baseline index scores (i.e., outliers) results in lower incremental QALY estimates that favour Enhanced Care. The results also reveal how incremental QALY-AUC estimates are strongly influenced by the magnitude of the between-group difference in baseline score (Figure 18b). In contrast, incremental QALY estimates from methods that adjust for baseline index score differences appear less influenced; however, the direction of treatment benefit for (adjusted) incremental estimates based on regression models differs from crude estimates (Figure 18b).

5.5 Discussion

The QALY is an outcome used in cost-utility analyses to assess the value of HRQL changes between groups, which is represented as the incremental cost divided by the incremental QALY [38,42]. End-users of generic HRQL instruments, who recognize the importance of a common end-point that captures people's preferences for health, may want to assess HRQL changes as a stand-alone study end-point. Therefore, QALY calculation methods that support the interpretability of HRQL changes at the point of application are potentially useful [20]. However, a new method must also be able to address issues common to calculating QALYs from patient-level data, namely adjusting for baseline HRQL scores, and standardizing comparisons between groups [132]. To this end, we have proposed a new QALY calculation method that adjusts for 'meaningful' change in HRQL scores (QALY-MID), comparing results from this method to standard QALY calculation methods.

When comparing group-level outcomes in the context of evaluating an intervention, an important challenge is isolating the incremental change in outcome that is due to group assignment (and the corresponding intervention) and not due to some external factor that is unequally distributed between groups [130]. In the TEAMCare trial, baseline HRQL scores were not identical across groups; therefore, adjusting for baseline scores and standardizing between-group differences in the QALY calculation may be necessary.

According to the results, the treatment of depression for patients with type 2 diabetes is expected to improve HRQL. The incremental QALY-CFB estimate differed in both direction and magnitude compared to the QALY-AUC. However, a negative correlation between weighted summary score change and baseline index score suggested that larger HRQL gains in the Enhanced Care Group might have resulted from a lower distribution of baseline index scores. A linear regression model was therefore used to standardize the distribution of baseline scores producing incremental QALY point estimates that are close to the null value of no difference between groups; however, incremental estimates varied over the range of observed baseline scores with incremental QALY-MID estimates that are more in favour of Collaborative Care (compared to QALY-R or QALY-CFB estimates).

Manca et al. identified the QALY-R method as a way to address 'measurement error' in HRQL based on the premise that there is intra-individual variability in scores due to error [67]. Measurement error is classically defined as the difference between the 'true score' and 'observed score'. For generic HRQL instruments, the true HRQL score is the *true* 'average of the public's preference for the patient-reported health state'. It is unclear what mechanism would be involved in causing erroneous intra-individual variability in scores to yield a difference between the observed HRQL score and the true HRQL score. Nonetheless, Manca et al. showed that statistical adjustment for differences in baseline HRQL scores addresses the problem of using AUC values; however, unlike the QALY-CFB method, regression also mitigated the effects of 'regression to the mean' [67]. In this regard, the negative correlation between the weighted HRQL summary change score and baseline HRQL score is consistent with regression to the mean phenomenon. Notably, regression to the mean does not offer a 'cause' for larger gains observed among patients with lower HRQL scores, rather the phenomenon is due to 'chance' (and doesn't reflect a change that is due to 'treatment' effect) [133].

A criticism of generic HRQL measures and the application of the ‘average of the general public’s preferences’ in the calculation of QALYs is that the resulting QALY, and subsequent use in a calculation of incremental QALYs, may lead to a sub-optimal decision concerning the benefits of an intervention as perceived by the patients experiencing the condition and treatment [64]. Previous research has shown that there are differences between HRQL instruments, experienced health states, ex-ante preferences from patients, and the average of the general public’s preferences [12,46,47,49,111,117]. This highlights the importance of understanding the ‘error’ or ‘uncertainty’ in HRQL measurement at the point of application as it relates to the meaning that patients give to observed changes in HRQL.

Evaluations of the psychometric properties of HRQL scores address issues in the reliability and responsiveness of the score at the point of application [46]. It is however unclear how these psychometric properties can be used to inform QALY calculations in order to better reflect ‘true scores’ and the meaning patients give to changes in HRQL. To this end, we proposed a new QALY-MID calculation that is based on the concept of MID for HRQL scores, representing the smallest change in score that is minimally important to patients. The MID has been proposed as a metric to support the interpretability of HRQL changes with empirical evidence derived from patients [31,32,80]. The MID is typically determined empirically using anchor-based methods, with supportive information from distribution-based methods [30,85,116]. Multiple anchors and multiple methods can be used to triangulate and/or give a plausible range for the MID [30,80,85,116]. However, it is important to recognize that MID estimates represent the ‘expected’ meaning that patients give to HRQL changes (i.e., group-level), which is not necessarily the same as the meaning that any one individual patient (in the study) may give to HRQL changes (i.e., individual-level) [60,80,85]. In addition, a QALY includes multiple HRQL change scores, which can be summarised by a weighted summary change from baseline. This summary change from baseline shows how the change from baseline to six months has two times the influence compared to the change from baseline to twelve months. Thus, the timing of HRQL measurement in a trial has a large influence on the QALY outcome. In this regard, the QALY-MID method is explicit in terms of how the responder-criteria are applied to change scores. This study applied criteria to the overall weighted summary change score and not to individual period change scores. The choice may reflect what is considered to be 'meaningful' for HRQL outcomes in the target population. In the treatment of depression, earlier changes that are maintained over one-year might be considered more important than later changes in HRQL, so using a weighted average may be appropriate.

The QALY-MID method applies 'responder criteria' to HRQL score changes within a patient to reflect what is an expected 'meaningful change' in HRQL from the patients' perspective, in terms of both improvement and deterioration. HRQL score changes that fall within the MID range are therefore considered not meaningful and assigned an observed value of zero in the QALY calculation, while HRQL score changes that are larger than the MID are used in the QALY calculation. This method is similar to 'duration of response' analyses (used in oncological studies), where 'non-responding' patients are assigned a duration of zero [134,135]. However, the 'crude' estimate of the incremental QALY does not account for differences in the distribution of baseline index scores (and other covariates) [130]. In this regard, patients with lower baseline index scores were more likely to experience meaningful improvement, while deterioration in HRQL was more common among patients with higher baseline scores. Therefore, regression is applied to standardize the comparison between groups, which results in a difference between crude and adjusted incremental QALY-MID estimates (adjusting for differences in index scores, sex and age). In this case, incremental QALY-MID estimates reflect the ‘average of the public’s preferences’; however, unlike the QALY-R method, the observed HRQL change scores used criteria of ‘meaningful’ improvement and deterioration in HRQL from the patients’ perspective. In summary, crude and adjusted QALY-MID estimates are larger than QALY-CFB and QALY-R estimates respectively, suggesting that ‘non-significant’ small changes in index score (that are more common in the Enhanced Care group) have the potential to influence conclusions with respect to treatment benefits (including downstream implications in cost-utility analyses) if included in the QALY calculation.

The stability of regression coefficients was assessed using recursive regression over the range of observed baseline index score values. Notably, the adjusted incremental QALY-MID estimates favor Collaborative Care as the range of baseline index scores increases until minimum index score values are included (specifically, the lowest observed baseline index score value in the Collaborative Care group). This suggests that an outlier may be responsible for the change in direction of the incremental effect estimate. In addition, the recursive regression technique also showed how the direction and magnitude of QALY-AUC estimates are positively correlated with the average difference in baseline index scores between groups. In contrast, crude incremental QALY-MID and QALY-CFB values moved in the opposite direction. Overall, this suggests that adjusting for differences in baseline scores (and other covariates) using a regression model may influence the assessment [136]. Further studies may consider assessing QALY outcomes between groups based on stratifications by baseline index scores and/or in the subset of ‘responders’ [134,137].

The main purpose of this study was to support end-users of generic HRQL instruments in non-economic assessments of HRQL changes. We presented a modified QALY calculation method that explicitly adjusts for ‘meaningful’ HRQL changes, potentially supporting the interpretability of QALY outcomes at the point of application. In addition, we considered common issues in the assessment of QALY outcomes from patient-level data including adjusting for baseline HRQL scores and standardizing differences between-groups. Importantly, regression can be used to adjust for ‘observed’ differences at baseline; however, unobserved factors may still exist. Furthermore, this study did not consider effects of missing data, wherein patients who are lost-to-follow up may differ from complete cases. Lastly, this study gives incremental QALY estimates for the purposes of assessing HRQL changes in non-economic applications; however, these estimates may still be useful in economic evaluations. Ultimately, the uncertainty in incremental QALY estimates reflects uncertainty in the value of small within individual change as well as the value of small differences between groups. Further research is required to show how uncertainty in incremental QALY estimates has downstream effects when evaluating the cost versus benefit of a technology, which, in turn, may help to inform value of information analyses [73].

Figure 14. Baseline EQ-5D-5L index score distributions for Enhanced Care and Collaborative Care groups.

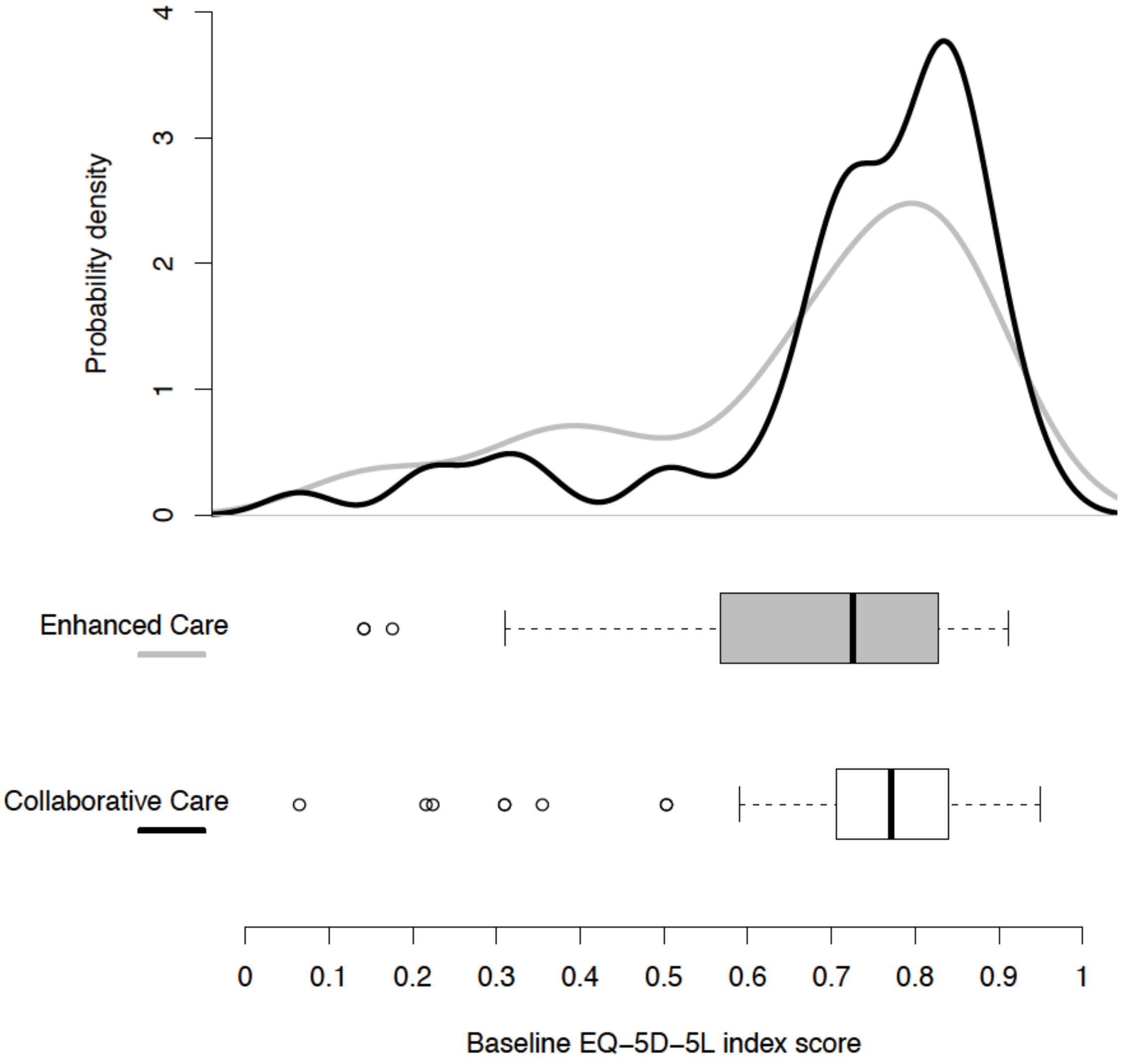
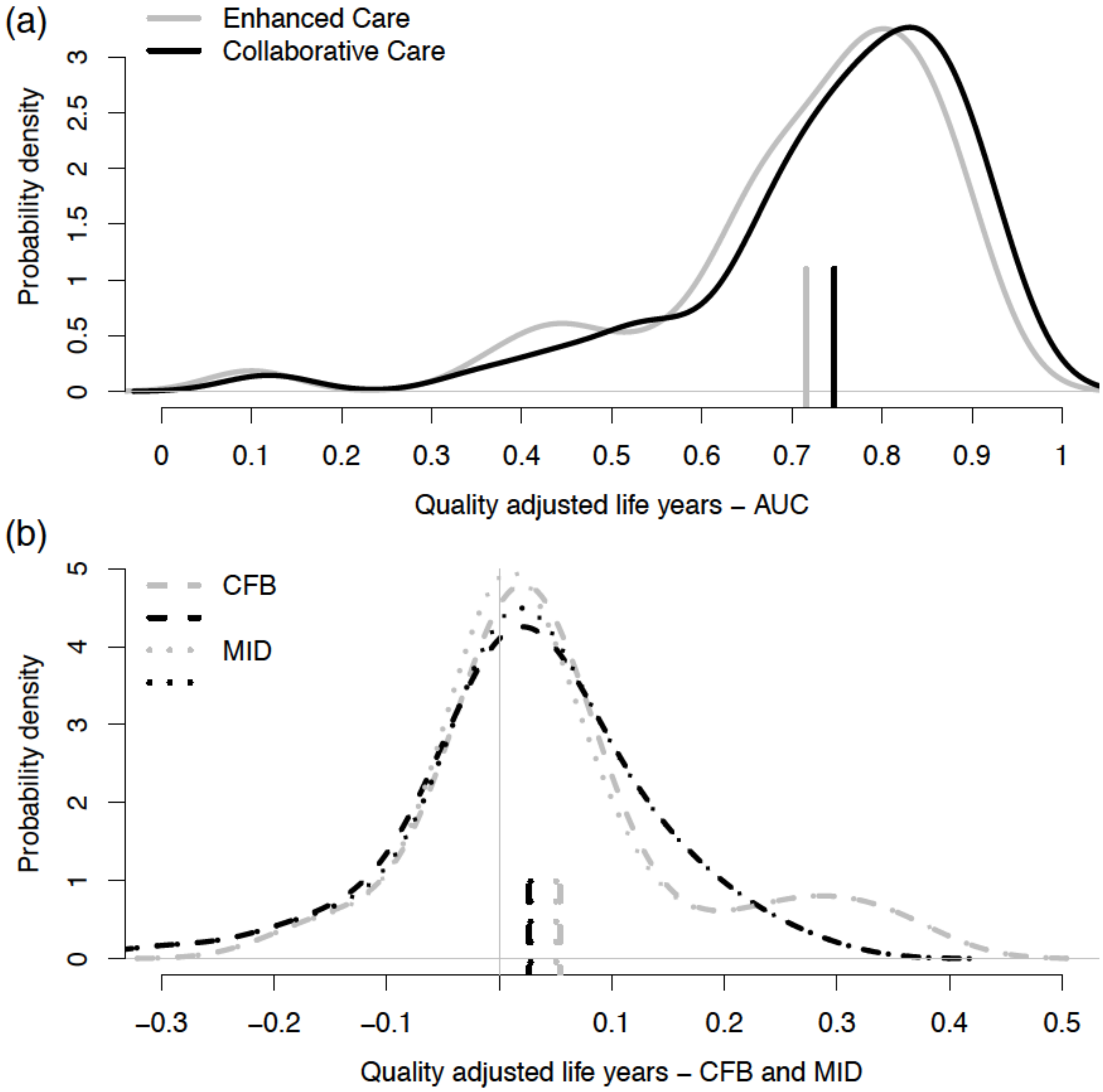
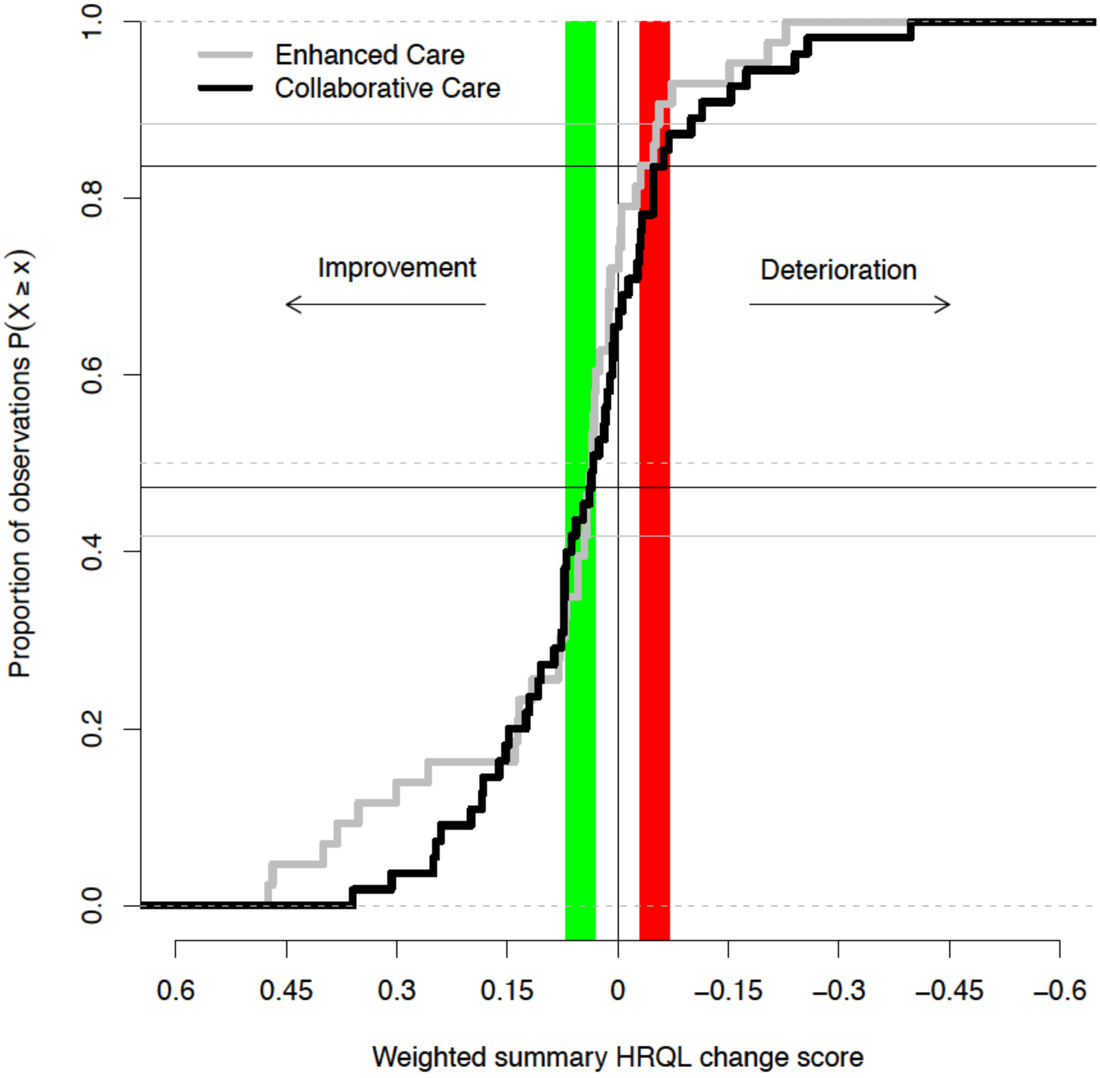


Figure 15. Quality-adjusted life year distributions for Enhanced Care (grey lines) and Collaborative Care (black lines) groups using different calculation methods: (a) Area-under curve (solid), and (b) change from baseline (dashed) and minimally important difference (dotted).



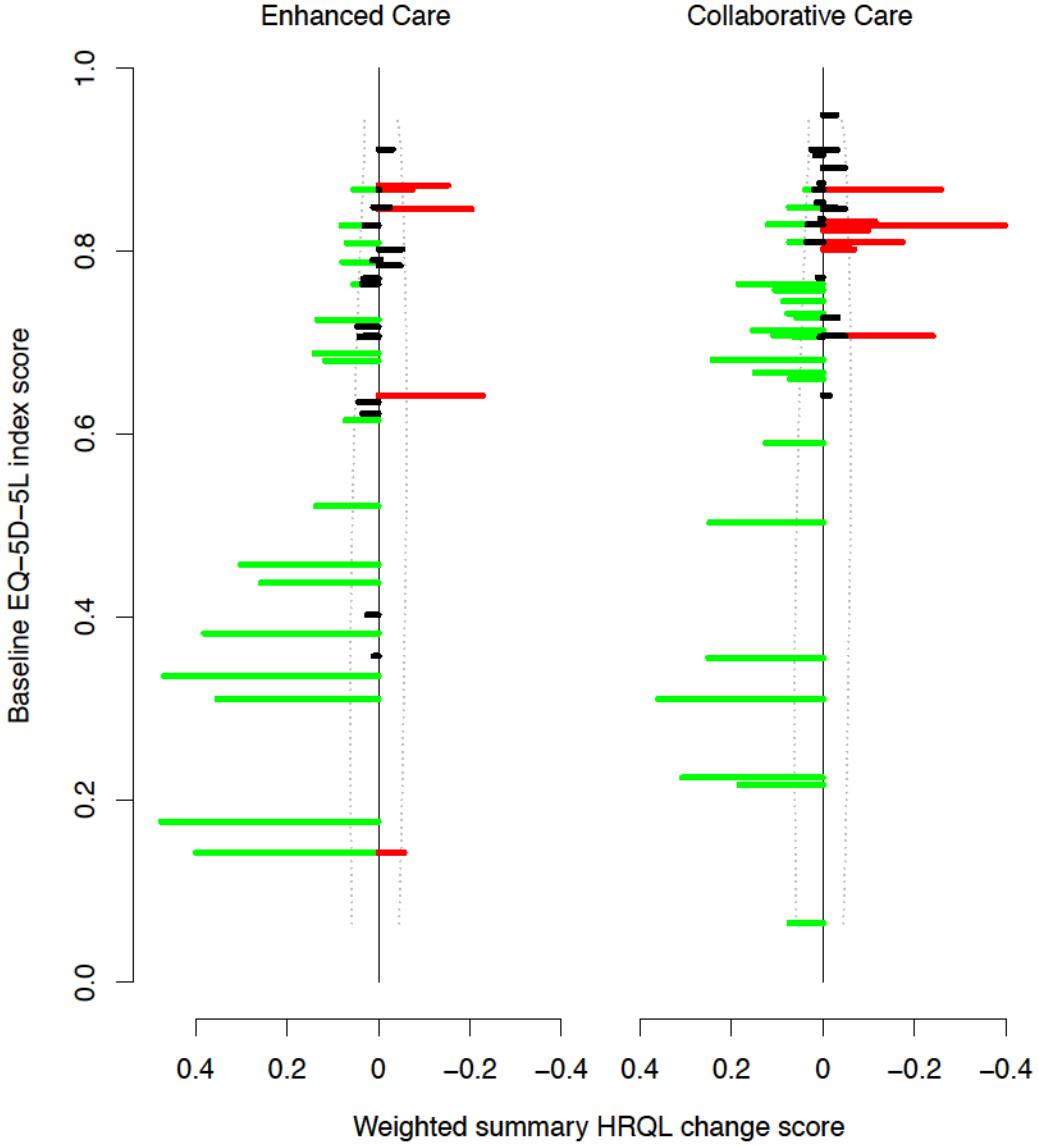
AUC, Area under curve; CFB, change from baseline; MID, minimally important difference. Vertical lines intersecting the x-axis represent the group-level averages for each quality-adjusted life year calculation method.

Figure 16. Empirical complementary cumulative distributions of the weighted summary health-related quality of life change score.



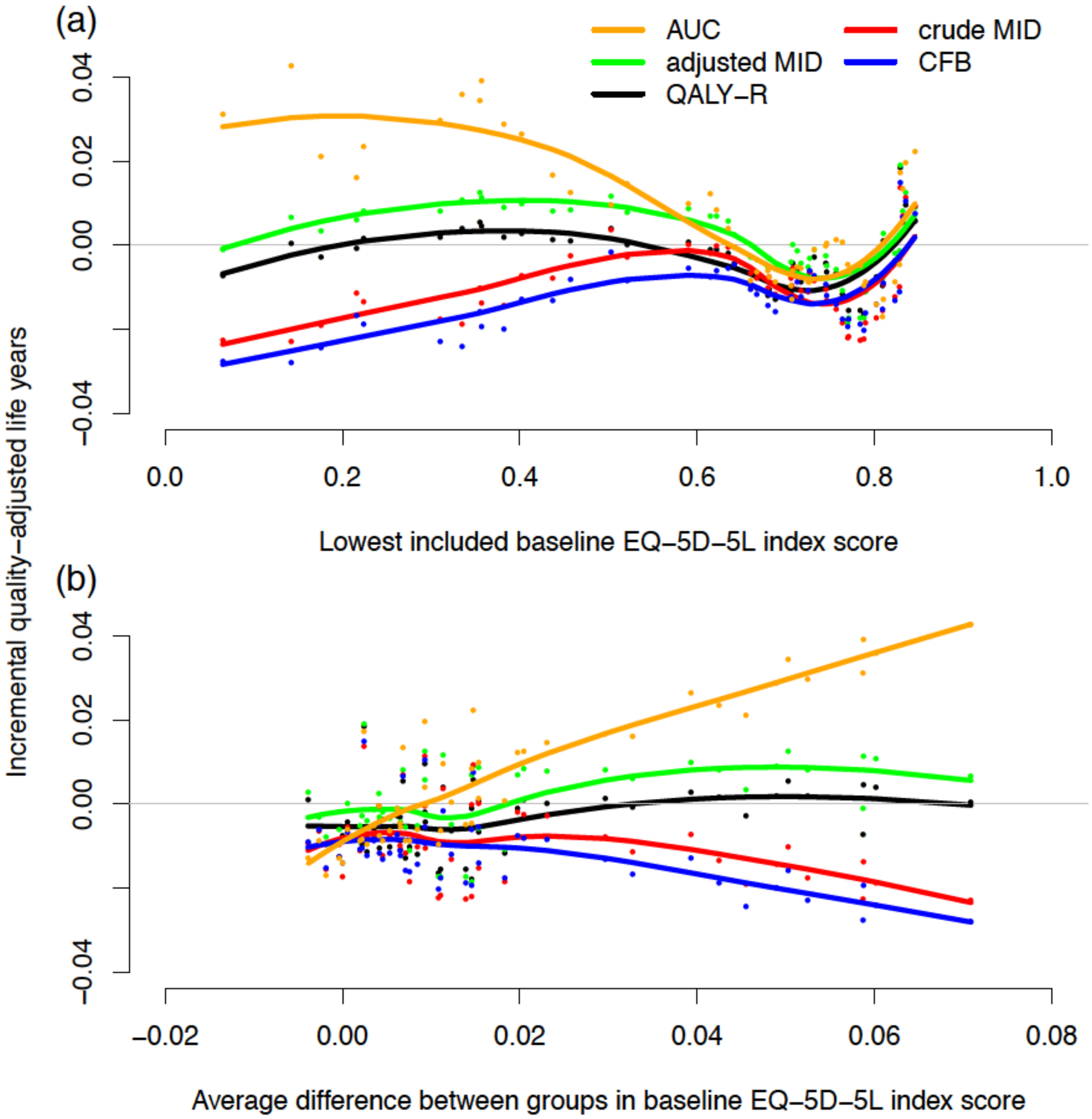
HRQL, Health-related quality of life. Green and red bars represent plausible range of minimally important improvement/deterioration. Horizontal solid lines intersecting the y-axis show the proportion of observations for Enhanced Care (grey line) and Collaborative Care groups (black line) that are expected to represent minimally important change.

Figure 17. Association between weighted summary health-related quality of life change score and baseline EQ-5D-5L index score.



HRQL, Health-related quality of life. Grey dotted line represents minimally important responder-criteria for meaningful improvement and deterioration. Green/red bars represent within-patient weighted summary HRQL change scores that are expected to represent meaningful improvement/deterioration in HRQL from the patients’ perspective.

Figure 18. Estimates of incremental quality-adjusted life year from recursive regression showing (a) the stability over the range of observed baseline EQ-5D-5L index scores, and (b) the relationship with average difference between groups in baseline EQ-5D-5L index scores.



AUC, Area under curve; CFB, change from baseline; MID, minimally important difference; QALY-R, incremental estimate from regression model. Adjusted estimates are based on regression models adjusting for differences in baseline index score, sex and age; crude estimates are based on average group-level differences. Positive values indicate larger values are observed in the Collaborative Care group. Smoothed lines are generated from local polynomial regression fitting (loess).

6. Discussion

The motivation for research focused on the minimally important difference (MID) concept is to support interpretation of generic health-related quality of life (HRQL) outcomes at the point of application. In contrast, the development of preference-based HRQL instruments such as the EQ-5D (i.e., that measure and value health) was motivated primarily by a desire to calculate quality-adjusted life years (QALYs), which in turn are interpreted in cost-utility analyses by way of between-group comparisons and cost-effectiveness thresholds [12,19,38,42]. The widespread adoption of HRQL instruments (e.g., EQ-5D) by policy-makers, clinicians, and researchers demonstrates how HRQL outcomes are important to a variety of end-users (not only in health economic applications) [1,9–11,13,45,58,138]. Indeed, the EQ-5D has become one of the most commonly used generic HRQL measures for regional and national patient-reported outcome measures programmes, informing patient and quality assurance decisions as an indicator on its own. However, there has been less development of methods (or guidance from instrument-developers) to support the appropriate interpretation of generic HRQL outcomes at the point of application (i.e., for different, non-economic evaluation, purposes) [10,60,81,139]. Moreover, end-users require methods to support interpretation of generic HRQL outcomes that also reflect the preferences of particular target populations (and not only the average of the general population) [48,58,64,80,107,140].

6.1 Overview of Research

Generic indirect preference-based measures, such as the EQ-5D-5L, use a multi attribute health descriptive system that reports the patient’s health status. To arrive at an

HRQL measure, a scoring algorithm is then applied to obtain an index score that represents the ‘average of the general population’s preferences’ for the patient-reported health state [9,22]. Thus, the chosen HRQL instrument represents a series of upstream decisions that impacts the contents of the health descriptive system (e.g., the domains and response options) and the development of the value set (e.g., source population, elicitation task, modelling assumptions, and model selection criteria) [12]. At the point of application, the psychometric properties of the HRQL instrument are assessed to demonstrate its ability to reliably detect meaningful change in the target patient population [29,46]. However, the fact that the index score represents the average of the general population’s preferences and not necessarily the preferences of patients, makes it challenging to determine if patients’ HRQL has improved, worsened or stayed the same [46,48,49,76,141].

This leads to the over-arching question that has motivated this thesis: is it possible to aid the interpretation of observed changes in HRQL index scores so that an end-user can determine if patients’ HRQL is expected to have improved or worsened? To this end, this thesis has investigated how MIDs for the EQ-5D-5L index score can be estimated and used to evaluate meaningful change in patients’ HRQL. The major take-away messages from each study are summarised below.

Instrument-defined MIDs for the EQ-5D-5L index score are specific to the general population preferences of a specific country. Chapter 2 estimated MIDs for the EQ-5D-5L index score using a method based on internal anchors defined by the health descriptive system (i.e., single-level transitions). As expected, this study found MID estimates varied by country-specific scoring algorithm. Moreover, depending on the scoring algorithm, MIDs can vary across the range of baseline index scores. Furthermore, removing maximum-valued scoring parameters for each dimension decreased the magnitude of MID estimates. These estimates can be used to give a plausible range of index score changes/differences that are expected to be minimally important, given the granularity of the EQ-5D-5L descriptive system and local population preferences for described health states.

Empirical evidence from multiple MID estimation methods shows that the smallest change in EQ-5D-5L index score that is expected to represent minimally important change in HRQL for adults with type 2 diabetes varies by direction of change and baseline score. Chapter 3 estimated MIDs for EQ-5D-5L index score changes over-time in adults with type 2 diabetes using the instrument-defined and anchor-based approaches. There was general agreement between approaches in what represents the smallest change in EQ-5D-5L index score that is meaningful to this patient population. However, we found differences in MID estimates based on direction of change, baseline index score, and clinically relevant subset of the population. These estimates can be used to evaluate the significance of observed group-level changes over-time with greater limitations when applied at the individual-level.

Small differences in EQ-5D-5L index score are ambiguous, such that a difference needs to be large enough to be meaningfully interpretable. Chapter 4 analysed the ordinal and cardinal responses of participants in the Canadian Valuation Study to determine how (small) EQ-5D-5L index score differences are interpreted. This study showed that there is variation in preferences between and within individuals for transitions between health states, which is particularly emphasized when the index score difference is small. Based on specified criteria, we found evidence that an index score difference needs to be larger than zero (e.g., >+/-0.05) to be expected to represent individuals' interpretation of the transition between health states. These results lend support to the applicability of MID estimates in supporting the interpretation of EQ-5D-5L index score changes.

The assessment of between-group QALY outcomes can use the MID to adjust for meaningful within-patient change in HRQL. Chapter 5 investigated QALY calculations and the importance of adjusting for between-group differences in baseline scores, as well as within-patient meaningful change in HRQL. To this end, a new QALY calculation method applied MID estimates to evaluate within-patient EQ-5D-5L index score changes over time using data from a trial for depression treatment in patients with type 2 diabetes. Different QALY calculation methods produced different results, demonstrating the challenges in assessing QALY outcomes and adjusting for errors and meaning in HRQL measurement. This emphasizes the uncertainty in QALY outcomes and the value of using multiple approaches based on different criteria to build understanding of HRQL changes that are relevant to end-users at the point of application.

6.2 General Implications of Research

There is growing interest in patient-centred care and shared decision-making [85,142]. In addition, the increasing burden of chronic illnesses requires redefining therapeutic end-points and reshaping the delivery of care; likewise, changing societal attitudes in many countries has resulted in the legalization of medically assisted dying [143,144]. Overall, this demonstrates the importance of understanding patients' HRQL.

Thus, there is a need to incorporate HRQL outcomes at all health system levels: micro (e.g., individual patient care), meso (e.g., guideline development for treatment of a specific condition) and macro (e.g., drug reimbursement decisions), such that HRQL can be evaluated to inform evidence-based decisions. However, this presents a salient issue, which is: can a generic HRQL measure be used to inform decision-making at these multiple levels?

Condition-specific measures exist because different aspects of health have varying relevance across health conditions [3]. In addition, patients with similar characteristics, such as experience living with a particular condition, may have preferences that are systematically different from others [47,49,64,141]. Therefore, different HRQL measures (differing in content and value sets) can be useful for different purposes [3]. However, it would not be possible to collate evidence across applications unless additional methods are used to map between scores from different HRQL measures, which would still require some ‘standard’ or a common HRQL measure [145]. The advantage of a generic HRQL measure is that evidence across multiple applications can be integrated to inform decisions [1,3]. For this reason, a generic measure is based on the average of the general population’s preferences [27]. It is further argued that these societal preferences ought to be based on hypothetical health states (as opposed to experienced health states) [49]. In practice, however, the goals of medicine and the provision of healthcare are to improve the health of patients and the general public [49,126]. Therefore, it is likely not possible to completely base decisions on an end-point that characterises a single ‘way of knowing’ [107]. The application of appropriate methods that address the shortcomings of a single HRQL measure may allow its usefulness to be extended. In this way, quantifying the smallest change in a generic indirect preference-based measure that is minimally important to patients might be a useful method of combining information from a generic measure with a fit-for-purpose criterion.

Similar to common clinical measures (e.g., blood pressure, weight, blood glucose), the scores that are produced by an HRQL instrument are used to ‘infer’ health (i.e., a health construct), and thus have associated errors that differ from the strict ‘objective’ properties of the score. Studying the association between scores and health outcomes of interest allows end-users to consider ‘meaningful’ properties of the score that are relevant to assessing changes in health (e.g., psychometric properties). However, the precise nature of these properties and magnitude of these ‘errors’ may differ across instruments and applications (e.g., different types of patients and/or health outcomes). As an example, in estimating the weight of adults and neonates, we would not be comfortable with the same degree of error in a measurement instrument. At one extreme, applying a HRQL measure at the level of the individual (i.e., micro-level) will result in the greatest uncertainty in terms of what a score (or score change) represents in terms of the patient’s health [107]. Importantly, this does not make a score useless; rather, using the same instrument to collect scores from patients in different applications allows for evidence integration that can be tailored to the context. The appropriate interpretation of scores requires the development of tailored methods and criteria that are relevant to the measurement objectives as well as the corresponding error in measurement at the point of application. To this end, this research has sought to inform the interpretation of HRQL scores at the point of application by way of the MID concept.

The interpretation of HRQL outcomes is challenging for the following reasons.

First, HRQL is a holistic all-encompassing construct that is important to all patients; however, defining it raises questions about what aspects of health are *judged most important* to HRQL [12,45,146]. Second, unlike health profile measures, index scores for HRQL require the elicitation of people's health preferences [22]. This presents a trade-off between obtaining the preferences of patients at the point of application (i.e., internal validity) and the generalizability of results to allow for comparisons involving other patients and treatments (external validity). Generic indirect preference-based HRQL measures arose to fill this need [19]; however, they are imperfect, thus the appropriate interpretation of scores require careful consideration [46]. The following sections discuss the implications of this thesis' results in regards to (6.2.1) estimating MIDs for generic preference-based measures of HRQL, (6.2.2) the interpretability of generic HRQL measures at multiple-levels of decision-making, and (6.2.3) equity implications from MIDs.

6.2.1 Estimating MIDs for Generic Preference-Based Measures of HRQL

As the name implies, the purpose of MID estimation is determining what is minimally important from the patients' perspective. In regards to HRQL, such an estimate will need to encompass the multiple dimensions of health that characterise patients' HRQL. In this way, estimating MIDs for a generic HRQL measure likely requires multiple anchors capturing various attributes of health, wherein each anchor has a change in score that represents, from the patients' perspective, a small yet meaningful improvement or deterioration in health. The MID estimate for the generic HRQL score is then the aggregation of individual MID estimates from different anchors that triangulates an overall estimate (or plausible range) for HRQL. In **Chapters 2 and 3**, MIDs for the

EQ-5D-5L index score were estimated using the instrument-defined approach based on simulation of all possible single-level transitions (**Chapter 2**), and only from EQ-5D-5L health states reported by a sample of patients with type 2 diabetes (**Chapter 3**). The following discusses the strengths and limitations of these MID estimation studies.

MID estimates of the EQ-5D-5L index score presented in **Chapter 2** are directly relevant to end-users of the EQ-5D instrument. These estimates appear to differ from MID estimates of the EQ-5D-3L [86]. This could be due to a variety of reasons, including the different range of scores, changes in attitudes toward health over-time, and differences in the health descriptive systems between the 3L and 5L. Since the instrument-defined approach simulates all possible single-level transitions with weights applied by the scoring algorithm representing the average of the general population's preferences, the resulting estimates provide plausible ranges for MIDs of the EQ-5D-5L index score, but are not specific to a particular patient population. Using MID estimates as criteria can help end-users of the instrument determine whether meaningful change in HRQL is expected at the point of application. Using the instrument-defined approach, it is however unclear if changes smaller than MID estimates represent 'trivial' changes in HRQL (i.e., 'about the same') or ambiguous changes that cannot meaningfully discriminate transitions to better states of health from transitions to worse states of health.

Unlike the MID estimates from Chapter 2 that were based on simulation, the MID estimates from **Chapter 3** reflect what change in EQ-5D index score is meaningful for adults with type 2 diabetes. Thus, these MID estimates are useful to end-users who are looking to support type 2 diabetic care and want to understand the expected significance of observed HRQL changes from the patients' perspective. MID estimates from the different approaches were in general agreement, providing a range of small changes in EQ-5D index score that are expected to represent minimally important change in HRQL outcomes for adults with type 2 diabetes. However, these MID estimates are based on the Canadian 5L value set, and as informed by the first study (Chapter 2), replication of these results in other countries (using other EQ-5D scoring algorithms) is required to determine the usefulness of different MID estimation methods, and the generalizability of HRQL MID estimates.

Similarly, an MID derived from aggregating changes reported by individuals to report an overall average change at the 'group-level' may have limitations when it is used to evaluate meaningful change at the 'individual-level'. One criticism of MID estimation methods that are based on averages of change scores is that a proportion of responses (e.g., 50%) with a meaningful change in the anchor will be below the 'average' of the change score identified by the MID. Any threshold can, to varying degrees, represent 'meaningful change', wherein a threshold of lower magnitude will have higher sensitivity (as it is more likely to capture all of the true positives), but will also have lower specificity, being unable to discriminate HRQL gains from HRQL losses or changes that are 'about the same'. To this end, the 'average' represents a threshold of change that is 'expected' to represent minimally important changes in HRQL outcomes for patients in the target population. Likewise, individual MID estimates from different anchors are aggregated to obtain an overall MID estimate for the generic HRQL score. However, an individual who 'improves' in one anchor may 'worsen' in another. Thus, the overall 'expected' MID for the generic HRQL score is determined by the average of individual MID estimates from different anchors.

6.2.2 The Interpretability of Generic HRQL Measures at Multiple-Levels of Decision-Making

As previously discussed, MID estimates obtained from **Chapters 2 and 3** are based on studies that aggregate responses (from multiple observations), such that MIDs are generally considered to be more applicable in the assessment of a group-level change in HRQL [35,60,147]. Of course, 'group-level' estimates of population parameters are important in evidence synthesis [147,148]. For instance, randomized controlled-trials, utilize group-level properties by equalising the distribution of unobserved confounders across treatment-arms [149]. In this way, the effects from unobserved confounders are, on average, assumed to be the same, which allows for the assessment of the expected treatment effect at the group-level [149]. However, this presents a gap when attempting to apply group-level results at more micro-levels of care such as informing on the most appropriate care for a patient [107]. Responder-analyses attempts to facilitate this gap in inference; however, these are not without issue [85,150]. In particular, responder-analyses suffer from loss of information through dichotomization, the poor identification of measurement error, and improperly attributing 'cause' to observed responses [151–153]. **Chapters 4 and 5** present methods showing how MID estimates for generic HRQL scores obtained from group-level analyses have relevance at multiple-levels of decision-making, specifically the individual-level. While HRQL scores represent the average of the general population's preference, which is used to inform macro-level decisions, **Chapter 4** shows how information in individuals' stated preferences is used to un-pack the meaning of observed small differences in EQ-5D-5L index score for individuals. **Chapter 5** then uses results from the previous Chapters to develop a QALY calculation method that incorporates responder-criteria in terms of what is expected meaningful change in generic HRQL scores for patients.

Chapter 4 provided a novel analysis of cardinal and ordinal preferences elicited from respondents in the Canadian EQ-5D-5L valuation study, demonstrating the variability in interpretation of small EQ-5D index score differences. This heterogeneity in preferences illustrates the ambiguity of small changes in EQ-5D index scores, suggesting that a change in EQ-5D index score needs to be 'large enough' to be expected to represent an improvement or deterioration in health. In this regard, it is not that small EQ-5D index score changes are not meaningful as different respondents interpret the same change differently (i.e., have different preferences); however, the usefulness of a score change is derived from its expected ability to represent the majority of respondents' preferences (i.e., from the general population) and discriminate what individuals consider to be better and worse HRQL outcomes. This interpretation of MIDs for HRQL index scores is similar to the interpretation from Johnston et al. (2010) who proposed that treatment may have an important impact on many patients even when pooled effect estimates are less than the MID, but the likelihood of benefit is progressively less [79]. Moreover, research has suggested that 'preference-weighting' in generic indirect preference-based HRQL outcomes (compared to an un-weighted or severity/misery index score) may be more important at the individual-level than the group-level in terms of discriminating meaningful differences in health [154]. The preference heterogeneity for small differences in generic HRQL scores, allowed for the quantification of a threshold difference in HRQL score that is expected to represent what individuals consider to be an improvement or deterioration in HRQL (according to their stated ex ante preferences).

Analysis of the preference heterogeneity resulted in threshold values similar to MID estimates. In this way, preference heterogeneity may link the interpretation of changes in EQ-5D index scores with the MID, such that MID values may be applicable at the micro (i.e., patient) level to suggest what change in index score is expected to represent a meaningful change in HRQL for a patient. Importantly, however, the use of generic HRQL scores to inform decision-making, especially at the patient-level, is not fixed or deterministic, but may be useful in communicating to the patient what can be expected based on the available evidence [107]. This has implications in the assessment of HRQL outcomes (including in cost-utility analyses), in which an observed change in EQ-5D-5L index score could be supplemented with analyses that elucidate preference heterogeneity, determining whether index score changes are expected to represent an improvement or deterioration in HRQL for patients (at the point of application) [74]. Further research is required to determine why preference heterogeneity exists; for example, does preference heterogeneity represent differences in attitudes toward HRQL or differences in response-scale interpretation [155]. This study was based on responses from the general population using hypothetical health states in standardized tasks; however, preference heterogeneity may exist in any target population using any type of elicitation task for experienced or hypothetical health states. Therefore, future analyses following a similar methodology may be useful in showing how changes/differences in generic HRQL scores are interpreted in different contexts. Further comparisons between estimates of the smallest meaningfully interpretable difference from preference heterogeneity and MID estimates for the target population may also help to support the link between expected within-patient change and group-level MID estimates.

Importantly, MIDs help to support interpretation by categorizing generic HRQL scores into changes that are expected to represent meaningful improvement versus deterioration from the patients' perspective. In this regard the MID is similar to a responder-criterion, which can be applied at the individual-level to assess meaningful within-patient change in HRQL. **Chapter 5** explored an approach to apply MID results from the previous studies to support the assessment of QALY outcomes in a study comparing two types of care for depression treatment in patients with type 2 diabetes. Currently, guidance from the U.S. Food and Drug Administration flags composite HRQL measures based on community preferences as problematic in assessing meaningful change [60]. This study is therefore useful for demonstrating how empirical studies of the MID can be applied in QALY calculations to adjust for meaningful within-patient HRQL changes; however, there are limitations in the inferences that can be made. Small changes in HRQL outcomes at both the individual and group levels are common; however, increasingly common are small group-level incremental differences between comparators [156]. To this end, interpretation of between-group differences in QALY outcomes requires additional approaches not captured by MIDs estimated from changes over-time [157]. Further MID studies are required to link the assessment of observed HRQL score changes with the assessment of between-group differences in HRQL change.

6.2.3 Equity Implications from MIDs

Overall, the results of this thesis support how MID estimates may be useful for assessing observed small changes in a generic HRQL score, in terms of identifying what is meaningful to patients' HRQL at the point of application. However, estimating and applying MIDs is not the same as quantifying an individual's true HRQL change. While ‘doing what is best for the patient’ is an important goal of individualized patient care, it is perhaps not a practical approach to take for assessing overall HRQL gains/losses at the population-level (e.g., in randomized controlled-trials) as this will involve inter-personal trade-offs in HRQL [48,107,145]. This is precisely the strength of a generic indirect measure of HRQL [1,3]. In this regard, applying MID estimates to interpret outcomes from generic HRQL scores will not subtract from its strength, rather it explicitly conveys the value of identifying HRQL changes that are meaningful to patients at the point of application. However, recognizing that MIDs for generic indirect preference-based HRQL scores may depend on the instrument (and characteristics of the value set), medical condition under investigation, direction of change, and baseline score, also raises some important equity considerations.

Generally, the measurement of HRQL by way of a generic instrument and the downstream calculation of QALYs are thought to align with a horizontal equity position, i.e., ‘equal treatment of equals’ [27,38]. The cardinal properties of the QALY are asserted in the statement ‘a QALY is a QALY is a QALY’ to mean that a QALY is the same value regardless of the identity or condition of the patient who receives it [27,38,157]. While this may be a reasonable default position to assume, there is much research to suggest that not all QALYs are the same. Specifically, decision-makers and the public may place importance on other observable characteristics such as inequities in health observed across age and socio-economic spectrums, pro-longing life near death, as well as the rarity of the condition and scarcity of available treatment [158–160]. In addition, empirical attributes of the QALY such as its magnitude, impact on length of life versus quality of life, direction of change, and baseline HRQL score can influence what value members of the general public or patients award a QALY [125,157,161]. These types of considerations imply that different QALYs are treated differently, depending on the type of QALY and the observable characteristics of those who are gaining or losing QALYs. Through the use of equity weightings, the framework for cost-utility analyses can be extended to include a vertical equity position; however, there may still exist challenges in 'weighting' the health that is displaced, as there may not be information on the identity of those whose health is forgone in allocation decisions [27,38,162].

In light of this previous description of equity, we might also consider the equity implications from estimating and applying different MIDs in different situations. In particular using different MIDs (based on observable characteristics of the patient population) suggests a vertical equity position. This is evident in **Chapter 5**, which used MIDs as responder-criteria applying a zero weighting to QALYs with HRQL changes less than the MID before calculating the incremental group-level QALY difference between comparators. In this case both comparator groups were for the same target population, patients with type 2 diabetes and depression; however, end-users may wish to compare HRQL changes from care programmes involving different target patient populations, for example management of type 2 diabetes, and hip and knee arthroplasty. In such a scenario, just as the responsiveness of the HRQL instrument might differ between the two patient populations, the MIDs used to evaluate HRQL changes may differ, such that patients belonging to one group have MIDs that are larger in magnitude than MIDs for patients in the other group. Ultimately, this suggests that not all small HRQL changes are valued equally, such that an equivalent small HRQL change may be expected to represent meaningful change for one patient, but not for another patient.

Further research is required to determine if differences in MIDs or responsiveness of HRQL scores cause there to be differences in how likely it is for different patients to achieve meaningful changes in HRQL.

Again, there is a difference between individual-level and group-level HRQL changes. For instance, two comparator groups may achieve the same group-level average change in HRQL but have differences in the distribution of individual-level HRQL changes. Suppose in one group there is heterogeneity in HRQL changes such that a few patients have HRQL gains larger than the MID responder-criteria, while the remaining individuals have HRQL index score changes less than the MID. Now suppose that in the other comparator group there is little variation in HRQL index score changes, such that every patient has positive index score changes that are less than the MID responder-criteria. In this case, the average group-level change in HRQL will be affected by the application of MID responder-criteria perhaps resulting in a larger incremental between-group difference. Further assuming some form of value-based pricing, larger changes in HRQL may be associated with increased costs. To this end, additional research is required to understand how the application of MIDs in the evaluation of HRQL changes might impact allocation decisions and/or incentivize producers (e.g., drug manufacturers). In the earlier example, if the incremental gain in HRQL is increased by application of MID responder criteria, it would suggest that greater value is placed on a few individuals achieving HRQL gains larger than MIDs compared to everyone achieving small positive changes in index scores that may be ambiguous. This emphasizes the need to make explicit our decision-criteria, and continue to develop methods that address gaps between evidence and what society expects from its healthcare: to maximize quality of life through patient-centred access and delivery of care [12,48,146,163].

6.3 Conclusion

> *Economics, like medicine, is imperfect. The challenge for practitioners of each is to ensure that the perfect does not drive out the good. Our practices may at times be imperfect, but that should not inhibit our drive to improve clinical practice and economic activity for the benefit of all our patients and citizens. We all must strive to avoid confused analysis in displays of modest understanding of each other's work.*
>
> -Parkin D, Appleby J, and Maynard A. Economics: the biggest fraud ever perpetrated on the world? Lancet 2013;382:11–5.

The measurement of HRQL and, in turn, the QALY, have important roles in health economics and outcomes research, in which the objective is to frame the value of health gains/losses in terms of what is meaningful to patients such that the potential trade-off between length and quality of life is captured. Because there is no single all-encompassing definition of HRQL, its measurement by any one instrument will reflect a series of judgments made over the course of the instrument's development [12]. These judgments shape decisions involving two main components common to all HRQL instruments: (1) the health descriptive system (i.e., content, question framing, reference period, etc.), and (2) preference value sets (i.e., elicitation task, hypothetical versus experience-based health states, target population, etc.) to produce an HRQL score. Even though psychometric properties of reliability and responsiveness can be evaluated, these are not sufficient criteria to determine if an HRQL score is 'valid' at the point of application, as there is no ‘gold-standard’ of HRQL to compare against [9,46]. Therefore, the assessment of observed HRQL changes also requires judgments. Common to all judgments are implied criteria to determine what is ‘important’, ‘appropriate’, ‘relevant’, ‘meaningful’, ‘reasonable’ or simply ‘good enough’ [107]. The multiple instruments currently available and continued development of new HRQL measures suggests that we are always looking to ‘do better’ [13].

The purpose of estimating and applying MIDs is to make explicit our *judgment* of observed HRQL changes: a criterion that can be expected to represent minimally important changes in HRQL as determined by patients at the point of application. It is an approach to assigning meaning to changes in HRQL scores: how big is big, how small is small? The fact that there are different MIDs for different HRQL instruments in different patient populations does not invalidate an instrument’s scores; rather it explicates the value of variation – the different perspectives on HRQL and the importance of assessing it. In turn, the methods to estimate and apply MIDs to support HRQL interpretation require careful scrutiny and continued development.

This research found MID estimates that reflect greater than zero change in EQ-5D-5L index score, suggesting that observed index score changes smaller than the MID (i.e., near 0) may not adequately represent HRQL improvement or deterioration from the patients’ perspective (i.e., the change in score is ambiguous). Therefore, the MID may be useful in determining whether or not the observed HRQL change is expected to represent meaningful change in patients’ HRQL. In doing so, there is an explicit incorporation of patients’ HRQL in the interpretation of HRQL outcomes from generic indirect preference-based measures.

References

[1] Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Ann Intern Med 1993;118:622–9.

[2] Canadian Institute for Health Information. Health outcomes of care: an idea whose time has come. Ottawa, Canada: CIHI; 2012.

[3] Ahmed S, Berzon RA, Revicki DA, Lenderking WR, Moinpour CM, Basch E, et al. The use of patient-reported outcomes (PRO) within comparative effectiveness research: implications for clinical practice and health care policy. Med Care 2012;50:1060–70. doi:10.1097/MLR.0b013e318268aaff.

[4] Hennessy CH, Moriarty DG, Zack MM, Scherr PA, Brackbill R. Measuring health-related quality of life for public health surveillance. Public Health Rep 1994;109:665–72.

[5] Romero M, Vivas-Consuelo D, Alvis-Guzman N. Is health related quality of life (HRQoL) a valid indicator for health systems evaluation? Springerplus 2013;2:664. doi:10.1186/2193-1801-2-664.

[6] Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res 2005;14:1523–32.

[7] Xie F, Pullenayegum E, Gaebel K, Bansback N, Bryan S, Ohinmaa A, et al. A time trade-off-derived value set of the EQ-5D-5L for Canada. Med Care 2016;54:98–105. doi:10.1097/MLR.0000000000000447.

[8] Brauer CA, Rosen AB, Greenberg D, Neumann PJ. Trends in the measurement of health utilities in published cost-utility analyses. Value Health 2006;9:213–8. doi:10.1111/j.1524-4733.2006.00116.x.

[9] Richardson J, McKie J, Bariola E. Multiattribute utility instruments and their use. In: Cuyler AJ, editor. Encyclopedia of Health Economics, vol. 2, San Diego, United States: Elsevier; 2014, p. 341–57. doi:10.1016/B978-0-12-375678-7.00505-8.

[10] Devlin NJ, Brooks R. EQ-5D and the EuroQol group: past, present and future. Appl Health Econ Health Policy 2017;15:127–37. doi:10.1007/s40258-017-0310-5.

[11] Devlin NJ, Appleby J. Getting the most out of PROMS: Putting health outcomes at the heart of NHS decision-making. London, United Kingdom: The King’s Fund; 2010.

[12] Pickles K, Lancsar E, Seymour J, Parkin D, Donaldson C, Carter SM. Accounts from developers of generic health state utility instruments explain why they produce different QALYs: a qualitative study. Soc Sci Med 2019;240:112560. doi:10.1016/j.socscimed.2019.112560.

[13] Brazier JE, Rowen D, Lloyd A, Karimi M. Future directions in valuing benefits for estimating QALYs: is time up for the EQ-5D? Value Health 2019;22:62–8. doi:10.1016/j.jval.2018.12.001.

[14] Boyce MB, Browne JP, Greenhalgh J. The experiences of professionals with using information from patient-reported outcome measures to improve the quality of healthcare: a systematic review of qualitative research. BMJ Qual Saf 2014;23:508–18. doi:10.1136/bmjqs-2013-002524.

[15] Hunter C, Fitzpatrick R, Jenkinson C, Darlington A-SE, Coulter A, Forder JE, et

al. Perspectives from health, social care and policy stakeholders on the value of a single self-report outcome measure across long-term conditions: a qualitative study. BMJ Open 2015;5:e006986. doi:10.1136/bmjopen-2014-006986.

[16] Boyce MB, Browne JP. Does providing feedback on patient-reported outcomes to healthcare professionals result in better outcomes for patients? A systematic review. Qual Life Res 2013;22:2265–78. doi:10.1007/s11136-013-0390-0.

[17] Schlesinger M, Grob R, Shaller D. Using patient-reported information to improve clinical practice. Health Serv Res 2015;50 Suppl 2:2116–54. doi:10.1111/1475-6773.12420.

[18] Black N, Burke L, Forrest CB, Ravens Sieberer UH, Ahmed S, Valderas JM, et al. Patient-reported outcomes: pathways to better health, better services, and better societies. Qual Life Res 2016;25:1103–12. doi:10.1007/s11136-015-1168-3.

[19] MacKillop E, Sheard S. Quantifying life: understanding the history of quality-adjusted life-years (QALYs). Soc Sci Med 2018;211:359–66. doi:10.1016/j.socscimed.2018.07.004.

[20] Kind P, Lafata JE, Matuszewski K, Raisch D. The use of QALYs in clinical and patient decision-making: issues and prospects. Value Health 2009;12 Suppl 1:S27-30. doi:10.1111/j.1524-4733.2009.00519.x.

[21] U.S. Food and Drug Administration. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims. Rockville, United States: FDA; 2009.

[22] Karimi M, Brazier J. Health, health-related quality of life, and quality of life: what is the difference? Pharmacoeconomics 2016;34:645–9. doi:10.1007/s40273-016-0389-9.

[23] Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011;20:1727–36. doi:10.1007/s11136-011-9903-x.

[24] Feng Y, Parkin D, Devlin NJ. Assessing the performance of the EQ-VAS in the NHS PROMs programme. Qual Life Res 2014;23:977–89. doi:10.1007/s11136-013-0537-z.

[25] Xie F, Gaebel K, Perampaladas K, Doble B, Pullenayegum E. Comparing EQ-5D valuation studies: a systematic review and methodological reporting checklist. Med Decis Mak 2014;34:8–20. doi:10.1177/0272989X13480852.

[26] Norman R, Cronin P, Viney R, King M, Street D, Ratcliffe J. International comparisons in valuing EQ-5D health states: a review and analysis. Value Health 2009;12:1194–200.

[27] Canadian Agency for Drugs and Technologies in Health. Guidelines for the economic evaluation of health technologies: Canada. 4th ed. Ottawa, Canada: CADTH; 2017.

[28] Payakachat N, Ali MM, Tilford JM. Can the EQ-5D detect meaningful change? A systematic review. Pharmacoeconomics 2015;33:1137–54. doi:10.1007/s40273-015-0295-6.

[29] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737–45. doi:10.1016/j.jclinepi.2010.02.006.

[30] Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008;61:102–9. doi:10.1016/j.jclinepi.2007.03.012.

[31] Francis DO, McPheeters ML, Noud M, Penson DF, Feurer ID. Checklist to operationalize measurement characteristics of patient-reported outcome measures. Syst Rev 2016;5:129. doi:10.1186/s13643-016-0307-4.

[32] Johnston BC, Ebrahim S, Carrasco-labra A, Furukawa TA, Patrick DL, Crawford MW, et al. Minimally important difference estimates and methods: a protocol. BMJ Open 2015;5:e007953. doi:10.1136/bmjopen-2015-007953.

[33] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. Qual Life Res 2010;19:539–49. doi:10.1007/s11136-010-9606-8.

[34] Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407–15.

[35] King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoecon Outcomes Res 2011;11:171–84. doi:10.1586/erp.11.9.

[36] Angst F, Aeschlimann A, Angst J. The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. J Clin Epidemiol 2017;82:128–36. doi:10.1016/j.jclinepi.2016.11.016.

[37] Turner D, Schünemann HJ, Griffith LE, Beaton DE, Griffiths AM, Critch JN, et al. The minimal detectable change cannot reliably replace the minimal important difference. J Clin Epidemiol 2010;63:28–36. doi:10.1016/j.jclinepi.2009.01.024.

[38] Olsen JA. Principles in health economics and policy. 2nd ed. New York, United States: Oxford University Press; 2017.

[39] Arrow KJ. Uncertainty and the welfare economics of medical care. Am Econ Rev 1963;53:941–73.

[40] Brouwer WB, Culyer AJ, van Exel NJ, Rutten FF. Welfarism vs. extra-welfarism. J Health Econ 2008;27:325–38. doi:10.1016/j.jhealeco.2007.07.003.

[41] Tappenden P. Problem structuring for health economic model development. In: Cuyler AJ, editor. Encyclopedia of Health Economics, vol. 3, San Diego, United States: Elsevier; 2014, p. 168–79. doi:10.1016/B978-0-12-375678-7.01410-3.

[42] Drummond M, Sculpher M, Torrance GW, O'Brien B, Stoddart GL. Methods for the economic evaluation of health care programmes. 3rd ed. New York, United States: Oxford University Press; 2005.

[43] Brazier J, Deverill M, Green C. A review of the use of health status measures in economic evaluation. J Health Serv Res Policy 1999;4:174–84. doi:10.1177/135581969900400310.

[44] Torrance GW. Measurement of health state utilities for economic appraisal: a review. J Health Econ 1986;5:1–30.

[45] Olsen JA, Misajon RA. A conceptual map of health-related quality of life dimensions: key lessons for a new instrument. Qual Life Res 2020;29:733–43. doi:10.1007/s11136-019-02341-3.

[46] Brazier J, Ara R, Rowen D, Chevrou-Severac H. A review of generic preference-based measures for use in cost-effectiveness models. Pharmacoeconomics 2017;35:21–31. doi:10.1007/s40273-017-0545-x.

[47] Menzel PT. Utilities for health states: whom to ask. In: Cuyler AJ, editor. Encyclopedia of Health Economics, vol. 3, San Diego, United States: Elsevier; 2014, p. 417–24. doi:10.1016/B978-0-12-375678-7.00508-3.

[48] Rogowski W, Payne K, Schnell-Inderst P, Manca A, Rochau U, Jahn B, et al. Concepts of 'personalization' in personalized medicine: implications for economic evaluation. Pharmacoeconomics 2015;33:49–59. doi:10.1007/s40273-014-0211-5.

[49] Helgesson G, Ernstsson O, Åström M, Burström K. Whom should we ask? a systematic literature review of the arguments regarding the most accurate source of information for valuation of health states. Qual Life Res 2020. doi:10.1007/s11136-020-02426-4.

[50] Arrow KJ. A difficulty in the concept of social welfare. J Polit Econ 1950;58:328–46.

[51] Gandjour A. Theoretical foundation of patient v. population preferences in calculating QALYs. Med Decis Mak 2010;30:E57-63. doi:10.1177/0272989X10370488.

[52] Claxton K, Paulden M, Gravell H, Brouwer W, Culyer AJ. Discounting and decision making in the economic evaluation of health-care technologies. Health Econ 2011;20:2–15. doi:10.1002/hec.1612.

[53] Birch S, Donaldson C. Valuing the benefits and costs of health care programmes: where's the 'extra' in extra-welfarism? Soc Sci Med 2003;56:1121–33.

[54] Mulhern B, Norman R, Street DJ, Viney R. One method, many methodological choices: a structured review of discrete-choice experiments for health state valuation. Pharmacoeconomics 2019;37:29–43. doi:10.1007/s40273-018-0714-6.

[55] Nord E. Quality-Adjusted Life-Years. In: Cuyler AJ, editor. Encyclopedia of Health Economics, vol. 3, San Diego, United States: Elsevier; 2014, p. 231–4. doi:10.1016/B978-0-12-375678-7.00510-1.

[56] Thokala P, Ochalek J, Leech AA, Tong T. Cost-effectiveness thresholds: the past, the present and the future. Pharmacoeconomics 2018;36:509–22. doi:10.1007/s40273-017-0606-1.

[57] Vallejo-Torres L, García-Lorenzo B, Serrano-Aguilar P. Estimating a cost-effectiveness threshold for the Spanish NHS. Health Econ 2018;27:746–61. doi:10.1002/hec.3633.

[58] Carr AJ. Measuring quality of life: are quality of life measures patient centred? BMJ 2001;322:1357–60. doi:10.1136/bmj.322.7298.1357.

[59] Sculpher M, Gafni A. Recognising diversity in public preferences: the use of preference sub-groups in cost-effectiveness analysis. Health Econ 2001;10:317–24. doi:10.1002/hec.592.

[60] U.S. Food and Drug Administration. Patient-focused drug development guidance public workshop-discussion document: incorporating clinical outcome assessments into endpoints for regulatory decision-making. Rockville, United States: FDA; 2019.

[61] Rashidi AA, Anis AH, Marra CA. Do visual analogue scale (VAS) derived standard gamble (SG) utilities agree with Health Utilities Index utilities? A

comparison of patient and community preferences for health status in rheumatoid arthritis patients. Health Qual Life Outcomes 2006;4:25. doi:10.1186/1477-7525-4-25.

[62] Pickard AS, Neary MP, Cella D. Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. Health Qual Life Outcomes 2007;5:70. doi:10.1186/1477-7525-5-70.

[63] Khan MA, Richardson J. Is the validity of cost utility analysis improved when utility is measured by an instrument with 'home-country' weights? Evidence from six western countries. Soc Indic Res 2019;145:1–15. doi:10.1007/s11205-019-02094-z.

[64] Karimi M, Brazier J, Paisley S. Are preferences over health states informed? Health Qual Life Outcomes 2017;15:105. doi:10.1186/s12955-017-0678-9.

[65] Round J. Once bitten twice shy: thinking carefully before adopting the EQ-5D-5L. Pharmacoeconomics 2018;36:641–3. doi:10.1007/s40273-018-0636-3.

[66] Kane M. Errors of measurement, theory, and public policy. Princeton, United States: Educational Testing Service; 2008.

[67] Manca A, Hawkins N, Sculpher MJ. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. Health Econ 2005;14:487–96. doi:10.1002/hec.944.

[68] Richardson J, Iezzi A, Khan MA, Chen G, Maxwell A. Measuring the sensitivity and construct validity of 6 utility instruments in 7 disease areas. Med Decis Mak 2016;36:147–59. doi:10.1177/0272989X15613522.

[69] Basu A, Meltzer D. Value of information on preference heterogeneity and individualized care. Med Decis Mak 2007;27:112–27. doi:10.1177/0272989X06297393.

[70] Khan MA, Richardson J. Variation in the apparent importance of health-related problems with the instrument used to measure patient welfare. Qual Life Res 2018;27:2885–96. doi:10.1007/s11136-018-1956-7.

[71] Fenwick E. Uncertainty in economic evaluation. In: Cuyler AJ, editor. Encyclopedia of Health Economics, vol. 1, San Diego, United States: Elsevier; 2014, p. 224–31. doi:10.1016/B978-0-12-375678-7.01419-X.

[72] Claxton K. The irrelevance of inference: a decision making approach to the stochastic evaluation of health care technologies. J Health Econ 1999;18:341–64.

[73] Claxton K. Value of information analysis. In: Cuyler AJ, editor. Encyclopedia of Health Economics, vol. 2, San Diego, United States: Elsevier; 2014, p. 53–60. doi:10.1016/B978-0-12-375678-7.01421-8.

[74] Espinoza MA, Sculpher MJ, Manca A, Basu A. Analysing heterogeneity to support decision making. In: Cuyler AJ, editor. Encyclopedia of Health Economics, vol. 1, San Diego, United States: Elsevier; 2014, p. 71–6. doi:10.1016/B978-0-12-375678-7.01420-6.

[75] Ramaekers BLT, Joore MA, Grutters JPC. How should we deal with patient heterogeneity in economic evaluation: A systematic review of national pharmacoeconomic guidelines. Value Health 2013;16:855–62. doi:10.1016/j.jval.2013.02.013.

[76] Grutters JP, Sculpher M, Briggs AH, Severens JL, Candel MJ, Stahl JE, et al. Acknowledging patient heterogeneity in economic evaluation: a systematic

literature review. Pharmacoeconomics 2013;31:111–23. doi:10.1007/s40273-012-0015-4.

[77] Sculpher M, Gafni A. Recognising diversity in public preferences: the use of preference sub-groups in cost-effectiveness analysis. Authors' reply. Health Econ 2002;11:653–4. doi:10.1002/hec.736.

[78] Espinoza MA, Manca A, Claxton K, Sculpher M. Social value and individual choice: the value of a choice-based decision-making process in a collectively funded health system. Health Econ 2018;27:e28–40. doi:10.1002/hec.3559.

[79] Johnston BC, Thorlund K, Schünemann HJ, Xie F, Murad MH, Montori VM, et al. Improving the interpretation of quality of life evidence in meta-analyses: the application of minimal important difference units. Health Qual Life Outcomes 2010;8:1–5. doi:10.1186/1477-7525-8-116.

[80] Jayadevappa R, Cook R, Chhatre S. Minimal important difference to infer changes in health-related quality of life—a systematic review. J Clin Epidemiol 2017;89:188–98. doi:10.1016/j.jclinepi.2017.06.009.

[81] Devlin NJ, Parkin D, Browne J. Patient-reported outcomes in the NHS: new methods for analysing and reporting EQ-5D data. Health Econ 2010;19:886–905. doi:10.1002/hec.1608.

[82] Feinstein AR. Indexes of contrast and quantitative significance for comparisons of two groups. Stat Med 1999;18:2557–81.

[83] Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:395–407.

[84] Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003;41:582–92. doi:10.1097/01.MLR.0000062554.74615.4C.

[85] Coon CD, Cook KF. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. Qual Life Res 2018;27:33–40. doi:10.1007/s11136-017-1616-3.

[86] Luo N, Johnson JA, Coons SJ. Using instrument-defined health state transitions to estimate minimally important differences for four preference-based health-related quality of life instruments. Med Care 2010;48:365–71.

[87] Ramos-Goñi JM, Pinto-Prades JL, Oppe M, Cabasés JM, Serrano-Aguilar P, Rivero-Arias O. Valuation and modeling of EQ-5D-5L health states using a hybrid approach. Med Care 2017;55:e51–8. doi:10.1097/MLR.0000000000000283.

[88] Coretti S, Ruggeri M, McNamee P. The minimum clinically important difference for EQ-5D index: a critical review. Expert Rev Pharmacoecon Outcomes Res 2014;14:221–33. doi:10.1586/14737167.2014.894462.

[89] 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;1:1–8.

[90] Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI ®): concepts, measurement properties and applications. Health Qual Life Outcomes 2003;1:1–13.

[91] Shiroiwa T, Fukuda T, Ikeda S, Igarashi A, Noto S, Saito S, et al. Japanese population norms for preference-based measures: EQ-5D-3L, EQ-5D-5L, and SF-6D. Qual Life Res 2016;25:707–19. doi:10.1007/s11136-015-1108-2.

[92] Luo N, Liu G, Li M, Rand-Hendriksen K. Estimating a Chinese EQ-5D-5L value set using nonlinear models. Society for Medical Decision Making Proceedings of the 38th Annual North American Meeting, Vancouver, Canada: 2016.

[93] Ikeda S, Shiroiwa T, Igarashi A, Shinichi N, Fukuda T, Saito S, et al. Developing a Japanese version of the EQ-5D-5L value set. J Natl Inst Public Health 2015;64:47–55.

[94] Devlin NJ, Shah KK, Feng Y. Valuing health-related quality of life: an EQ-5D-5L value set for England. London, United Kingdom: Office of Health Economics; 2016.

[95] Augustovski F, Rey-Ares L, Irazola V, Garay OU, Gianneo O, Fernández G, et al. An EQ-5D-5L value set based on Uruguayan population preferences. Qual Life Res 2016;25:323–33. doi:10.1007/s11136-015-1146-9.

[96] Nolan CM, Longworth L, Lord J, Canavan JL, Jones SE, Kon SSC, et al. The EQ-5D-5L health status questionnaire in COPD: validity, responsiveness and minimum important difference. Thorax 2016;71:493–500. doi:10.1136/thoraxjnl-2015-207782.

[97] World Health Organization. Global report on diabetes. Geneva, Switzerland: WHO; 2016.

[98] Public Health Agency of Canada. Diabetes in Canada: facts and figures from a public health perspective. Ottawa, Canada: PHAC; 2011. doi:HP35-25/2011E-PDF.

[99] Al Sayah F, Majumdar SR, Soprovich A, Wozniak L, Johnson ST, Qiu W, et al. The Alberta's caring for diabetes (ABCD) study: rationale, design and baseline characteristics of a prospective cohort of adults with type 2 diabetes. Can J Diabetes 2015;39:S113–9. doi:10.1016/j.jcjd.2015.05.005.

[100] Janssen MF, Lubetkin EI, Sekhobo JP, Pickard AS. The use of the EQ-5D preference-based health status measure in adults with type 2 diabetes mellitus. Diabet Med 2011;28:395–413. doi:10.1111/j.1464-5491.2010.03136.x.

[101] Gutacker N, Street A. Use of large-scale HRQoL datasets to generate individualised predictions and inform patients about the likely benefit of surgery. Qual Life Res 2017;26:2497–505. doi:10.1007/s11136-017-1599-0.

[102] Fleishman JA, Selim AJ, Kazis LE. Deriving SF-12v2 physical and mental health summary scores: a comparison of different scoring algorithms. Qual Life Res 2010;19:231–41. doi:10.1007/s11136-009-9582-z.

[103] McClure NS, Sayah F Al, Xie F, Luo N, Johnson JA. Instrument-defined estimates of the minimally important difference for EQ-5D-5L index scores. Value Health 2017;20:644–50. doi:10.1016/j.jval.2016.11.015.

[104] Yost KJ, Eton DT, Garcia SF, Cella D. Minimally important differences were estimated for six PROMIS-Cancer scales in advanced-stage cancer patients. J Clin Epidemiol 2011;64:507–16. doi:doi:10.1016/j.jclinepi.2010.11.018.

[105] Tabberer M, Brooks J, Wilcox T. A meta-analysis of four randomized clinical trials to confirm the reliability and responsiveness of the Shortness of Breath with Daily Activities (SOBDA) questionnaire in chronic obstructive pulmonary disease. Health Qual Life Outcomes 2015;13:177. doi:10.1186/s12955-015-0369-3.

[106] Cleveland RB, Cleveland WS, McRae JE, Terpenning I. STL: a seasonal-trend decomposition procedure based on loess. J Off Stat 1990;6:3–33.

[107] Manski CF. Reasonable patient care under uncertainty. Health Econ 2018;27:1397–421. doi:10.1002/hec.3803.

[108] Wyrwich K, Senn S, Herdman M, Whitehurst D, Johnson JA. Workshop on: legitmacy, estimation, and uses of the minimal important difference (MID) with EQ-5D. EuroQol Academy Meeting, Noordwijk: Office of Health Economics; 2017.

[109] Briggs A, Pickard A, Lloyd A. Issue panel: minimal clinically important difference in EQ-5D: we can calculate it - but does that mean we should? International Society for Pharmacoeconomics and Outcomes Research 22nd Annual Meeting, Boston: ISPOR; 2017.

[110] Whitehurst DG, Bryan S. Trial-based clinical and economic analyses: the unhelpful quest for conformity. Trials 2013;14:421. doi:10.1186/1745-6215-14-421.

[111] Dirksen CD. The use of research evidence on patient preferences in health care decision-making: issues, controversies and moving forward. Expert Rev Pharmacoeconomics Outcomes Res 2014;14:785–94. doi:10.1586/14737167.2014.948852.

[112] EQ-5D. EuroQol Group Research Foundation 2020. http://www.euroqol.org/ (accessed April 14, 2020).

[113] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: The R Foundation for Statistical Computing; 2020.

[114] Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67:1–48. doi:10.18637/jss.v067.i01.

[115] Leeden R van der, Meijer E, Busing FMTA. Resampling multilevel models. In: Leeuw J de, Meijer E, editors. Handbook of Multilevel Analysis, New York, United States: Springer; 2007, p. 401–29.

[116] McClure NS, Sayah F Al, Ohinmaa A, Johnson JA. Minimally important difference of the EQ-5D-5L index score in adults with type 2 diabetes. Value Health 2018;21:1090–7. doi:10.1016/j.jval.2018.02.007.

[117] Ogorevc M, Murovec N, Fernandez NB, Rupel VP. Questioning the differences between general public vs. patient based preferences towards EQ-5D-5L defined hypothetical health states. Health Policy 2019;123:166–72. doi:10.1016/j.healthpol.2017.03.011.

[118] Rand-Hendriksen K, Augestad LA, Kristiansen IS, Stavem K. Comparison of hypothetical and experienced EQ-5D valuations: relative weights of the five dimensions. Qual Life Res 2012;21:1005–12. doi:10.1007/s11136-011-0016-3.

[119] Mann R, Brazier J, Tsuchiya A. A comparison of patient and general population weightings of EQ-5D dimensions. Health Econ 2009;18:363–72. doi:10.1002/hec.1362.

[120] Gandhi M, Tan RS, Ng R, Choo SP, Chia WK, Toh CK, et al. Comparison of health state values derived from patients and individuals from the general population. Qual Life Res 2017;26:3353–63. doi:10.1007/s11136-017-1683-5.

[121] Sullivan T, Hansen P, Ombler F, Derrett S, Devlin N. A new tool for creating personal and social EQ-5D-5L value sets, including valuing 'dead.' Soc Sci Med 2020;246:112707. doi:10.1016/j.socscimed.2019.112707.

[122] Craig BM, Busschbach JJ V, Salomon JA. Keep it simple: ranking health states

yields values similar to cardinal measurement approaches. J Clin Epidemiol 2009;62:296–305. doi:10.1016/j.jclinepi.2008.07.002.Keep.

[123] Craig BM, Busschbach JJV, Salomon JA. Modeling ranking, time trade-off, and visual analog scale values for EQ-5D health states: a review and comparison of methods. Med Care 2009;47:634–41. doi:10.1097/MLR.0b013e31819432ba.

[124] Purba FD, Hunfeld JAM, Timman R, Iskandarsyah A, Fitriana TS, Sadarjoen SS, et al. Test-retest reliability of EQ-5D-5L valuation techniques: the composite time trade-off and discrete choice experiments. Value Health 2018;21:1243–9. doi:10.1016/j.jval.2018.02.003.

[125] Taylor M, Chilton S, Ronaldson S, Metcalf H, Nielsen JS. Comparing increments in utility of health: an individual-based approach. Value Health 2017;20:224–9. doi:10.1016/j.jval.2016.12.009.

[126] Parkin D, Appleby J, Maynard A. Economics: the biggest fraud ever perpetrated on the world? Lancet 2013;382:11–5. doi:10.1016/S0140-6736(13)61178-2.

[127] Johnson JA, Al Sayah F, Wozniak L, Rees S, Soprovich A, Chik CL, et al. Controlled trial of a collaborative primary care team model for patients with diabetes and depression: rationale and design for a comprehensive evaluation. BMC Health Serv Res 2012;12. doi:10.1186/1472-6963-12-258.

[128] Johnson JA, Al Sayah F, Wozniak L, Rees S, Soprovich A, Qiu W, et al. Collaborative care versus screening and follow-up for patients with diabetes and depressive symptoms: results of a primary care-based comparative effectiveness trial. Diabetes Care 2014;37:3220–6. doi:10.2337/dc14-1308.

[129] Johnson JA, Lier DA, Soprovich A, Sayah F Al, Qiu W, Majumdar SR. Cost-effectiveness evaluation of collaborative care for diabetes and depression in primary care. Am J Prev Med 2016;51:e13–20. doi:10.1016/j.amepre.2016.01.010.

[130] Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology 2009;20:488–95. doi:10.1097/EDE.0b013e3181a819a1.

[131] Mills TC. Testing for stability in regression models. In: Patterson K, Mills TC, editors. Analysing Economic Data A Concise Introduction, New York, United States: Palgrave Macmillan; 2014, p. 231–43.

[132] Hunter RM, Baio G, Butt T, Morris S, Round J, Freemantle N. An educational review of the statistical issues in analysing utility data for cost-utility analysis. Pharmacoeconomics 2015;33:355–66. doi:10.1007/s40273-014-0247-6.

[133] Pearl J, Mackenzie D. From buccaneers to guinea pigs: the genesis of causal inference. In: The Book of Why: The New Science of Cause and Effect, New York, United States: Basic Books; 2018.

[134] Korn EL, Othus M, Chen T, Freidlin B. Assessing treatment efficacy in the subset of responders in a randomized clinical trial. Ann Oncol Off J Eur Soc Med Oncol 2017;28:1640–7. doi:10.1093/annonc/mdx197.

[135] European Medicines Agency. Committee for medicinal products for human use (CHMP) - guideline on the evaluation of anticancer medicinal products in man. London, United Kingdom: EMA; 2017.

[136] van Breukelen GJP. ANCOVA versus change from baseline in nonrandomized studies: The difference. Multivariate Behav Res 2013;48:895–922. doi:10.1080/00273171.2013.831743.

[137] U.S. Food and Drug Administration. Draft guidance - patient-focused drug development: collecting comprehensive and representative input guidance for industry, Food and Drug Administration staff, and other stakeholders. Maryland, United States: FDA; 2018.

[138] Drummond M. Introducing economic and quality of life measurements into clinical studies. Ann Med 2001;33:344–9. doi:10.3109/07853890109002088.

[139] Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality of life research: how meaningful is it? Pharmacoeconomics 2000;18:419–23.

[140] Krabbe PFM, van Asselt ADI, Selivanova A, Jabrayilov R, Vermeulen KM. Patient-centered item selection for a new preference-based generic health status instrument: CS-Base. Value Health 2019;22:467–73. doi:10.1016/j.jval.2018.12.006.

[141] Disher T, Beaubien L, Campbell-Yeo M. Are guidelines for measurement of quality of life contrary to patient-centred care? J Adv Nurs 2018;74:2677–84. doi:10.1111/jan.13820.

[142] Weinfurt KP. Viewing assessments of patient-reported heath status as conversations: implications for developing and evaluating patient-reported outcome measures. Qual Life Res 2019;28:3395–401. doi:10.1007/s11136-019-02285-8.

[143] Sav A, King MA, Whitty JA, Kendall E, Mcmillan SS, Kelly F, et al. Burden of treatment for chronic illness: A concept analysis and review of the literature. Health Expect 2015;18:312–24. doi:10.1111/hex.12046.

[144] Emanuel EJ, Onwuteaka-Philipsen BD, Urwin JW, Cohen J. Attitudes and practices of euthanasia and physician-assisted suicide in the United States, Canada, and Europe. JAMA 2016;316:79–90. doi:10.1001/jama.2016.8499.

[145] Brazier J, Tsuchiya A. Improving cross-sector comparisons: going beyond the health-related QALY. Appl Health Econ Health Policy 2015;13:557–65. doi:10.1007/s40258-015-0194-1.

[146] Harvard S, Werker G, Silva D. Social, ethical, and other value judgments in health economics modelling. Soc Sci Med 2020;253:112975. doi:10.1016/j.socscimed.2020.112975.

[147] McLeod LD, Cappelleri JC, Hays RD. Best (but oft-forgotten) practices: expressing and interpreting associations and effect sizes in clinical outcome assessments. Am J Clin Nutr 2016;103:685–93. doi:10.3945/ajcn.115.120378.

[148] Zhang Y, Coello PA, Guyatt GH, Yepes-Nuñez JJ, Akl EA, Hazlewood G, et al. GRADE guidelines: 20. Assessing the certainty of evidence in the importance of outcomes or values and preferences—inconsistency, imprecision, and other domains. J Clin Epidemiol 2019;111:83–93. doi:10.1016/j.jclinepi.2018.05.011.

[149] Pearl J, Mackenzie D. The Book of Why: The New Science of Cause and Effect. New York, United States: Basic Books; 2018.

[150] Kunz M. On responder analyses when a continuous variable is dichotomized and measurement error is present. Biometrical J 2011;53:137–55. doi:10.1002/bimj.201000069.

[151] Senn S. The measurement of treatment effects. In: Statistical Issues in Drug Development, 2nd ed., Wiley; 2008, p. 113–31. doi:10.1002/9780470723586.ch8.

[152] Atkinson G, Williamson P, Batterham AM. Issues in the determination of ‘responders’ and ‘non-responders’ in physiological research. Exp Physiol 2019:1215–25. doi:10.1113/EP087712.

[153] Snapinn SM, Jiang Q. Responder analyses and the assessment of a clinically relevant treatment effect. Trials 2007;8:1–6. doi:10.1186/1745-6215-8-31.

[154] Lamu AN, Gamst-Klaussen T, Olsen JA. Preference weighting of health state values: what difference does it make, and why? Value Health 2017;20:451–7. doi:10.1016/j.jval.2016.10.002.

[155] Knott RJ, Black N, Hollingsworth B, Lorgelly PK. Response-scale heterogeneity in the EQ-5D. Health Econ 2017;26:387–94. doi:10.1002/hec.3313.

[156] Wisløff T, Hagen G, Hamidi V, Movik E, Klemp M, Olsen JA. Estimating QALY gains in applied studies: a review of cost-utility analyses published in 2010. Pharmacoeconomics 2014;32:367–75. doi:10.1007/s40273-014-0136-z.

[157] Lancsar E, Gu Y, Gyrd-Hansen D, Butler J, Ratcliffe J, Bulfone L, et al. The relative value of different QALY types. J Health Econ 2020:102303. doi:10.1016/j.jhealeco.2020.102303.

[158] Gu Y, Lancsar E, Ghijben P, Butler JRG, Donaldson C. Attributes and weights in health care priority setting: a systematic review of what counts and to what extent. Soc Sci Med 2015;146:41–52. doi:10.1016/j.socscimed.2015.10.005.

[159] Ghijben P, Gu Y, Lancsar E, Zavarsek S. Revealed and stated preferences of decision makers for priority setting in health technology assessment: a systematic review. Pharmacoeconomics 2018;36:323–40. doi:10.1007/s40273-017-0586-1.

[160] Paulden M, Stafinski T, Menon D, McCabe C. Value-based reimbursement decisions for orphan drugs: a scoping review and decision framework. Pharmacoeconomics 2015;33:255–69. doi:10.1007/s40273-014-0235-x.

[161] Lipman SA, Brouwer WBF, Attema AE. A QALY loss is a QALY loss is a QALY loss: a note on independence of loss aversion from health states. Eur J Health Econ 2018;0:0. doi:10.1007/s10198-018-1008-9.

[162] Round J, Paulden M. Incorporating equity in economic evaluations: a multi-attribute equity state approach. Eur J Health Econ 2018;19:489–98. doi:10.1007/s10198-017-0897-3.

[163] Devlin NJ, Sussex J. Incorporating multiple criteria in HTA: methods and processes. London, United Kingdom: Office of Health Economics; 2011.

Appendix

Items wording of selected measures: SF-12 (A1 to A7), EQ-5D-5L (A8), PAID5 (A9), and PHQ8 (A10).

A1. In general, would you say your health is:

1 □ Excellent **2** □ Very Good **3** □ Good **4** □ Fair **5** □ Poor

A2. The following two questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

	Yes, limited a lot	Yes, limited a little	No, not limited at all
A2A. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	**1** □	**2** □	**3** □
A2B. Climbing several flights of stairs	**1** □	**2** □	**3** □

A3. During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of your physical health?

	All of the time	Most of the time	Some of the time	Little of the time	None of the time
A3A. Accomplished less than you would like	**1** □	**2** □	**3** □	**4** □	**5** □
A3B. Were limited in the kind of work or other activities	**1** □	**2** □	**3** □	**4** □	**5** □

A4. During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)

	All of the time	Most of the time	Some of the time	Little of the time	None of the time
A4A. Accomplished less than you would like	**1** □	**2** □	**3** □	**4** □	**5** □
A4B. Did work or other activities less carefully than usual	**1** □	**2** □	**3** □	**4** □	**5** □

A5. During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?

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

A6. The following three questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past 4 weeks...

	All of the time	Most of the time	Some of the time	Little of the time	None of the time
A6A. Have you felt calm and peaceful?	**1** □	**2** □	**3** □	**4** □	**5** □
A6B. Did you have a lot of energy?	**1** □	**2** □	**3** □	**4** □	**5** □

A6C. Have you felt downhearted and depressed?	1 □	2 □	3 □	4 □	5 □

A7. During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting friends, relatives, etc.)?

1 □ All of the time
2 □ Most of the time
3 □ Some of the time
4 □ Little of the time
5 □ None of the time

A8. Indicate which statement best describes your own health status today

A8A. Mobility:

1 □ I have no problems walking
2 □ I have slight problems walking
3 □ I have moderate problems walking
4 □ I have severe problems walking
5 □ I am unable to walk

A8B. Self-Care:

1 □ I have no problems washing or dressing myself
2 □ I have slight problems washing or dressing myself
3 □ I have moderate problems washing or dressing myself
4 □ I have severe problems washing or dressing myself
5 □ I am unable to wash or dress myself

A8C. Usual Activities (e.g. work, study, housework, family or leisure activities):

1 □ I have no problems doing my usual activities
2 □ I have slight problems doing my usual activities
3 □ I have moderate problems doing my usual activities
4 □ I have severe problems doing my usual activities
5 □ I am unable to do my usual activities

A8D. Pain or Discomfort:

1 □ I have no pain or discomfort
2 □ I have slight pain or discomfort
3 □ I have moderate pain or discomfort
4 □ I have severe pain or discomfort
5 □ I have extreme pain or discomfort

A8E. Anxiety/Depression:

1 □ I am not anxious or depressed
2 □ I am slightly anxious or depressed
3 □ I am moderately anxious or depressed
4 □ I am severely anxious or depressed
5 □ I am extremely anxious or depressed

A9. To what extent are the following diabetes issues currently problems for you?
(***Circle*** *the number that applies to you).*

	Not a problem	Minor problem	Moderate problem	Somewhat serious problem	Serious problem
A9A. Feeling scared when you think about living with diabetes	**0**	**1**	**2**	**3**	**4**
A9B. Feeling depressed when think about living with diabetes	**0**	**1**	**2**	**3**	**4**
A9C. Worrying about the future and the possibility of serious complications	**0**	**1**	**2**	**3**	**4**
A9D. Feeling that diabetes is taking up too much of your mental and physical energy every day	**0**	**1**	**2**	**3**	**4**
A9E. Coping with complications of diabetes	**0**	**1**	**2**	**3**	**4**

A10. Over the last 2 weeks, how often have you been bothered by any of the following problems? (***Circle*** *the number that applies to you*).

	Not at all	Several days	More than half the days	Nearly everyday
A10A. Little interest or pleasure in doing things	**0**	**1**	**2**	**3**
A10B. Feeling down, depressed, or hopeless	**0**	**1**	**2**	**3**
A10C. Trouble falling or staying asleep, or sleeping too much	**0**	**1**	**2**	**3**
A10D. Feeling tired or having little energy	**0**	**1**	**2**	**3**
A10E. Poor appetite or overeating	**0**	**1**	**2**	**3**
A10F. Feeling bad about yourself, or that you are a failure, or have let yourself or your family down	**0**	**1**	**2**	**3**
A10G. Trouble concentrating on things, such as reading the newspaper or watching television	**0**	**1**	**2**	**3**
A10H. Moving or speaking so slowly that other people could have noticed. Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual	**0**	**1**	**2**	**3**

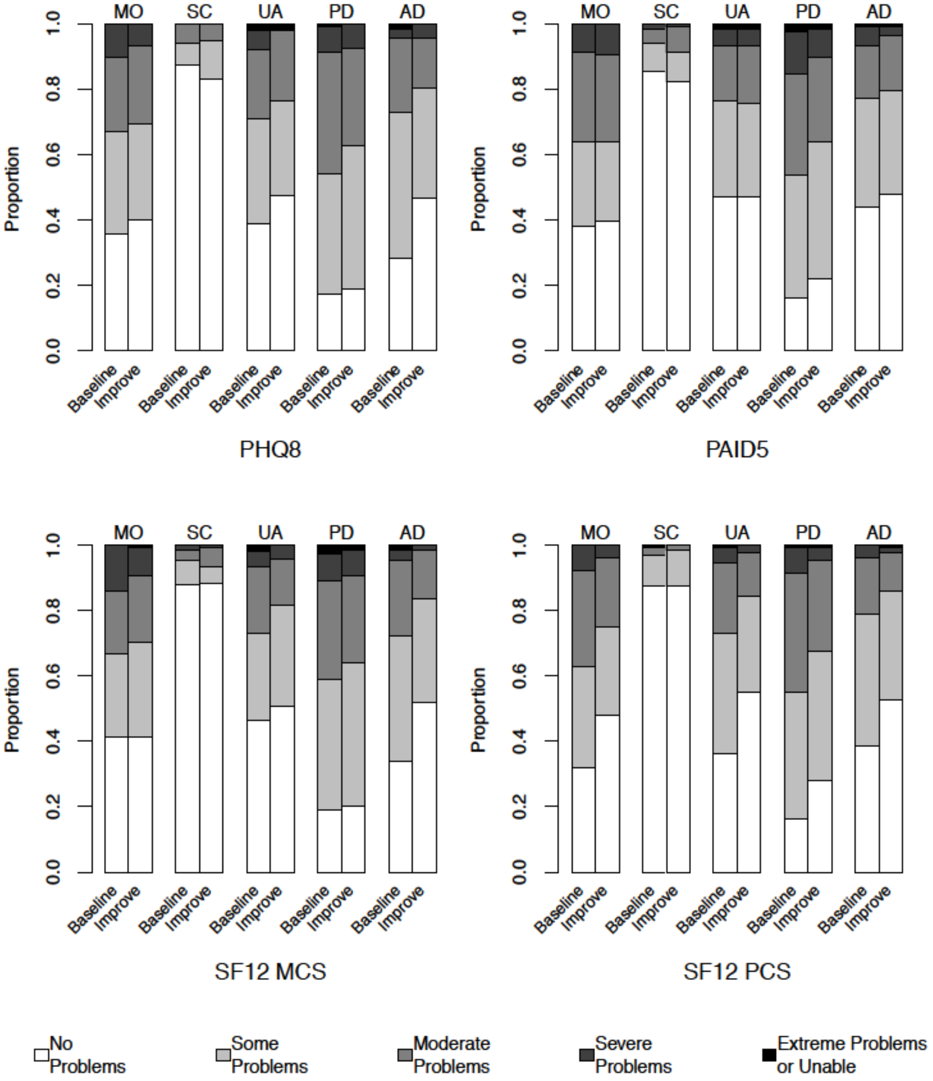
Table A1. Paretian classification of changes in EQ-5D-5L health state according to change in anchor score.

Anchor	Change in anchor score	Change in EQ-5D-5L health state					
		n	worse	better	mixed	same	11111
PHQ8	large det	91	57%	8%	32%	2%	1%
	large imp	77	5%	60%	32%	1%	1%
	zero change	341	24%	28%	11%	14%	23%
	no det	339	39%	21%	19%	15%	6%
	no imp	279	24%	36%	15%	14%	10%
	small det	164	55%	16%	18%	9%	2%
	small imp	137	17%	45%	23%	12%	3%
PAID5	large det	94	44%	16%	28%	11%	2%
	large imp	98	12%	56%	22%	4%	5%
	zero change	380	30%	23%	11%	17%	19%
	no det	269	37%	25%	18%	14%	6%
	no imp	325	31%	33%	19%	10%	6%
	small det	130	41%	16%	25%	11%	7%
	small imp	136	21%	38%	24%	11%	7%
SF-12 MCS	large det	128	66%	6%	19%	4%	5%
	large imp	95	6%	54%	26%	3%	11%
	zero change	0	0%	0%	0%	0%	0%
	no det	367	40%	21%	14%	15%	10%
	no imp	380	19%	36%	14%	16%	15%
	small det	170	45%	15%	21%	12%	6%
	small imp	148	14%	47%	26%	10%	3%
SF-12 PCS	large det	72	71%	4%	17%	7%	1%
	large imp	53	6%	66%	23%	2%	4%
	zero change	0	0%	0%	0%	0%	0%
	no det	455	36%	21%	17%	14%	12%
	no imp	414	21%	34%	16%	16%	14%
	small det	165	58%	11%	19%	6%	6%
	small imp	129	9%	57%	21%	9%	4%

EQ-5D-5L, EuroQol five-dimensional five-level questionnaire; PHQ8, patient health questionnaire 8 items; PAID5, problem areas in diabetes 5 items; SF-12, short-form medical survey 12 item; MCS, mental health component score; PCS, physical component score; det, deterioration; imp, improvement. Note changes in anchor score are: large (>1 SD), small (≥½ and ≤1 SD), zero change (0), no det/imp (<½ SD); n is the sample size of respondents included in each category; changes in EQ-5D-5L health states are based on Paretian classification: worse, change to a worse level of problems in at least one dimension and no improvement in any other dimension; better, change to a better level of problems in at least one dimension and no worsening in any other dimension; mixed, change to a worse (or better) level of problems in at least one dimension and the opposite change in at least one other dimension; same, no change in health state (excluding

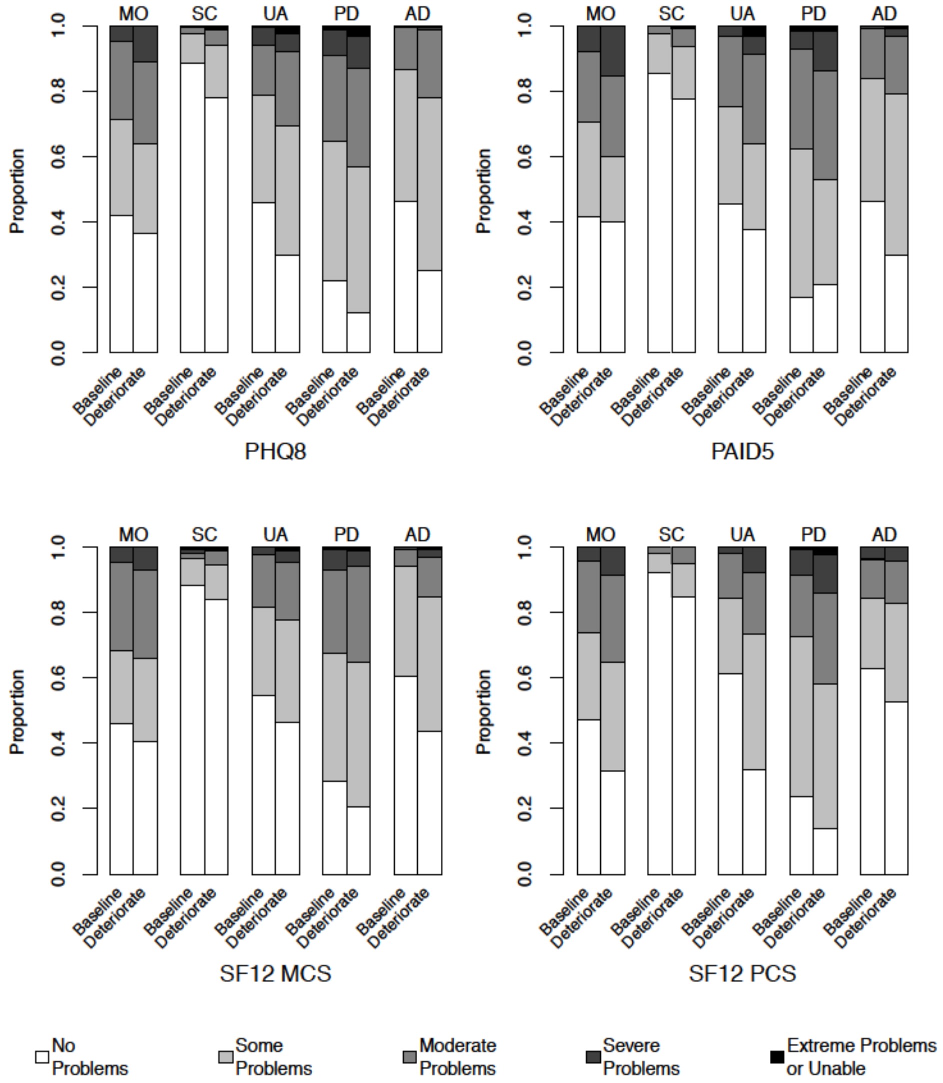
11111); 11111, unchanging health state with no problems in all dimensions.

Figure A1. Distribution of EQ-5D-5L health states for the small improve change group by dimension and level comparing baseline classification to classification at follow-up.



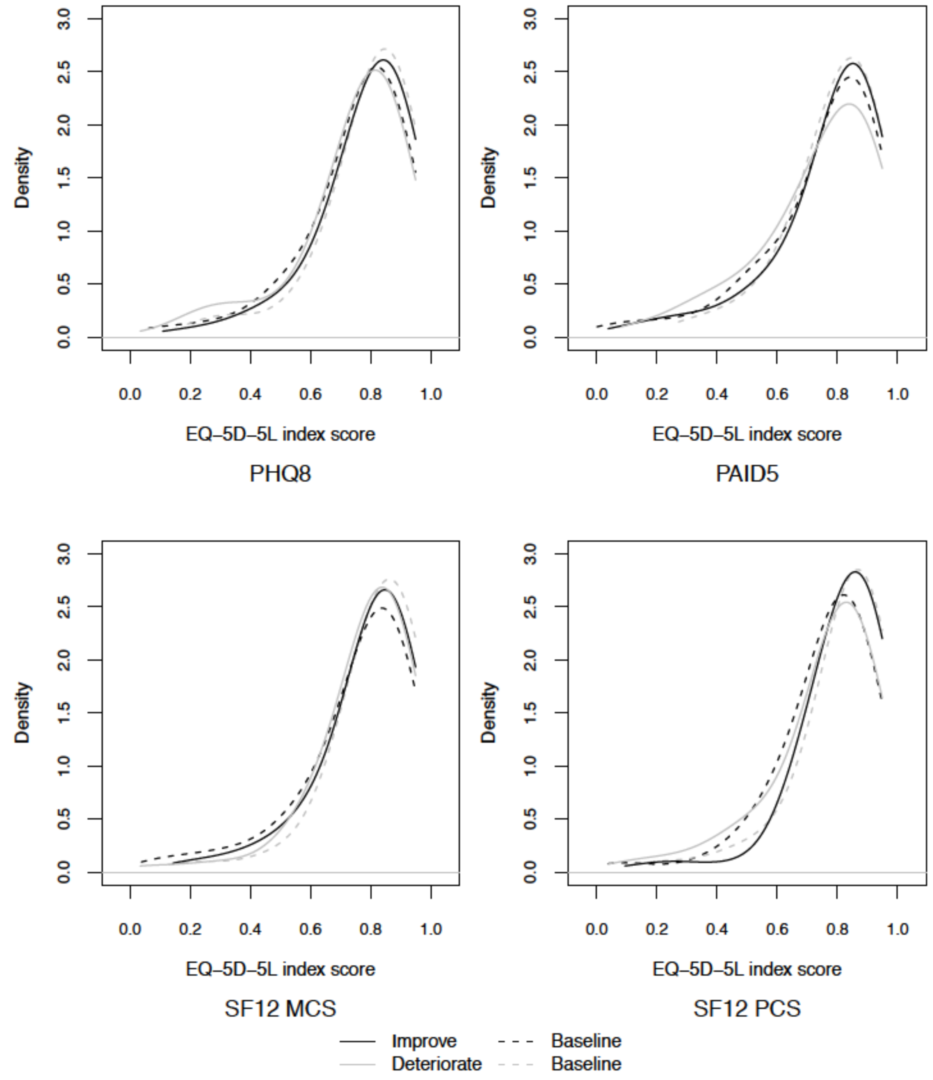
EQ-5D-5L, EuroQol five-dimensional five-level questionnaire; PHQ8, patient health questionnaire 8 items; PAID5, problem areas in diabetes 5 items; SF-12, short-form medical survey 12 item; MCS, mental health component score; PCS, physical component score; MO, mobility; SC, self-care; UA, usual activities; PD, pain/discomfort; AD, anxiety depression.

Figure A2. Distribution of EQ-5D-5L health states for the small deteriorate change group by dimension and level comparing baseline classification to classification at follow-up.



EQ-5D-5L, EuroQol five-dimensional five-level questionnaire; PHQ8, patient health questionnaire 8 items; PAID5, problem areas in diabetes 5 items; SF-12, short-form medical survey 12 item; MCS, mental health component score; PCS, physical component score; MO, mobility; SC, self-care; UA, usual activities; PD, pain/discomfort; AD, anxiety depression.

Figure A3. Distribution of EQ-5D-5L index score for the small change group by direction of change (improve versus deteriorate) comparing baseline score (dashed-line) to score at follow-up (solid lines).



EQ-5D-5L, EuroQol five-dimensional five-level questionnaire; PHQ8, patient health questionnaire 8 items; PAID5, problem areas in diabetes 5 items; SF-12, short-form medical survey 12 item; MCS, mental health component score; PCS, physical component score.

Table A2. Minimally important difference estimates of the EQ-5D-5L index score for patient subgroups by estimation method and direction of change.

Subgroup	Direction of change	Method	Sample size	Mean^{+}	SD^{+}	MID	95% CI	ES	SRM
Duration <10 years	All	Instrument-defined:							
		idMID	687	0.814	0.155	0.047	0.046–0.048	0.301	0.448
		idMID*	687	---	---	0.037	0.037–0.037	0.240	0.358
		Anchor-based:							
		PHQ8	129	0.012	0.086	0.039	0.018–0.061	0.250	0.373
		PAID5	106	-0.001	0.096	0.032	0.008–0.056	0.204	0.303
		MCS	147	0.015	0.087	0.031	0.018–0.046	0.201	0.300
		PCS	116	0.016	0.085	0.060	0.042–0.077	0.386	0.575
		Pooled	---	0.009	0.087	0.040	0.037–0.044	0.260	0.388
	Improve	Instrument-defined:							
		idMID	687	---	---	0.042	0.041–0.043	0.272	0.405
		idMID*	687	---	---	0.038	0.038–0.038	0.245	0.365
		Anchor-based:							
		PHQ8	53	0.004	0.081	0.018	-0.015–0.054	0.116	0.173
		PAID5	52	-0.009	0.094	0.016	-0.009–0.042	0.104	0.155
		MCS	73	0.012	0.091	0.029	0.010–0.050	0.190	0.283
		PCS	48	0.014	0.075	0.064	0.036–0.096	0.412	0.613
		Pooled	---	0.000	0.083	0.032	0.005–0.060	0.205	0.306
	Deteriorate	Instrument-defined:							
		idMID	687	---	---	0.051	0.049–0.052	0.326	0.486
		idMID*	687	---	---	0.037	0.037–0.038	0.241	0.360
		Anchor-based:							
		PHQ8	76	0.012	0.083	0.053	0.029–0.081	0.344	0.512
		PAID5	54	0.012	0.089	0.046	0.008–0.088	0.300	0.446
		MCS	74	0.019	0.083	0.033	0.014–0.052	0.213	0.317
		PCS	68	0.018	0.093	0.057	0.036–0.078	0.368	0.549
		Pooled	---	0.019	0.084	0.047	0. 022–0.075	0.306	0.456

Duration >=10 years	All	Instrument-defined:							
		idMID	773	0.783	0.167	0.050	0.049–0.051	0.296	0.416
		idMID*	773	---	---	0.037	0.037–0.037	0.222	0.312
		Anchor-based:							
		PHQ8	148	0.016	0.089	0.044	0.028–0.062	0.265	0.373
		PAID5	144	-0.002	0.099	0.044	0.024–0.064	0.263	0.370
		MCS	152	0.017	0.083	0.036	0.018–0.054	0.213	0.300
		PCS	160	0.018	0.086	0.064	0.048–0.079	0.380	0.534
		Pooled	---	0.012	0.089	0.047	0.046–0.048	0.280	0.394
	Improve	Instrument-defined:							
		idMID	773	---	---	0.043	0.042–0.044	0.258	0.363
		idMID*	773	---	---	0.038	0.037–0.038	0.225	0.316
		Anchor-based:							
		PHQ8	72	0.009	0.082	0.038	0.014–0.061	0.227	0.319
		PAID5	72	-0.011	0.092	0.029	0.001–0.056	0.175	0.246
		MCS	68	0.017	0.081	0.037	0.010–0.062	0.223	0.313
		PCS	73	0.022	0.087	0.055	0.032–0.077	0.326	0.458
		Pooled	---	0.009	0.086	0.040	0.014–0.064	0.237	0.334
	Deteriorate	Instrument-defined:							
		idMID	773	---	---	0.055	0.053–0.057	0.328	0.461
		idMID*	773	---	---	0.038	0.037–0.038	0.225	0.317
		Anchor-based:							
		PHQ8	76	0.012	0.088	0.051	0.021–0.079	0.302	0.425
		PAID5	72	0.016	0.096	0.059	0.034–0.086	0.352	0.495
		MCS	84	0.016	0.084	0.034	0.011–0.060	0.206	0.289
		PCS	87	0.015	0.085	0.071	0.050–0.094	0.425	0.598
		Pooled	---	0.015	0.089	0.054	0. 029–0.080	0.321	0.452
Age<65 years	All	Instrument-defined:							
		idMID	969	0.794	0.173	0.048	0.047–0.049	0.277	0.416
		idMID*	969	---	---	0.037	0.037–0.037	0.215	0.323
		Anchor-based:							

		PHQ8	166	0.011	0.085	0.049	0.030–0.068	0.280	0.421
		PAID5	147	0.003	0.099	0.028	0.006–0.046	0.160	0.240
		MCS	162	0.014	0.088	0.031	0.017–0.045	0.178	0.268
		PCS	156	0.014	0.083	0.066	0.052–0.081	0.380	0.571
		Pooled	---	0.010	0.089	0.043	0.040–0.046	0.249	0.375
	Improve	Instrument-defined:							
		idMID	969	---	---	0.043	0.042–0.044	0.250	0.376
		idMID*	969	---	---	0.038	0.038–0.038	0.218	0.328
		Anchor-based:							
		PHQ8	79	0.003	0.085	0.037	0.013–0.062	0.212	0.318
		PAID5	73	-0.005	0.095	0.021	-0.013–0.050	0.120	0.180
		MCS	87	0.013	0.098	0.037	0.018–0.057	0.211	0.317
		PCS	73	0.012	0.069	0.066	0.046–0.088	0.382	0.575
		Pooled	---	0.006	0.087	0.040	0.016–0.065	0.231	0.347
	Deteriorate	Instrument-defined:							
		idMID	969	---	---	0.052	0.051–0.053	0.301	0.452
		idMID*	969	---	---	0.038	0.037–0.038	0.217	0.326
		Anchor-based:							
		PHQ8	87	0.013	0.076	0.059	0.032–0.088	0.343	0.515
		PAID5	74	0.010	0.090	0.034	0.010–0.059	0.199	0.299
		MCS	75	0.015	0.077	0.024	0.003–0.045	0.140	0.210
		PCS	83	0.016	0.094	0.065	0.045–0.087	0.377	0.567
		Pooled	---	0.013	0.084	0.046	0. 023–0.070	0.265	0.398
Age>=65 years	All	Instrument-defined:							
		idMID	906	0.788	0.164	0.049	0.048–0.050	0.301	0.453
		idMID*	906	---	---	0.037	0.037–0.037	0.226	0.340
		Anchor-based:							
		PHQ8	134	0.013	0.088	0.035	0.019–0.053	0.215	0.324
		PAID5	118	-0.005	0.093	0.049	0.028–0.074	0.298	0.449
		MCS	155	0.017	0.086	0.032	0.015–0.049	0.193	0.290
		PCS	137	0.019	0.087	0.047	0.029–0.065	0.287	0.433

Age>=65 years		Pooled	---	0.011	0.089	0.041	0.040–0.043	0.248	0.374
	Improve	Instrument-defined:							
		idMID	906	---	---	0.043	0.042–0.045	0.264	0.398
		idMID*	906	---	---	0.037	0.037–0.038	0.226	0.341
		Anchor-based:							
		PHQ8	57	0.004	0.078	0.022	0.000–0.047	0.135	0.204
		PAID5	62	-0.013	0.092	0.020	-0.007–0.046	0.120	0.181
		MCS	60	0.010	0.083	0.028	-0.001–0.057	0.173	0.261
		PCS	56	0.017	0.082	0.037	0.005–0.066	0.226	0.341
		Pooled	---	0.005	0.084	0.027	-0.001–0.054	0.164	0.247
	Deteriorate	Instrument-defined:							
		idMID	906	---	---	0.055	0.053–0.056	0.332	0.500
		idMID*	906	---	---	0.038	0.037–0.038	0.229	0.345
		Anchor-based:							
		PHQ8	77	0.012	0.089	0.045	0.022–0.069	0.273	0.412
		PAID5	56	0.014	0.088	0.081	0.045–0.122	0.495	0.746
		MCS	95	0.024	0.090	0.034	0.011–0.056	0.205	0.309
		PCS	81	0.021	0.092	0.054	0.032–0.078	0.330	0.497
		Pooled	---	0.018	0.090	0.054	0.027–0.081	0.326	0.491
Number of comorbidities<4	All	Instrument-defined:							
		idMID	825	0.858	0.109	0.043	0.043–0.044	0.397	0.483
		idMID*	825	---	---	0.037	0.037–0.037	0.341	0.415
		Anchor-based:							
		PHQ8	101	0.011	0.070	0.035	0.017–0.056	0.321	0.390
		PAID5	97	-0.007	0.079	0.010	-0.011–0.031	0.089	0.109
		MCS	118	0.007	0.072	0.031	0.016–0.048	0.284	0.345
		PCS	101	0.017	0.071	0.057	0.040–0.075	0.517	0.629
		Pooled	---	0.007	0.073	0.033	0.032–0.036	0.303	0.368
	Improve	Instrument-defined:							
		idMID	825	---	---	0.041	0.040–0.042	0.377	0.458
		idMID*	825	---	---	0.039	0.038–0.039	0.353	0.429

Number of comorbidities<4		Anchor-based:							
		PHQ8	44	0.003	0.067	0.025	0.003–0.050	0.229	0.279
		PAID5	48	-0.011	0.078	0.000	-0.039–0.034	0.000	0.000
		MCS	58	0.002	0.074	0.041	0.018–0.066	0.376	0.457
		PCS	44	0.017	0.064	0.048	0.024–0.077	0.441	0.536
		Pooled	---	0.003	0.071	0.029	0.002–0.057	0.262	0.318
	Deteriorate	Instrument-defined:							
		idMID	825	---	---	0.046	0.045–0.047	0.421	0.511
		idMID*	825	---	---	0.037	0.037–0.037	0.341	0.414
		Anchor-based:							
		PHQ8	57	0.010	0.069	0.043	0.017–0.074	0.391	0.476
		PAID5	49	0.007	0.071	0.019	0.001–0.043	0.177	0.215
		MCS	60	0.011	0.071	0.021	0.005–0.038	0.194	0.236
		PCS	57	0.017	0.078	0.063	0.041–0.086	0.576	0.700
		Pooled	---	0.011	0.072	0.037	0. 016–0.060	0.335	0.407
Number of comorbidities>= 4	All	Instrument-defined:							
		idMID	1102	0.740	0.190	0.053	0.052–0.053	0.277	0.416
		idMID*	1102	---	---	0.037	0.037–0.037	0.195	0.293
		Anchor-based:							
		PHQ8	200	0.013	0.099	0.047	0.030–0.064	0.247	0.371
		PAID5	169	0.003	0.109	0.053	0.033–0.074	0.280	0.422
		MCS	200	0.023	0.097	0.032	0.018–0.046	0.167	0.251
		PCS	193	0.015	0.096	0.057	0.042–0.073	0.302	0.454
		Pooled	---	0.013	0.100	0.047	0.045–0.050	0.249	0.374
Number of comorbidities>= 4	Improve	Instrument-defined:							
		idMID	1102	---	---	0.045	0.044–0.046	0.235	0.353
		idMID*	1102	---	---	0.037	0.037–0.037	0.194	0.292
		Anchor-based:							
		PHQ8	93	0.004	0.094	0.034	0.011–0.059	0.181	0.272
		PAID5	88	-0.007	0.106	0.032	0.008–0.057	0.170	0.256
		MCS	90	0.019	0.101	0.029	0.006–0.051	0.153	0.230

		PCS	85	0.012	0.085	0.056	0.033–0.077	0.297	0.446
		Pooled	---	0.007	0.096	0.038	0.014–0.061	0.200	0.301
	Deteriorate	Instrument-defined:							
		idMID	1102	---	---	0.059	0.057–0.060	0.310	0.465
		idMID*	1102	---	---	0.038	0.038–0.038	0.199	0.299
		Anchor-based:							
		PHQ8	107	0.015	0.096	0.058	0.033–0.082	0.304	0.457
		PAID5	81	0.017	0.102	0.076	0.047–0.108	0.400	0.601
		MCS	110	0.026	0.093	0.034	0.013–0.056	0.179	0.269
		PCS	108	0.019	0.106	0.058	0.040–0.078	0.306	0.459
		Pooled	---	0.019	0.099	0.056	0.033–0. 081	0.297	0.447
Female	All	Instrument-defined:							
		idMID	866	0.777	0.178	0.050	0.049–0.051	0.280	0.428
		idMID*	866	---	---	0.037	0.037–0.037	0.208	0.319
		Anchor-based:							
		PHQ8	146	0.010	0.087	0.039	0.022–0.059	0.218	0.334
		PAID5	121	0.001	0.103	0.040	0.018–0.062	0.226	0.346
		MCS	139	0.017	0.089	0.025	0.009–0.040	0.141	0.216
		PCS	133	0.011	0.085	0.055	0.040–0.072	0.308	0.471
		Pooled	---	0.010	0.091	0.040	0.038–0.042	0.223	0.342
	Improve	Instrument-defined:							
		idMID	866	---	---	0.044	0.043–0.045	0.245	0.376
		idMID*	866	---	---	0.037	0.037–0.038	0.209	0.320
		Anchor-based:							
		PHQ8	62	0.000	0.082	0.030	0.003–0.057	0.166	0.253
		PAID5	61	-0.010	0.098	0.023	-0.012–0.056	0.128	0.196
		MCS	67	0.010	0.087	0.032	0.008–0.055	0.178	0.273
		PCS	58	0.014	0.083	0.041	0.017–0.062	0.229	0.351
Female		Pooled	---	0.004	0.088	0.031	0.004–0.058	0.175	0.268
	Deteriorate	Instrument-defined:							
		idMID	866	---	---	0.055	0.053–0.056	0.308	0.471

		idMID*	866	---	---	0.038	0.037–0.038	0.211	0.323
		Anchor-based:							
		PHQ8	84	0.017	0.083	0.046	0.022–0.073	0.257	0.393
		PAID5	60	0.021	0.096	0.058	0.034–0.086	0.325	0.498
		MCS	72	0.024	0.091	0.019	0.000–0.040	0.107	0.164
		PCS	75	0.009	0.086	0.066	0.042–0.090	0.368	0.563
		Pooled	---	0.018	0.089	0.047	0.025–0. 072	0.264	0.405
Male	All	Instrument-defined:							
		idMID	1048	0.802	0.162	0.048	0.047–0.048	0.294	0.439
		idMID*	1048	---	---	0.037	0.037–0.037	0.230	0.343
		Anchor-based:							
		PHQ8	155	0.013	0.086	0.047	0.029–0.066	0.289	0.431
		PAID5	144	-0.003	0.091	0.035	0.014–0.057	0.214	0.319
		MCS	178	0.014	0.085	0.036	0.020–0.053	0.222	0.332
		PCS	161	0.020	0.085	0.059	0.044–0.076	0.364	0.543
		Pooled	---	0.011	0.087	0.044	0.043–0.047	0.272	0.406
	Improve	Instrument-defined:							
		idMID	1048	---	---	0.043	0.042–0.044	0.266	0.396
		idMID*	1048	---	---	0.038	0.037–0.038	0.233	0.348
		Anchor-based:							
		PHQ8	75	0.006	0.081	0.033	0.010–0.056	0.203	0.303
		PAID5	74	-0.008	0.090	0.018	-0.009–0.047	0.113	0.168
		MCS	80	0.012	0.094	0.035	0.014–0.056	0.214	0.319
		PCS	71	0.014	0.070	0.064	0.039–0.089	0.396	0.590
		Pooled	---	0.006	0.083	0.037	0.014–0.062	0.231	0.345
Male	Deteriorate	Instrument-defined:							
		idMID	1048	---	---	0.052	0.051–0.053	0.321	0.479
		idMID*	1048	---	---	0.037	0.037–0.038	0.232	0.346
		Anchor-based:							
		PHQ8	80	0.009	0.083	0.060	0.032–0.088	0.369	0.550
		PAID5	70	0.005	0.081	0.052	0.015–0.090	0.320	0.477

		MCS	98	0.015	0.075	0.037	0.014–0.060	0.230	0.343
		PCS	90	0.027	0.099	0.055	0.034–0.077	0.339	0.506
		Pooled	---	0.014	0.085	0.051	0. 024–0.079	0.314	0.469

Patient subgroups were defined by diabetes duration (< 10 years or ≥ 10 years), number of comorbidities (< 4 or ≥ 4), age (< 65 years old or ≥ 65 years old), and sex (male or female). MID, minimally important difference; EQ-5D-5L, EuroQol five-dimensional five-level questionnaire; idMID, instrument-defined minimally important difference (*excluding maximum-valued scoring parameters); PHQ8, patient health questionnaire 8 items; PAID5, problem areas in diabetes 5 items; SF-12, short-form medical survey 12 item; MCS, mental health component score; PCS, physical component score; Pooled, average of anchor-based estimates; SD, standard deviation; CI, confidence interval based on 1000 bootstrap replicates; ES, effect size; SRM, standardized response mean. Note that values for direction of change "deteriorate" have been multiplied by negative one (-1); plus symbol [$^+$] represents a statistic for the baseline data set (Instrument-defined) or for the no change group (Anchor-based); the "no change" group by direction of change includes responses with anchor change scores between 0 and the corresponding limit of change (i.e., trivial improvement/deterioration); sample size is the number of respondent scores used in the calculation of the MID.