Towards a Pluralistic Conception of Value in Health: Theoretical Issues and Practical Tools

Paul Schneider

A thesis submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy

The University of Sheffield
Faculty of Medicine, Dentistry and Health
School of Health and Related Research

May 2023

Abstract

In this thesis, I present my work on theoretical and methodological issues related to the valuation of health (states) in the context of health technology assessment. It is written in publication format, comprising an introduction and literature review, followed by seven substantive chapters (written in journal article style), organised into two parts, and, finally, a discussion and conclusion. The first part explores the normative foundations for interpersonal utility comparisons and provides a critical examination of current practices, including the ubiquitous use of the arithmetic average to aggregate individual health preferences into social value sets, setting 'being dead' to a utility value of zero, and the assignment of negative utility values to health states considered 'worse than dead'. The second part reports on the development of a practical tool called Online elicitation of Personal Utility Functions (OPUF). It is a new type of online survey for creating value sets using compositional preference elicitation methods, which allow for the construction of value functions for small groups and even on the individual level. The thesis covers the design, pilot testing and application of OPUF to derive value sets from the general population in the UK and from patients with rheumatic diseases in Germany. Both parts of the thesis call into question current standard practices and develop new starting points for a more pluralistic approach to determine the value of health in a society, consisting of individuals with unique preferences, values, and perspectives.

Acknowledgments

I would like to thank my supervisors, Ben van Hout and John Brazier, for their encouragement, trust, and support, and for giving me the freedom to pursue many different avenues of thought, even if they disagreed with them, and Anju Keetharuth for stepping in at the last minute to help me with the final stages of my PhD project.

I am incredibly grateful to the many people who have helped me with my work and made it the pleasant and enjoyable experience it has been. In particular, I would like to thank Stefano Amorelli, Katharina Blankart, Jen Boyd, Nathan Bray, Alan Brennan, Jen-Yu (Amy) Chang, Jürgen Clausen, Richard Cookson, Siobhan Daley, Nancy Devlin, Barry Dewitt, Simon Dixon, Jack Dowie, Irene Ebyarimpa, Job van Exel, Naomi Gibbs, Wolfang Greiner, Nils Gutacker, Anthony Hatswell, Marieke Heisen, Berendice Herandez, Dan Howdon, Michal Jakuczyk, Paul Kind, Johanna Kokot, Artur Kowalski, Nick Latimer, James Leigh, Stefan Lipman, James Love-Koh, Daria Lucchesi, Kristina Ludwig, Ole Marten, Emily McDool, Simon McNamara, Dave Mott, Clara Mukuria, Dimitry Nohrin, Monica Oliveira, Phil Powell, Donna Rowen, Mandy Ryan, Chris Sampson, Erik Schokkaert, Koonal Shah, Nico Silva-Illanes, Rob Smith, Niall Stewart, Elly Stolk, Mark Strong, Bob Sugden, Tazeen Tahsina and Aki Tsuchiya, for helping, encouraging, challenging, inspiring, guiding and/or otherwise supporting my PhD journey.

Finally, I would like to express my gratitude to my family: Amalie, Levin Nepomuk, and Fridolin. Your patience, tolerance, and understanding have been beyond measure, and for that, and so much more, I am truly thankful.

This work was supported by the Wellcome Trust Doctoral Training Centre in Public Health Economics and Decision Science (108903/Z/19/Z) and by the University of Sheffield.

Acknowledgement of collaborative work

The candidate confirms the work submitted is their own, except where that work has formed part of jointly authored publications. In those circumstances, the contribution of other authors is explicitly indicated. The candidate confirms appropriate credit is given within the thesis where reference is made to the work of others.

Copyright and permissions

Research outputs

The following research outputs are based on the work presented in this thesis. Consistent with the alternative 'publication format' of this thesis, published and submitted manuscripts are included as chapters verbatim and without additional editing.

Publications:

Schneider P. The QALY is ableist: on the unethical implications of health states worse than dead. *Quality of Life Research.* 2022 May;31(5): 1545-52. https://doi.org/10.1007/s11136-021-03052-4 (Chapter 5).

Schneider P, van Hout B, Heisen M, Brazier J, Devlin N. The Online Elicitation of Personal Utility Functions (OPUF) tool: a new method for valuing health states. *Wellcome Open Research.* 2022 Jan 14;7: 14. https://doi.org/10.12688/wellcomeopenres.17518.1 (Chapter 6).

Schneider P. Social tariffs and democratic choice—Do population-based health state values reflect the will of the people? *Health Economics.* 2021 Jan;30(1): 104-12. https://doi.org/10.1002/hec.4179 (Chapter 2).

Manuscripts under review:

Schneider P, Brazier J, van Hout B, Devlin N. Not just another EQ-5D-5L value set for the UK: using the OPUF approach to study health preferences on the societal, group and individual level. *Health Economics* – resubmitted, under review (Chapter 7).

Schneider P, van Hout B, Brazier J. Fair interpersonal utility comparison in the valuation of health: a relative utilitarian preference aggregation method. *European Journal of Health Economics* – submitted (Chapter 3).

Schneider P, Blankart K, Brazier J, van Hout B, Devlin N. Using the Online Elicitation of Personal Utility Functions (OPUF) approach to derive a patient-based EQ-5D-5L value set: a study in 122 patients with rheumatic diseases from Germany. *Value in Health* – submitted (Chapter 8).

Data and code repositories:

Schneider P. *Source code of the EQ-5D-5L OPUF survey tool.* 2022. Available from: https://github.com/bitowaqr/opuf_demo.

Bray N, Tudor Edwards R, Schneider P. *Data and code repository for publication (Developing preference-based value sets for the MobQoL-7D: Practical application of the Online Elicitation of Personal Utility Functions (OPUF) tool).* 2022. https://github.com/bitowaqr/mobqol.

Schneider P, van Hout B, Brazier J. *Code and Data Repository for: Fair interpersonal utility comparison in the valuation of health: a relative utilitarian preference aggregation method.* 2021. https://github.com/bitowaqr/unpac.

Conference presentations, seminars, and invited talks:

Schneider P, van Hout B, Heisen M, Devlin N. *Online Elicitation of Personal Utility Functions: Understanding What Brings Value to the Individual.* [Workshop] May 2023. ISPOR. Boston, USA.

Schneider P. *OPUF – an open platform for eliciting health preferences and values.* [Virtual seminar] April 2023. Swedish HEOR Research Network. Virtual.

Schneider P. *Using the Online Elicitation of Personal Utility Functions (OPUF) approach to derive patient-based value sets.* [Presentation] March 2023. dggö Conference. Hannover, Germany.

Schneider P. *OPUF – an open platform for eliciting health preferences and values.* [Virtual seminar] February 2023. Lumanity, Knowledge Series Seminar. Virtual.

Schneider P, Blankart K, Brazier J, van Hout B, Devlin N. *Using the Online Elicitation of Personal Utility Functions (OPUF) approach to derive a patient-based EQ-5D-5L value set.* [Paper discussion by Mandy Ryan] January 2023. HESG. Manchester, UK.

Schneider P, van Hout B, Heisen M, Devlin N. *Personalized Values for Health: Introducing OPUF.* [Workshop] November 2022. ISPOR Europe. Vienna, Austria.

Schneider P. *Online elicitation of Personal Utility Functions (OPUF): an open health valuation platform.* [Presentation] November 2022. Society for Benefit-Cost Analysis (SBCA). Paris, France.

Schneider P. *The QALY is ableist.* [Invited virtual talk] November 2022. PenTAG Seminar. University of Exeter, Exeter, UK.

Schneider P. *Online elicitation of Personal Utility Functions (OPUF): a new method for eliciting patient preferences.* [Presentation] November 2022. EQ-London Meeting. London, UK.

Schneider P. *Online elicitation of Personal Utility Functions (OPUF): a new method for eliciting patient preferences.* [Plenary] October 2022. ISOQOL. Prague, Czech Republic.

Schneider P. *The Online elicitation of Personal Utility Functions (OPUF) Tool: an open health valuation platform* [Presentation] October 2022. EQ-PhD Network Seminar. Virtual.

Schneider P. *Online elicitation of Personal Utility Functions (OPUF) – an open, modular health valuation platform.* [Presentation] September 2022. International Association for Health preference Research (IAHPR). Berlin, Germany.

Schneider P. *Online elicitation of Personal Utility Functions (OPUF) – an open, modular health valuation platform.* [Presentation] July 2022. European Health Economics Association (EuHEA). Oslo, Norway.

Schneider P. *Online elicitation of Personal Utility Functions (OPUF) – an open, modular health valuation platform.* [Presentation] June 2022. HTAi. Utrecht, Netherlands.

Schneider P. *The QALY is ableist.* [Presentation] June 2022. Brocher Summer School. Geneva, Switzerland.

Schneider P, Brazier J, van Hout B, Devlin N. *Not just another value set for the UK: using the OPUF approach to study health preferences on the societal- group- and individual-level.* [Paper discussion by Anthony Hatswell] June 2022. HESG. Sheffield, UK.

Schneider P, Brazier J, van Hout B, Devlin N. *Not just another value set for the UK: using the OPUF approach to study health preferences on the societal- group- and individual-level.* [Paper discussion by Job van Exel] May 2022. lolaHESG. Maastricht, Netherlands.

Schneider P. *Online elicitation of Personal Utility Functions (OPUF) – an open, modular health valuation platform.* [Virtual poster] May 2022. ISPOR. Washington, USA.

Schneider P. *The Online elicitation of Personal Utility Functions (OPUF) Tool: a new method for valuing health states.* [Invited talk] May 2022. CINCH. University Duisburg-Essen, Germany.

Schneider P. *Online elicitation of Personal Utility Functions (OPUF) – an open, modular health valuation platform.* [Virtual round table] April 2022. ISPH Priorities. Bergen, Norway.

Schneider P, van Hout B, Brazier J, Devlin N. *Not just another value set for the UK: using the OPUF approach to study health preferences on the societal- group- and individual-level.* [Poster] April 2022. EuroQol Academy Meeting. Noordwijk, Netherlands.

Schneider P, van Hout B, Brazier J, Devlin N. *Not just another EQ-5D-5L value set for the UK: Using the OPUF approach to study preferences on the societal, group, and individual level.* [Presentation] March 2022. dggö Conference. Hamburg, Germany.

Schneider P. *The Online elicitation of Personal Utility Functions (OPUF) Tool: a new method for valuing health states.* [Virtual seminar] March 2022. University of Manchester, MCHE Seminar Series. Manchester, UK.

Schneider P. *The Online elicitation of Personal Utility Functions (OPUF) Tool: a new method for valuing health states.* [Virtual seminar] March 2022. University of Oxford, HERC Seminar Series. Oxford, UK.

Schneider P. *The Online elicitation of Personal Utility Functions (OPUF) Tool: a new method for valuing health states.* [Virtual presentation] February 2022. University Duisburg/Essen, Wasem Jour Fix. Duisburg, Germany.

Schneider P. *The online elicitation of personal utility functions (OPUF) Tool: a new method for valuing health states.* [Invited virtual talk] February 2022. Institute for Quality and Efficiency in Health Care (IQWIG). Cologne, Germany.

Schneider P. *The OPUF tool: a new approach for eliciting EQ-5D-5L health state preferences on the societal, group, and individual level.* [Virtual seminar] January 2022. Health Economics Association Ireland (HEAi) Seminar Series. Dublin, Ireland.

Schneider P. *The Online elicitation of Personal Utility Functions (OPUF) Tool: a new method for valuing health states.* [invited talk] January 2022. University of Bielefeld, School of Public Health. Bielefeld, Germany.

Schneider P. *A New Tool for Valuing Health States: Online elicitation of Personal Utility Functions (OPUF) for the EQ-5D-5L.* [Podium discussion] November 2021. ISPOR Europe. Virtual.

Schneider P, van Hout B, Brazier J, Heisen M, Devlin N. *The Online elicitation of Personal Utility Functions (OPUF) Tool for EQ-5D-5L.* [Presentation] November 2021. dggö workshop: resource allocation. Mannheim, Germany.

Schneider P, van Hout B, Brazier J, Heisen M, Devlin N. *Valuing EQ-5D-5L health states using a compositional approach.* [Virtual poster] October 2021. SMDM Virtual conference.

Schneider P, van Hout B, Brazier J, Heisen M, Devlin N. O*nline Elicitation of Personal Utility Functions (OPUF) for EQ-5D-5L Health States.* [Virtual presentation] September 2021. Valuation working group. EuroQol Topical Online Meeting. Virtual.

Schneider P. *Online Elicitation of Personal Utility Functions (OPUF) for EQ-5D-5L Health States.* [Virtual seminar] September 2021. International Health Economics Association (iHEA) online seminar series. Virtual.

Schneider P. *Online Elicitation of Personal Utility Functions (OPUF) for EQ-5D-5L Health States.* [Virtual seminar] September 2021. Office of Health Economics (OHE) Brown Bag Seminar. London, UK.

Schneider P. *The QALY is ableist.* [Virtual paper discussion by Elly Stolk] September 2020. lolaHESG, NL.

Schneider P, van Hout B, Brazier J. *Fair interpersonal utility comparison in the valuation of health: a relatively utilitarian preference aggregation method.* [Paper discussion by Michal Jakubczyk] Virtual EuroQol Topical Meeting 2020. September 2020. EuroQol.

Schneider P, van Hout B, Brazier J. *Fair interpersonal utility comparison in the context of health valuation studies.* [Paper discussion by Stefan Lipman] February 2020. EuroQol Academy Meeting. Prague, Czech Republic.

Schneider PP. *Interpersonal comparability of health state utilities: why it is unfair to measure preferences in units of full-health-time, and what we can do about it.* [Paper discussion by Koonal Shah] January 2020. HESG Newcastle, UK.

Schneider P. *Democratic social tariffs.* [Paper discussion by Daniel Howdon] July 2019. HESG East Anglia, UK.

Table of contents

INTRODUCTION

‘Have you guessed the riddle yet?’ the Hatter said, turning to Alice again.

‘No, I give it up,’ Alice replied: ‘What’s the answer?’

‘I haven’t the slightest idea,’ said the Hatter.

– Lewis Caroll, Alice in Wonderland

Preamble and thesis outline

The Quality-Adjusted Life Year (QALY) is made up. It cannot be observed. It cannot be measured. Unlike a physical property, such as the weight of a helium atom or the speed of light, health state utilities cannot be quantified in any objective way. That is to say, the value of health is a social construct based on subjective judgments, conventions, and norms.

There are many different ways in which health *could* be valued, in which social value sets *could* be modelled. So choices have to be made. Some of those choices are pragmatic and technical (e.g. which statistical model to use, how to recruit participants), but others are distinctly normative: *How are preferences elicited? From whom? And how are individual preferences compared and aggregated?*

Evidently, these choices have to be made, if QALYs are to be used for decision-making. Notwithstanding, the contingency of these choices is often overlooked, as a consequence of a very technical, seemingly objective, approach to *measuring* preferences. The underlying (strong) normative assumptions are often not made explicit, inhibiting critical discussions about the ethical implications of current practices.

This thesis seeks to advance the understanding and practice of health valuation in the context of health technology assessment by exploring important normative *and* methodological issues. It examines the ways in which individual health (state) preferences are elicited, aggregated and used to create value sets, and explores alternative approaches.

The thesis is written in a ‘publication format’ style[1] (not to be confused with a ‘thesis by publication’). This means, it mainly comprises a collection of publica-

[1] See *https://www.sheffield.ac.uk/research-services/code/thesis/preparation/formats#publication*

tion-formatted papers. Three papers have already been published in peer-reviewed journals; the others are currently under review or in preparation. Publication format necessitates some repetition. Slight variations in the way the same issue is presented in different papers are also unavoidable. Some of the chapters are preceded by a short introduction to provide additional context and (where needed) to specify contributions.

The unifying theme across all this thesis's papers is the challenge of determining the value of health in a plural society, consisting of many individuals, each with unique preferences, values, and perspectives. Overall, this work calls current standard practices into question, and develops new starting points for a better, more pluralistic approach to the valuation of health.

The thesis is structured as follows: it begins with a broad overview of normative issues in the valuation of health, and the current practices of valuing health in the UK (Chapter 1). The main body of the thesis is organised into two parts. Part I focuses on the question of how individual preferences are aggregated into a social value set. It uses an applied ethics approach to explore related normative issues (Chapters 2-5). Part II is more practical, reporting on the development and pilot testing of a new preference elicitation method, which – potentially – can better account for the heterogeneity of preferences between individuals (Chapters 6-8). Finally, the thesis concludes with a discussion of the main findings and their implications for future research (Chapter 9). A more detailed outline of the thesis is provided below.

Thesis outline

In this **first Chapter** (*Literature review: health valuation for HTA*), I provide a narrative scoping review of the normative foundations of health valuation, key normative choices, and related current practices.

In **Part I** (N*ormative issues in the valuation of health*), I challenge the seemingly unconscious ways in which individual preferences are aggregated into a social value set, without giving due consideration to the comparability of preferences between individuals. While this has been a topic of much debate in other fields, it has received little attention in health economics. The aim is to advance the understanding of the ethical issues underpinning current methods, and to identify potential alternative approaches.

In **Chapter 2** (*Social tariffs and democratic choice*) I argue that social value sets should be understood as an instrument of democratic participation. I discuss the implications of this view for the method used to aggregate individual preferences and explore alternative tariff specifications and decision rules.

A different approach towards preference aggregation, based on the notion of relative utilitarianism, is proposed in **Chapter 3** (*Fair interpersonal utility comparisons in the valuation of health*). It is presented as a potentially fairer alternative to the current practice of aggregating preferences by taking the arithmetic average.

In **Chapter 4** (*Setting dead at zero?*), I examine the widely accepted practice of setting dead to a utility value of zero and argue that, despite its wide adoption, there is no theoretical imperative for doing so. I rebut four arguments commonly used to justify setting dead at zero and conclude that setting dead to a different value may well be permissible.

Chapter 5 (*The QALY is ableist*) discusses the (un)ethical implications of valuing the lives of people with disabilities and in poor health by applying a social value set with negative utility values assigned to states considered (by the general public) to be worse than dead. I argue that this practice is unequivocally ableist, and should be stopped.

In **Part II** (*Practical Tools*), I address the problem of relying on a single reference case to derive social value sets, which, in the case of the National Institute for Health and Care Excellence (NICE) in the UK, is supposed to reflect the average general population preference for EQ-5D health states. Given the arbitrary nature of this choice, and in view of the abundance of potentially equally valid alternative perspectives, this definition of social value seems too narrowly prescriptive. I argue a more pluralistic approach is needed: particularities of different contexts and different groups of people, or even individuals, should be taken into account or at least taken into consideration. For this purpose, a practical tool, called 'Online elicitation of Personal Utility Functions (OPUF)' is proposed. It is a new method for eliciting health state preferences, based on compositional preference elicitation techniques, allowing for the construction of value functions for small groups of patients, and even on the individual level. OPUF is implemented in a modular open source software package to provide researchers with more flexibility in how they elicit preferences and from whom, so that context-specific preference information can be made more widely available to decision makers. Part II of this thesis reports on the development and pilot testing of OPUF. It consists of the following chapters:

Chapter 6 (*The OPUF tool*) reports on the initial development of the OPUF tool as a new type of online survey for creating value sets using compositional preference elicitation methods, implemented for the EQ-5D-5L instrument. The tool was refined using a series of iterative design cycles and piloted in a sample of 50 participants from the UK.

Chapter 7 (*Not Just Another EQ-5D-5L Value Set for the UK*) presents the findings of a study that used OPUF to elicit EQ-5D-5L health state preferences from a representative sample of the UK general population. The main objective was to explore the variability and heterogeneity of preferences between individuals. The results demonstrate that preferences vary greatly between individuals and that demographic characteristics explain only a small proportion of the variability between subgroups.

In **Chapter 8** (*Using OPUF to derive a patient-based value set*), I present the results of a study that tested the OPUF approach to derive an EQ-5D-5L value set from a relatively small sample of patients with rheumatic diseases in Germany. It was found that OPUF was generally well received and that a plausible, logically consistent, EQ-5D-5L value set was derived with good precision, despite the small sample size.

A joint discussion of the results from both parts is provided in **Chapter 9** (*Discussion and Conclusion*). It will outline the many limitations of the work presented in this thesis and identify avenues for future research.

Chapter 1

Literature review: health valuation for HTA

BACKGROUND

Health valuation is a small, but essential, cog in the large and complicated machinery of ‘health economic decision modelling’, which, in turn, informs societal decisions about the allocation of health care resources. To understand the significance of health valuation, it is helpful to begin by considering its role in the wider context of health technology assessment (HTA).

HTA is a process of evaluating the medical, economic, social, and ethical implications of using new and old health technologies such as drugs, medical devices, and diagnostic tests. In countries like the UK, the Netherlands, and Canada, it involves a systematic process that includes gathering and analysing evidence, deliberation by an independent appraisal committee, developing recommendations, and communicating the results. HTA is increasingly adopted by health care systems around the world and is used to inform decisions about the reimbursement of health technologies.

Health economic evaluation is an important component within HTA that aims to provide a systematic and transparent evaluation of the incremental cost-effectiveness of health technologies (Charlton, 2022). For this purpose, the costs and benefits of different courses of action are assessed and compared. While costs can be observed more or less directly, for example, in clinical trials or obtained from administrative data, the assessment of health benefits is more difficult.

Central to the measurement and valuation of health benefits in HTA in the UK, as well as many other countries, is the concept of the QALY (Weinstein et al., 2009). Nowadays, the QALY is defined by the National Institute for Health and Care Excellence (NICE) as a "measure of the state of health of a person or group in which the benefits, in terms of length of life, are adjusted to reflect the quality of life. One [QALY] is equal to 1 year of life in perfect health" (NICE, 2016). On a more technical level, it is the arithmetic product of length and health-related quality of life (HRQoL).

Measuring the length of a person's life is relatively straightforward. Data on survival times can be obtained from clinical trials or observational studies (although in practice, survival times often have to be extrapolated beyond the observed period, which can introduce considerable uncertainty). A person's HRQoL, on the other hand, cannot be directly observed. Because it is generally not based on the person's own account of their experienced quality of life, it needs to be constructed in a rather complicated process, I refer to as 'health valuation'. I will describe the intricacies of this process in more detail below, but here, it will suffice to say it broadly involves two steps:

First, a descriptive system is devised that defines mutually exclusive health states. The EQ-5D-5L is NICE's current reference case, but there are many other systems that could be used.

Secondly, a value set is constructed that maps any health state onto a utility value. These values - often also referred to as indices, scores, or weights - are preference-based as they are derived by eliciting health state preferences from a group of participants (e.g. a representative sample of the population, or patients with a specific disease). The resulting values lie on a scale that is anchored at one, equal to full health, and zero, equal to being dead. Negative values are also possible and are assigned to health states considered worse than dead.

Once these two components (the descriptive system and the value set) are in place, QALYs can be calculated by observing what health state a person is in and for how long. The number of years of life lived in a given health state is multiplied by the respective utility value from the value set. QALYs are then fed into health economic evaluations as inputs for the cost-utility analysis (a special type of cost-effectiveness analysis, in which the effectiveness is measured in terms of QALYs), which ultimately returns incremental cost-per-QALY estimates.

Furthermore, a decision threshold is needed to determine whether an intervention is cost-effective. This threshold specifies the social value of an additional QALY gained in the health care system. Conceptually, this is the monetary value of (one year in) full health. It can be based on one of four approaches: 1) opportunity costs, i.e. marginal productivity in the health care system; 2) population-based willingness to pay, e.g. from surveys; 3) the value of a statistical life, derived from revealed preferences for avoiding risk of death; 4) precedent, i.e. inferred from previous decisions (McCabe et al., 2008; Robinson et al., 2017).

As described above, social value sets play a crucial role in health economic evaluation, with a potentially large impact on the assessment of whether or not a given health technology is considered cost-effective.

But what is a social value set, exactly? What do the values represent? What do they measure? What are the underlying assumptions?

Formally, a value set can be defined as a mapping from a set of health states to a set of values, which are then used to compute QALYs. Conventionally, these values are expressed on a scale that is anchored at 0 (death) and 1 (perfect health), with negative values assigned to health states considered worse than death. Unit-less 0-1 scales are also used.

This technical definition, however, does not provide much insight into what a value set actually represents. Depending on the methods used and the underlying theoretical framework, value sets can represent different constructs and be subject to different normative considerations.

It is not a coincidence that the values of a value set are referred to by a number of different terms: utilities, indices, scores, or preference weights, social values, HRQoL, health state values, or QALY weights (Wisløff et al., 2014). I take this to reflect the plurality of views and perspectives on what a value set is.

Therefore, this chapter provides a narrative scoping review of the literature to identify the underlying normative foundations and paradigms through which a value set should be assessed. This is essential for a proper understanding of value sets and their implications in an HTA context.

The purpose of the review is to determine the normative foundations and paradigms of health valuation. It is, as a research question, deliberately broad. Therefore, the review takes a two-pronged approach:

In the first part, I briefly revisit the history of the ideas which have shaped the current understanding of health valuation, and I then outline key theoretical concepts and paradigms. I call this approach ‘upstream’. In the second part, which I call ‘downstream’, I give an overview of alternative operational definitions of a social value set – in other words, different ways in which a value set can be derived – and then I highlight the key normative choices associated with these different approaches. Finally, I discuss the review findings and their potential implications.

Given the interdisciplinary nature of the topic, a systematic search strategy was not feasible. Instead, various papers from different fields were chosen as starting points, and then a snowballing and ad-hoc manual focused search in PubMed and Google Scholar was carried out to identify additional relevant literature, or to follow up on specific questions. The analysis was performed inductively by identifying common themes and concepts, and with no formal classification scheme. The results are presented in the form of a narrative review, following the historical development of ideas and to juxtapose the different concepts and approaches.

This review is not meant to be exhaustive, nor are the different theories or approaches discussed in much detail. The literature on welfare economics can be especially technical, with an emphasis on formal mathematical proofs, which is beyond the scope of this thesis. Instead, I try to give a brief and accessible account of the main ideas, providing an overview of the ‘research landscape’ and illustrating the vast diversity of views and perspectives on health valuation.

Classical utilitarianism

Utilitarianism was to an extent already present in the works of earlier philosophers (e.g. Hutcheson and Hume), but it was Jeremy Bentham (Bentham, 1789) who developed the first systematic account of utilitarianism in the 18th century (Driver, 2022). Bentham (and others, subsequently) sought to build a general ethical theory based on the maximisation of the "greatest happiness for the greatest numbers". Utility was conceived as a measure of the pleasure or happiness that a person derives from a particular action or state of affairs. Every individual was assumed to have a utility function, measuring the level of pleasure or satisfaction associated with each possible state of the world. The summation of those utilities across all individuals then yielded a global evaluation of any given state, what Bentham referred to as the "Hedonic Calculus", allowing for a quantitative comparison of alternative courses of action (Barberà et al., 2004, pp. 1181).

The concept of a social value function was implied in Bentham's work, albeit without any elaborate specification or justification. Individuals were simply assumed to measure utility in a common unit because doing so allowed full cardinal interpersonal comparability. This was a requirement because the arithmetic mean (or sum) of individual utilities determined which social states were considered good or bad (Harsanyi, 1988; Barberà et al., 2004, p. 1129).

The ordinal revolution

The main challenge of classical utilitarianism was how to measure utility. Some economists argued utility was directly measurable. Edgeworth, most notably, thought that future discoveries in physio-psychology would enable development of a "hedonimeter" to measure individual utilities (Colander, 2007). Overall, however, economists were unable to find a common viewpoint on the meaning of, and conditions for, utility measurement, and Classical Utilitarianism was eventually abandoned in favour of the more structural approach of "new welfare economics" pioneered by Pareto, Hicks, Allen, and Samuelson. It was a revised demand theory, built on a more basic concept of ordinal preference relations and individual indifference curves. This new theory only required individuals to express their preferences in binary terms: either they preferred state A over B, or B over A, or they were indifferent. This was all the information needed. Utility values were used as secondary numerical indices only and were no longer related to the experience of satisfaction or happiness. Also, utilities were now interpersonally incomparable, and could not be aggregated across individuals (Barberà et al., 2004, p. 1181).

The main premise of the so-called 'ordinal revolution' was expressed by Robbins (1932) in a seminal essay, in which he claimed that interpersonal comparisons of utility were 'unscientific': "It is a comparison which necessarily falls outside the scope of any positive science. [...] It involves an element of conventional valuation. Hence it is essentially normative. It has no place in pure science".

The 'Pareto criterion' became the key concept for evaluating changes in social states: any change that makes at least one individual better off without making anyone else worse off should be considered socially desirable, arguably without making any subjective value judgements, or at least only requiring minimal normative assumptions (Buchanan, 1959). In the new paradigm, evaluating social states thus did not involve aggregating individual utilities in a social value function, but, individual preferences were taken as points in an Euclidean space with dimensions equal to the number of individuals involved. There could then be many Pareto optimal states, all located within an efficiency frontier. Yet, the new welfare economic theory could not provide any guidance on how to choose between different Pareto optimal states, even if in some states only one individual was only marginally better off, and in others many benefited greatly – comparing these scenarios is not possible without making interpersonal comparisons.

To overcome the shortcomings of the ordinal utility paradigm, two new approaches were developed: the Kaldor-Hicks compensation school and the Bergson-Samuelson social welfare function approach.

The Kaldor-Hicks compensation criterion

The Kaldor-Hicks compensation criterion allowed the evaluation of changes in social states by considering potential compensations. A change was desirable if those who gain could compensate the losers. Interestingly, the compensation did not actually have to take place, but only be hypothetically possible (Kaldor, 1939).

The compensation-criterion was received with considerable criticism. Robbins (1981) argued that the "the fact that such compensation is conceivable is not sufficient: if it is not actual, the fundamental Paretian condition is violated." Similarly, Sen (2018, p. 56) argued that unrealised compensations could not determine whether an improvement had actually taken place because the 'losers' may be poorer, needier or more deserving than the 'winners'. If the compensation was paid, however, it would be an actual Pareto improvement, and a compensation test was no longer needed.

Notwithstanding the criticism, contemporary benefit-cost analysis conventionally refers to the Kaldor-Hicks criterion as their theoretical foundation from which they arguably draw their legitimacy (Adler and Posner, 1999).

The Bergson-Samuelson Social Welfare Function

Bergson and Samuelson formulated a different approach, which acknowledged the Pareto criterion as a necessary, but not sufficient, requirement for policy evaluation. They argued that it would need to be supplemented by some explicit value judgment, specified as a social welfare function, to choose between Pareto optimal states (Bergson, 1938; Harsanyi, 1988). Their social welfare function was a rather theoretical construct, which provided much flexibility, as it did not specify at all how the function should be constructed. Its only requirement was the function should always increase or decrease when an individual's ordinal preference increases or decreases, all else being equal. This flexible framework allowed for a wide range of possible social welfare functions.

Arrow's Impossibility Theorem

Arrow's celebrated impossibility theorem (Arrow, 1983) was a damaging critique. It showed there can be no democratic procedure for aggregating individual preferences into a collective decision which simultaneously satisfies the conditions of collective rationality, the Pareto principle, the independence of irrelevant alternatives, and non-dictatorship. Put simply, it proved a social welfare function could not be constructed based on ordinal preferences without violating seemingly fundamental axioms of rationality.

Arrow's influential result also led to the development of the field of social choice theory, which studies the aggregation of (ordinal) individual preferences into collective decisions. This field's main focus is to identify the formal conditions under which collective decisions are consistent with individual preferences, and how the properties of the decision-making process influence the outcome (Barberà et al., 2004, p. 1195).

Sen: Broadening the informational base

Building on the developments in welfare economics and social choice theory outlined above, Sen (2014) developed an entirely new approach for evaluating social states. A key motivation for his framework was the observation that people living in poverty often become resigned to their circumstances and may even be content with having fewer resources. However, this does not mean that their situation is acceptable. With reference to this example, he argued that welfare economics had been too narrow in its focus on utility. Rather than trying to measure and compare subjective states of mind or subjective wellbeing, he advocated for considering other, more objective, information to assess individual welfare (Sen, 1995).

With Nussbaum, Sen developed the capability approach, which is a conceptual framework for assessing individual well-being and evaluating social arrangements. It focuses on what people are able to do, on the capabilities that people have to "lead the lives they have reason to value" (Nussbaum and Sen, 1993; Sen, 2014), rather than simply assessing subjective states of mind.

This "broadening of the informational base" was no mere modification of an existing framework, but a change in the paradigmatic foundation of welfare economics, allowing once more for interpersonal comparisons of levels and differences - not necessarily in units of utility, but of other indicators (Vanberg, 2018). It provided a flexible framework, in which the construction of social value functions was permitted based on a variety of indicators. As will be described below, Sen's approach appears to have been the main point of reference from which health economists have drawn to develop Extra- or Non-Welfarist approaches to healthcare value frameworks (Cookson et al., 2012; Daniels, 2010).

The QALY framework

The term QALY was first used by Bush et al. (1972), but the idea for an outcome measurement unit that combines duration and quality of life had already been shaped by previous work, which took place in parallel to other developments in welfare economics. In fact, research on health status measurement and quantification in medical literature dates back to the 1930s, when Thorndike (1937) asked respondents to indicate how much money they would be willing to accept to suffer "certain pains, deprivations, and frustrations", such as a headache or a broken leg. In the 1940s through 1960s, the quantification of health and quality of life received evermore attention. Tools and methods were developed by medical professionals, but other disciplines including philosophy, operations research and psychology took an interest in the topic as well (Klarman and Rosenthal, 1968; MacKillop and Sheard, 2018). In the early 1970s this field of research really gained traction, with seminal contributions from Torrance et al. (1970) and Fanshel and Bush (1970), who proposed the time trade-off (TTO), the person trade-off, and the standard gamble (SG) methods for eliciting health state preferences. Later, the discrete choice experiment (DCE) was developed by Louviere and Hensher (1982).

In the UK, widely acknowledged contributions were made by Alan Williams, who, in collaboration with many other researchers, most prominently Rosser and Kind (1978), had decisive influence on the development of the QALY framework and its implementation into health policy decision making (Brazier et al., 2017b; Culyer, 2007; Kind, 2005; Williams, 1985). A key driver of the changes were the increasing pressures on health care budgets and the need for more efficient use of resources. Williams also co-founded the QoL Measurement Group, leading to the formation of the EuroQoL group in 1993, which is now a leading organisation in the field of HRQoL measurement (Devlin and Brooks, 2017). A more extensive account of the history of the QALY framework, based on a review of the literature and a total of 44 semi-structured interviews with academics and civil servants was produced by MacKillop and Sheard (2018).

The QALY under welfarism

Despite the popularity of the QALY and its apparent usefulness in making and communicating health policy decisions, it is important to note that it was not directly derived from economic theory. There are no explicit theoretical justifications for the QALY derived from first principles or welfare economic literature (Garber and Phelps, 1997). A welfarist interpretation would require the number of QALYs to be proportionate to the level of an individual's overall, or at least health-related (cardinal), utility (Brazier et al., 2017b). However, Pliskin et al. (1980) convincingly demonstrate that QALY maximisation is consistent with the maximisation of an individual lifetime health only under very restrictive and (it might be said) unrealistic assumptions as follows:

1. Utility independence between life years and health status: This comprises, first, the utility of health status in relation to life expectancy and, second, the utility of life expectancy in relation to health status.

2. Constant proportional tradeoff: This assumes the life years a person would give up for better health would not be affected by the number of years they have left to live.

3. Risk neutrality on life years: This assumes that, for any given health level, each prospect is equal to its expected value. For example, it assumes an individual would be indifferent between a 50-50 chance of immediate death and 10 years of full health, and a guaranteed 5 years in full health.

Accordingly, QALY maximisation would be consistent with the maximisation of overall utility (i.e. including preferences over consumption and any other non-health aspects of life) only under even more restrictive assumptions (Bleichrodt and Quiggin, 1999).

The QALY under extra-welfarism

A potentially more plausible justification of the QALY framework can be derived from what has been called Extra-Welfarism (some also refer to it as Non-Welfarism). It provides an alternative approach to modern welfare economics, which, analogous to Sen's capability approach, allows for a broader informational base and rejects the idea that individual utility is the sole indicator of social welfare. In an influential publication, Culyer (1989) referred to it as the "undue information restriction in welfarism". Thereby, extra-welfarism enables the construction of social value functions on the basis of other indicators beyond utility, such as health, independently of how it is valued by the individuals themselves (Brazier et al., 2017b). In this context, health may also be considered a merit good, which Musgrave (1959) defined as a good "which, due to imperfect knowledge, individuals would choose to consume too little.".

Showing further similarities to Sen's capability approach, extra-welfarism provides a very flexible framework, under which a wide range of approaches may fit. In fact, any value framework that is not strictly welfarist could be considered extra-welfarist. Brouwer et al. (2008) did however identify four areas in which welfarism differs from extra-welfarism:

1. Outcomes considered relevant in an evaluation: Welfarism only considers individual utility, whereas extra-welfarism permits consideration of utility, health and other outcomes.

2. Sources of valuation of relevant outcomes: In welfarism, only the utilities of affected individuals matter, whereas in extra-welfarism, the valuation of individuals who are not practically affected may also count.

3. Basis of weighting relevant outcomes: In welfarism, the weights are only determined by individual utility, whereas in extra-welfarism, the weights may be determined by criteria such as age (see 'fair innings': (Williams, 1999), deprivation (see e.g. (McNamara et al., 2020)), or remaining life expectancy (see e.g. (Shah et al., 2015)).

4. Interpersonal comparisons: In welfarism, permissible comparisons between individuals depend on the school of thought. They are either not permitted at all, permitted using the potential Pareto improvement, or permitted only to select points on the Pareto frontier. In extra-welfarism, interpersonal comparisons of well-being are permitted in a variety of dimensions.

An often referred to variant of extra-welfarism is the 'decision maker approach', which emphasises the role of health economic evaluation as a tool to aid decision making. It aims to achieve whatever goals are given by the relevant decision makers. The role of health economists is then limited to supporting them in making decisions consistent with given objectives – rather than suggesting objectives themselves. In principle, the decision maker approach does not relate to any particular theoretical framework, although, in practice, it is often equated with the current extra-welfarist QALY framework (Brouwer and Koopmanschap, 2000; Coast, 2004).

Some confusion is often caused be the imprecise use of the term utility within Extra-Welfarism. While in welfarism, utility refers to the individual's own assessment of their own welfare, in extra-welfarism, utility theory is used to derive measures of characteristics of individuals that correspond to entirely different concepts. Yet, confusingly, these are generally also called ‘utility’ measures (Cookson et al., 2012, p. 52).

Proponents of extra-welfarism argue that it provides a more policy relevant analytical framework. Especially in situations where equity is a key concern alongside efficiency, or where the explicit policy objective is to maximise a concept other than welfare (such as population health) (Culyer, 1989), it may seem more appropriate to complement or even replace welfarist assessment of utility with non-utility information.

As a criticism of extra-welfarism, one could argue that it is inherently paternalistic (Brouwer et al., 2008). Instead of relying on the individuals to decide for themselves what is valuable to them, it requires some authority to make a range of normative judgements to construct a social value function (which will be discussed below). The authority will effectively determine the basis on which societal decisions are made (Brouwer et al., 2008). A further criticism of extra-welfarism is that it does not provide any guidance on who ought to make these decisions. It could be a policy maker, an appointed committee, a citizens’ jury, or some other organ. It remains unclear how the authority should be chosen, what its legitimacy is, or how it should be held accountable for its decisions (Culyer, 1989).

Further paradigms

Besides welfarism and extra-welfarism, there are other theoretical frameworks and pertinent paradigms under which methods for valuing health can be considered. These include deliberative democracy, communitarianism, multi-criteria decision analysis, psychometrics, and behavioural economics.

Deliberative democracy

Deliberative democracy is a political theory, supported by a substantive body of literature, which provides a normative framework for collective decision making through open and reasoned discourse (Davies et al., 2006). In the context of health valuation, or preference elicitation more generally, deliberation usually entails providing respondents with relevant information, and encouraging an open and reflective discussion. Communication and debate are considered essential. Often, trained facilitators are involved to guide the discussion in a constructive way.

Normative frameworks for deliberative health valuation have been proposed by Hausman (2015, 2010) in the form of 'public values', which, he argues, require a broad public debate ('collective deliberation'), and by Baker et al. (2021), building on work by Sunstein (1994), in the form of 'incompletely theorised agreements'.

There are two types of arguments for the use of deliberation: instrumental arguments claim that deliberation helps people construct their own preferences better. It educates them, reduces mistakes in reasoning, and improves the quality of the decisions that are made (Karimi et al., 2019). Ultimately, this means the outcomes will be better. The second type of argument maintains that deliberation has an intrinsic value. It provides legitimacy to the decisions that are made (Elster and Przeworski, 1998) and it is more respectful of the autonomy of the respondents, as decisions are taken collectively, and based on a shared understanding of the relevant issues. In this regard, deliberation can help to mediate, or even 'transform' disagreement, as an alternative to aggregation and/or as a way of acknowledging pluralistic values (Baker et al., 2021; Knight and Johnson, 1994).

A number of studies have investigated the effect of deliberation on health valuation. The results are mixed: some studies found that deliberation improved the ‘quality’ of decisions, while others found no significant effect (Gansen and Klinger, 2020; Karimi et al., 2019). The main drawback of deliberation is that it is time and resource consuming. Usually, it requires a number of sessions, and only a small number of respondents can be involved (Davies et al., 2006).

Communitarianism

An alternative basis for the allocation of health care resources, referred to as communitarianism, was proposed by Mooney (Mooney, 1998). It states that social welfare is more than the aggregation of individual preferences over individual programmes, and that it would be important to consider community values, such reciprocity, sharing and caring. A key concept of communitarianism is that ‘claims’ are neither welfarist nor extra-welfarist but allow the community to decide what is owed to each member. The theory has not been developed into any actionable approach, and thus it is not clear how it would operate in practice.

Multi-criteria Decision Analysis

Multi-criteria decision analysis (MCDA) is a methodological framework to inform decisions involving multiple, often conflicting, objectives. It was developed in the 1960s as a decision support tool and entails a broad set of methods to structure and analyse decision problems and to elicit preferences and values from decision makers. Trade-offs are expressed in a multi-attribute utility or value function (Keeney et al., 1993). MCDA has been applied in a range of different areas, including health care, to increase the consistency, transparency, and legitimacy of decisions (Thokala et al., 2016).

MCDA is principally agnostic about the type of decision problem or the content of the value function. The framework does not presuppose any specific normative theory, nor does it aim to find the ‘right’ solution. It merely provides a structured way to analyse the decision problem, and to organise and weigh relevant information. The validity of the decision and the methodology of the decision making process are ultimately judged by the decision makers themselves: principally, did they find the method useful (Belton and Stewart, 2002)? In this regard, MCDA seems compatible with the decision maker approach, and can be seen as a way to operationalise it.

It may be interesting to note, that Torrance's seminal work on health preference elicitation, which introduced the time trade-off and standard gamble method, was explicitly based on an MCDA framework (Torrance et al., 1970). Torrance specifically refers to the 'operations research methodology' laid down by Ackoff (Ackoff and Sasieni, 1968), which is a precursor of modern MCDA. Later, Torrance et al. (1982) provided a detailed account of the application of the MCDA framework to health preference elicitation, and developed the Health Utility Index (HUI) instrument (formulated as a multi-attribute utility function) (Feeny et al., 2002; Torrance et al., 1995).

Psychometrics

Another paradigm, whose influence on health valuation can hardly be overestimated, is psychometric evaluation. Psychometrics is the science of measuring mental processes and behaviour, and it provides a battery of methods to assess the quality of a measurement instrument. It is used to assess properties such as validity, reliability, and the responsiveness or sensitivity of a measure (Hays et al., 1993). Psychometric testing has been established as a standard practice and should be considered a prerequisite for the use of any health descriptive instrument (Brazier and Deverill, 1999; Finch et al., 2018).

Notwithstanding its importance, psychometrics may also have some unintended, negative side effects. Although it may not immediately seem to convey any particular normative implications, I would argue that psychometrics does provides a certain scientistic frame of reference. There is a risk that utility measures are evaluated predominantly in terms of their technical properties, while a debate about the normative foundations of the measure is avoided. Uncritical application of psychometric assessment may come at the expense of considering more fundamental conceptual issues, which are more difficult or impossible to quantify.[2]

Behavioural economics and preference construction

A final consideration is the epistemological conception of individual preferences in different fields. In the welfare economics tradition, individuals are often assumed to have pre-existing, stable, and consistent preferences. They are further assumed to be fully rational agents, who seek to maximise their personal utility. In any elicitation task, only these preferences need to be articulated (Fischhoff, 1991). Behavioural economics findings, however, have challenged this view. Framing effects, other cognitive 'biases', or (more generally) context and the

[2] Recently, Devlin et al. (2018) published an EQ-5D-5L social value set for England, which was supposed to replace the old EQ-5D-3L value set from 1995 (Group, 1995) as the reference value set for NICE. However, a methodological review by Hernández-Alava et al. (2018) flagged a number of potential issues: in a 68 page report, they criticised the data quality and some model specifications. In response, NICE invited four independent reviewers to evaluate the new value set. The reviewers were asked whether the new data was likely to reflect the preferences of the English public. They were also asked to comment on 'accuracy'. One of the reviewers, Werner Brouwer, provided the following insightful response: "We believe this question cannot be answered unequivocally because of the lack of a golden standard. Put simply: which test (on the existing data) would prove beyond any doubt that the current preference data or valuation set does or does not reflect the preferences [of] the English general public adequately?" (n.d.). NICE nevertheless decided to reject the new value set, with the consequence that the old EQ-5D-3L value set, which never underwent similar scrutiny, remains in use. Sampson (2022) also commented on this episode: "Normative statements cannot be tested. The validity of methods to derive a value set can be tested. However, they cannot—or at least should not—be tested against some unattainable perfection. In lieu of a methodological 'gold standard', perfection became the comparator."

overall choice architecture play an important role in decision making and the construction of preferences. Individuals may, for example, fall back on simple heuristics during a complex preference elicitation task, or they may be influenced by the way in which the task is presented. In fact, respondents can often be inconsistent or intransitive in their preferences. This has led to a growing body of literature that emphasises the importance of context and framing effects in the construction of preferences (Slovic, 1995). Fischhoff (1991) referred to the whole spectrum, ranging from full rationality to full irrationality, as "a continuum of philosophies".

Considerations regarding people's cognitive capacity and framing effects are of practical importance when designing a health valuation study. Particular attention should be paid to the ways in which preferences are constructed 'on the fly', and how the elicitation task itself may influence this process (Dolan, 1997). This includes, but is not limited to, the wording of questions; the presentation and amount of information provided; colours used to highlight certain options; the number of tasks and the time required to complete those tasks (Himmler et al., 2021; Jonker et al., 2018; Peasgood et al., 2021).

DOWNSTREAM: OPERATIONAL DEFINITIONS OF A SOCIAL VALUE SET

Having looked at the ‘upstream’ theoretical foundations of social value sets, we now turn to the ‘downstream’ operational definitions and consider their implied normative implications. There are many other plausible ways to derive a value set and, depending on the choices made in the process, they can represent different constructs. I will focus on two key normative choices: 1) what is being valued?; and 2) whose values are being elicited? I will then briefly outline further methodological questions with important normative implications: 3) additional degrees of freedom. The review is not meant to be exhaustive, but is intended to provide a broad overview of the ‘landscape’ and to illustrate the variety of practical approaches, which may allude to the diversity of the underlying concepts and paradigms.

1. What is being valued?

A key distinction between alternative approaches to derive a value set is the evaluative space, i.e. what is being valued. One can look at this on two levels: First, on a conceptual level, one can try to identify the construct that is being valued; and, secondly, on an operational level, one can consider the actual instrument/descriptive system used to capture it.

Conceptual dimension: the construct behind the value set

The QALY is usually described as a measure of health or health-related quality of life (HRQoL). Both terms are broad concepts, which do not lend themselves to simple definitions. For health, one may only consider its holistic definition as “a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity” proposed by the World Health Organisation (WHO, 1948), or the more recent, dynamic conception of health as "the ability to adapt and to self-manage" (Huber et al., 2011), to name just two examples to illustrate the diversity of views. Many more definitions have been proposed in the last few decades (see e.g. (Simmons, 1989)).

The definition of HRQoL seems similarly ambiguous. Karimi and Brazier (2016) identified at least four definitions of HRQoL commonly used in the literature. These include: 1) 'how well a person functions in their life and his or her perceived wellbeing in physical, mental, and social domains of health'; 2) 'quality of life is an all-inclusive concept incorporating all factors that impact upon an individual's life. Health-related quality of life includes only those factors that are part of an individual's health'; 3) 'those aspects of self-perceived well-being that are related to or affected by the presence of disease or treatment'; and lastly, 4] a purely instrumental definition of HRQoL as 'values assigned to different health states' which are used to construct QALYs.

NICE, however, uses the term HRQoL to refer to the health state profiles themselves. In their latest method guide, they state, "the valuation of health-related quality of life measured by patients [...] should be based on a valuation of public preferences" (NICE, 2022). According to their interpretation, the descriptive system measures HRQoL, which is then valued by assigning preference weights to the profiles in a separate step.

In the context of health valuation, concepts of HRQoL usually entail a subjective component, which is linked to a person's experience or value, but even more objective naturalistic conceptions (e.g. health as functional efficiency) have been proposed (Hausman, 2015, p. 29).

Overall, there is no consensus on a single concept of HRQoL. Various interpretations are used concurrently, and QoL and HRQoL are frequently used interchangeably despite being distinct concepts (Karimi and Brazier, 2016). Consequently, it has been questioned whether the term HRQoL should be used at all, as it may be a source of considerable confusion (Ronen, 2017).

Moreover, there have been calls to consider outcomes beyond health in health economic evaluation, viz. to include other domains such as social outcomes, consumption, and/or well-being (Brazier and Tsuchiya, 2015; Cookson et al., 2021). The main motivation for doing so is to allow for a more comprehensive assessment of the value of interventions and to enable comparisons across sectors (e.g. between health, education, and social care). While this could be useful to better assess complex interventions (e.g. public health policies to reduce childhood obesity or alcohol-related harm), it would entail an expansion of the evaluative space beyond health or HRQoL.

Operational dimension: the instrument

Despite the challenges of defining HRQoL and the evaluative space of the QALY in theory, there is a breadth of instruments - so called (health) descriptive systems – that aim to operationalise these concepts in practice. They provide a classification of (health) states or profiles, usually with multiple dimensions or attributes. There is considerable overlap with the concept of patient-reported outcome measures (PROMs), which are used in clinical practice and research to assess the impact of a disease on the health of an individual. PROMs for which preference weights have been derived, and which can be used to compute QALYs, are called preference-based, whereas PROMs without such weights are referred to as profile-based measures (Al Sayah et al., 2021).

Measures can be further divided into generic and condition-specific measures. Generic preference-based measures of HRQoL are designed to assess a person's overall health and to be applied to a wide range of settings, allowing for comparisons across different conditions. In contrast, condition-specific instruments aim to capture information on health problems that are more specific to a particular disease.

There are a number of generic preference-based measures of HRQoL available for use in health economic evaluation.[3] A useful overview is provided by Brazier et al. (2017a), who compare seven well established measures: 15D, AQoL-8D, EQ-5D-3L, EQ-5D-5L, HUI3, SF-6D, and QWB-SA. A potentially interesting new development was the introduction of PROMIS (Patient-Reported Outcomes Measurement Information System), which provides a more adaptive item bank, consisting of a core set of items (e.g. the Global Health 10), with additional items that can be selected based on the setting and/or individual's responses (Hays et al., 2009). This may allow for a more personalised assessment of HRQoL, but also poses challenges for the derivation of preference weights (Craig et al., 2014).

Despite the apparent overlap between generic instruments, there are considerable differences in some of the domains covered: first, the number of dimensions and levels, meaning the size of the descriptive system varies greatly between instruments. It ranges from 243 mutually exclusive health states (EQ-5D-3L) to more than 2 sextillion, that is 10 to the power of 23 (AQoL-8D). Secondly, they capture different domains. While most instruments have multiple items covering

[3] The most popular generic measure of HRQoL is probably the EQ-5D. It is a fairly simple instrument, consisting of just five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), with three (3L) or five (5L) levels of severity on each dimension, ranging from no problems to extreme problems. EQ-5D is available in about 170 languages (Rabin et al., 2014), and value sets have been produced for more than 25 countries (Roudijk et al., 2022), with more studies currently underway. It is the instrument that is most often mentioned in the method guides of HTA organisations around the world (Kennedy-Martin et al., 2020), and NICE in the UK prescribes its use for submissions (NICE, 2022). Other generic measures, e.g. the HUI3, have also been accepted by NICE, but only in exceptional cases (Brazier et al., 2017a).

physical health (physical activity, bodily functioning, pain, etc.), there is considerable variation in the coverage of mental health and psychosocial domains.[4]

Ultimately, different generic measures of HRQoL should be understood as capturing different constructs (Brazier et al., 2017a; Whitehurst and Bryan, 2011). Yet, even for a given instrument, it is often unclear what underlying construct it is supposed to measure. The EuroQol group, which is responsible for the development of the EQ-5D, acknowledged at their recent internal group meeting that it remains undetermined what the EQ-5D actually measures (Sampson, 2023).

Notwithstanding, there is a plethora of research assessing the construct validity, responsiveness, practicality, and reliability of various generic instruments in different populations and settings (McNamee and Seymour, 2005). However, a gold standard has not been established, and a consensus, or any conclusive guide for selecting the most appropriate generic instrument in any given setting, is lacking (Brazier et al., 2017b). It is interesting to note that the only justification NICE provides for explicitly recommending the EQ-5D is "the need for consistency across evaluations" (NICE, 2022). Further elaboration on what makes the EQ-5D better suited for this purpose than the other generic instruments is not provided.

Following calls to expand the evaluative space beyond health, generic preference-based measures have been developed to capture other domains, such as social care, mental health, and well-being. Notable examples are the ASCOT, to assess social care services (Netten et al., 2012), the EQ-HWB-S, to assess interventions

[4] Generic instruments are usually designed with an adult target population in mind, but it is increasingly recognised that certain domains, such as self-care (in the EQ-5D), social functioning (in the SF-6D), or sexual activity (in the 15D) may be less or not relevant for children or older adults, and/or the wording of the questions may not be appropriate for these age groups. A number of age-specific instruments have been developed to address this issue, such as the CHU9D and the EQ-5D-Y-3L (which only deviates slightly in the wording from the original EQ-5D-3L) for children, or the ICECAP-O for older adults. These measures raise a number of issues, not only regarding consistency of results across age groups, but also regarding the approach used to derive preference weights (see below).

across healthcare, social care, and public health (Brazier et al., 2022), and the ICECAP-A, as a measure of capability wellbeing (Al-Janabi et al., 2012). To emphasise the broadening of the evaluative space, QALYs based on these instruments are sometimes referred to as wellbeing-adjusted life years (WELBYs) (Brazier and Tsuchiya, 2015).

For completeness, mapping should also be briefly mentioned. Mapping, or ‘cross-walking’, is a statistical method that can be used when data on a target generic preference-based measure is not available (NICE, 2022; Alava et al., 2020). However, if the underlying construct of the target generic instrument is too different from that of the source instrument, the link between the two may be too weak to draw sound inferences (Round and Hawton, 2017). Franklin et al. (2018) recently came to this conclusion when trying to map the ICECAP-O to the EQ-5D-3L.

Condition-specific preference-based measures

A large number of condition-specific instruments have been developed, encompassing a vast range of disorders, ranging from uni-dimensional measures to more symptomatic and complex, multi-faceted instruments (Goodwin and Green, 2016; Rowen et al., 2017). In a recent update of a systematic review, Rowen et al. (2017) identified 36 condition-specific preference-based measures of HRQoL, across a range of 29 different conditions.

A generic instrument can also be adapted for a specific condition by using a bolt-on. Bolt-ons are condition-specific items that are added (‘bolted-on’) to a generic instrument to improve the validity and responsiveness for specific conditions, while preserving the generic dimensions of the measure (Brazier et al., 2017b; Mukuria et al., 2019).

Another approach to derive condition-specific utilities is to use a vignette study. Vignettes are simplified descriptive systems that can be used to describe health states, including treatments, usually in a narrative form. They are often used for rare conditions, or for conditions for which no condition-specific instrument or bolt-on is available (Matza et al., 2021). The main advantage of vignettes is that they are relatively easy to develop, but their restricted scope may limit their validity and generalisability (NICE, 2022).

The main rationale for using condition-specific instruments is that they are more sensitive to the impact that a particular condition has on the patient than generic instruments (Payakachat et al., 2015; Rowen et al., 2017; Versteegh et al., 2012). This means a condition-specific measure may pick up a significant benefit or harm, while a generic measure shows no overall change in HRQL. Divergences can occur in both baseline HRQoL and the change in HRQoL over time due to an intervention (Lorgelly et al., 2017). This divergence may well be legitimate, because condition-specific and generic instruments are capturing different constructs (Longworth et al., 2014). Patients themselves may also prefer a condition-specific instrument because it allows them to report on the problems they experience, whereas a generic instrument may appear less relevant to them. However, the limited comparability across different patient groups and interventions is, by many, considered a major drawback (Rowen et al., 2017).

In their method guide, NICE recommends the use of condition-specific instruments as the basis to derive QALY estimates only in circumstances where the EQ-5D (and other generic instruments) are inappropriate, e.g. when they lack content validity because a key dimension is missing (NICE, 2022). However, an internal review at NICE found the EQ-5D-3L to be appropriate in almost all cases considered: "[It] works well for most diseases and conditions except for sensory disorders and some mental health conditions. For conditions where there is mixed evidence that EQ-5D performs well, [...] it has been possible for committees to make recommendations based on EQ-5D" (NICE, 2020). Given the simplistic design of the EQ-5D-3L with just five dimensions with three levels each, this seems somewhat surprising. However, it may also reflect NICE's preference to maintain consistency across appraisals.

2. Whose values are being elicited?

Preferences differ between individuals. Evidence from previous studies suggest that experience of a particular disease, adaptation, age, or education, among other factors, may influence health state preferences (Brazier, 2005). Whose preferences are being elicited to form the basis of QALY estimates therefore matters.

In the literature, two sources of preferences are generally distinguished: patients and the general public (Versteegh and Brouwer, 2016). Practically, the 'general public' refers to a representative sample of the adult general population. 'Patients' as a source is more ambiguous. Depending on the context, it can refer to a) patients who have experience with a particular health condition or disease, b) patients who currently live in, or who have had experience with, a particular health state; or c) patients who will be directly affected by a particular policy decision.

Sometimes the two perspectives are equated with ex-ante (decision utility) and ex-post (experience utility) preferences respectively. However, the demarcation may not be clear cut and may depend on the exact context (Versteegh and Brouwer, 2016) and I shall avoid using this terminology here.

Most HTA agencies around the world, including NICE, recommend the use of preferences elicited from the general population (NICE, 2022; Kennedy-Martin et al., 2020). Even for condition-specific instruments, researchers tend to ask the general public for their preferences rather than patients with the condition, even though the general public may not have any experience with, or may lack understanding of, the health problems being valued (Goodwin and Green, 2016; Rowen et al., 2017). A notable exception is Sweden, where value sets are experience-based, i.e. derived from people's valuations of their own health states (Burström et al., 2020).

Arguments for and against the use of general population preferences

The use of preferences from the general public is supposedly in line with the societal perspective and the insurance principle (Gold, 1996). The highly influential Report on the Washington Panel also recommended the use of general population values (Russell et al., 1996). They argued that this would most accurately reflect societal values. Members of the general public would also – generally speaking – be at risk of various health problems. This would provide a veil of ignorance, which means the elicitation would not be affected by self-interest. Another often cited argument is that the general public pays for healthcare through taxes, and thus, their preferences should be taken into account when making decisions about the allocation of the healthcare budget (Versteegh and Brouwer, 2016). This is sometimes referred to as the ‘taxpayer perspective’ (Helgesson et al., 2020), which seems to be misleading as it suggests that it is only the tax payer who should have a say in the allocation of funds (a person who pays more taxes may then even have a greater say in the allocation of funds). In democratic societies, however, the public is generally considered sovereign not by virtue of paying taxes, but on the basis of basic democratic principles and their rights as citizens.

Arguments against the use of general population values include the fact that the general public may not have any, or only limited, experience with the health problems they are asked to evaluate (Brazier et al., 2005). Moreover, there is a risk of discrimination against individuals with disabilities or chronic conditions. Finally, it may be considered ethically objectionable to apply the preferences of the general public to answer questions about the desirability of a particular intervention for a patient group without considering the preferences of the patients themselves (Schneider, 2022).

Arguments for and against the use of patient preferences

Arguments for the use of patient preferences are, conversely, that patients are the ones who are directly affected by the decision, and that they are the ones who may best be able to evaluate the impact of a particular health problem on their lives (Brazier et al., 2005). Some authors argue that patient evaluations would be more accurate, yet it is not entirely clear what this means, as ex-ante preferences elicited from the general public may actually represent an entirely different construct than ex-post preferences elicited from patients.

There are also arguments against the use of patient preferences. First, patients may adapt to their condition, which can affect their preferences, resulting in a different (that is, higher) valuation of their own health state. This would put them at a disadvantage (Versteegh and Brouwer, 2016). However, as convincingly argued by Menzel et al. (2002), adaptation is a complex phenomenon. It has several elements, some of which (e.g. skill enhancement, learning) should potentially be taken into account. Secondly, in some cases, patients may simply not be able to participate in a preference elicitation study, for example, because they are too ill or too young. Some authors also mention potential ethical objections to the elicitation of preferences from patients (Brazier et al., 2017b, p. 81). Yet, it could be argued that this is paternalistic and disregarding the autonomy of patients.

Alternative perspectives

Different perspectives apply depending on whose preferences are elicited. At least four perspectives are distinguished in the literature: first, personal perspective, i.e. respondents are asked to imagine being in a particular health state themselves. If preferences are elicited from patients, they may also be asked to consider their current health state; Secondly, other person's perspective, i.e. respondents are asked to imagine someone else in a particular health state and to make decisions on their behalf; and thirdly, social perspective, i.e. respondents are asked to imagine a group of people in a particular health state and to make decisions as a social planner; lastly, social inclusive perspective, i.e. respondents are asked to imagine a group of people in a particular health state to which they belong (note, in this perspective, preferences are usually elicited with uncertainty, i.e. the group, of which the respondent is a member, faces a risk of being in a particular health state). Further nuance can be introduced by considering who the bearer of the opportunity costs is (Tsuchiya and Watson, 2017).

Compromises between the two perspectives

Helgesson et al. (2020) provides a useful overview of the different perspectives and their advantages and disadvantages. They convincingly show that neither the general public nor the patient perspective is without flaws. However, there may be different ways to compromise between the two perspectives. One is to require members of the general population to be more informed about the health problems they assess (Brazier et al., 2005). Another variant, originating from deliberative democracy, is based on the idea that arguments are exchanged in an open and transparent manner. Finally, Versteegh and Brouwer (2016) advocated for taking both perspectives into account separately. In the case of agreement, it would only provide further validation. In the case of disagreement, it would provide additional - and arguably valuable - information to the decision maker.

Further theoretical considerations

In addition to the two main groups (general public and patients), other stakeholder preferences could be considered relevant, including carers, health professionals, and policy makers. Children and adolescents are a special case, as they pose a number of conceptual and practical challenges for the current standard approach for valuing health states: if a generic instrument like the EQ-5D is used to assess HRQoL in children, the domains may not be applicable. If, however, a child specific instrument is used, it causes inconsistencies in the perspective. Children themselves may not be able to form and express well-informed preferences, and so preferences for child-specific health states are often elicited from adults. Yet, in that situation, respondents then automatically imagine another person (which happen to be a child) in that state. In fact, the current guideline for the EQ-5D-Y specifies that respondents should imagine ‘a 10-year-old child’. This signifies a change in perspective, from self/first-person, to other/third-person. Similar issues might arise from instruments specifically designed for an elderly population (e.g. ICECAP-O). It remains unclear how these inconsistencies could be reconciled (Rowen et al., 2020).

3. Additional degrees of freedom

Apart from the two questions discussed above (what to value and whose values to elicit), there are several other important, essentially normative choices that need to be made when deriving a social value set. It is beyond the scope of this thesis to identify and discuss them all, but, in what follows, three important choices are briefly outlined, and they are discussed in more detail in later chapters. These are: How to elicit preferences? How to scale preferences? And how to aggregate preferences across individuals?

How to elicit preferences?

There are many methods which can and have been used to elicit health preferences: Soekhai et al. (2019b) conducted a systematic literature review and identified a total of 19 different methods, of which five were ‘exploration’, i.e. qualitative, and 14 were ‘elicitation’, i.e. quantitative methods. In the context of HTA, the most commonly used methods were time trade-off (TTO), standard gamble (SG), and discrete choice experiment (DCE).

SG is supposedly the only technique compatible with von Neumann-Morgenstern expected utility theory, as it elicits preferences under conditions of uncertainty (Brazier et al., 2017b, p. 53). On a practical level, however, SG is often considered difficult to apply.

TTO was then developed by Torrance (Torrance et al., 1970), as an alternative, which would supposedly produce similar results, but that was easier to apply. There have been (ex-post) attempts to fit the TTO into the theoretical utility framework, but studies have consistently shown that individuals often make choices that do not adhere to underlying assumptions, e.g. due to biases arising from time preferences, loss aversion, constant proportional trade-offs, or maximal endurable time (Bleichrodt et al., 2003; Buckingham and Devlin, 2006; Dolan and Stalmeier, 2003; Lugnér and Krabbe, 2020).

DCE was initially developed by Louviere and Hensher (Louviere and Hensher, 1982) in the context of transport and market research. The theoretical foundation of DCE is the random utility theory of McFadden (McFadden, 1981). As in TTO and

SG, it is assumed that people have a utility function which maps the attributes of a health state to a utility value. Yet, the evaluation is assumed to be a stochastic process, with some (random) errors.

Another interesting method worth mentioning is the person trade-off (PTO) technique, developed by Nord (1994). PTO places the respondent in the position of a social decision maker, who has to choose among a series of alternative health care interventions. The method involves the trade-offs between the number of people that may benefit from the intervention, and the type of health improvement they receive. PTO would seem to align much better with the notion of the 'taxpayer perspective' than TTO, SG, or DCE. Unfortunately, however, PTO has rarely been widely used in practice.

The four methods discussed above are all choice-based (i.e. respondents have to consider trade-offs). This type of method can be contrasted with 'choice-less' methods, such as visual analogue scales (VAS), where respondents rate a certain health state, or attribute directly. VAS are relatively easy to apply in a wide range of settings. However, they 'lack a theoretical foundation' and have no basis in economic theory (Nord, 1991). A noteworthy rebuttal to this common criticism has been made by Parkin and Devlin (2006), who argued that the theoretical appeal of SG and other choice-based methods lie only in individual-level applications. On the societal level, and especially within an extra-welfarist framework, choice-based methods are not necessarily superior, as choices are not (only) based on individual utilities, but can take other factors into account.

An important point to note is that all methods do – or at least can – yield different values for the same health state. This applies not only to comparisons between different methods; even different variations of the same method (e.g. classical TTO vs. lead-time TTO) can yield different results (Brazier et al., 2017b, p.

72). At the same time, there does not seem to be any consensus on which method is best: there is simply no gold standard. Interestingly, in practice researchers usually just follow the convention of the instrument they seek to value. For example, the EQ-5D-5L is usually valued using the EQ-VT protocol, i.e. TTO with or without DCE (Devlin et al., 2022), and the SF-6D is valued using SG (Wang and Poder, 2023). Given the significance of the values derived from the elicitation of health state utilities for the calculation of QALYs, a better understanding of the strengths and weaknesses of each method, and clearer guidance on how to choose a method, would be desirable.

How to scale individual utility values?

By definition, utilities for estimating QALYs are anchored at full health, set to 1, and being dead, set to 0. This has been called the zero condition, and it is widely accepted, even though the paper from which it originates actually only shows that the origin of the utility scale needs to be anchored at a common point; zero is simply taken as the most intuitive and practical choice (Miyamoto et al., 1998) (also see chapter 3).

Far more controversial is the question of how to deal with negative values assigned to states worse than dead. Beside empirical anomalies, namely the absence of a clear relationship between severity and utilities for states worse than dead (Gandhi et al., 2019), there are significant conceptual problems with negative utilities. In contrast to their negative counterparts, negative utilities do not have a lower bound. They can range from 0 to minus infinity, which can cause considerable problems when aggregating utilities, as the resulting mean value set will be heavily influenced by the few respondents who assign very low negative values. Some researchers have dealt with this problem by rescaling negative values to have a lower limit of -1 (without giving any theoretical justification). Other researchers have applied an experimental design that does not allow for values below -1 (e.g. lead-time TTO).

How to aggregate individual preferences?

MacKillop et al. (2018) report that the decision to aggregate individual preferences using the unweighted arithmetic mean instead of medians or other (weighted) measure of central tendency can be traced back to a meeting of the QoL measurement steering group in 1995, involving Alan Williams, where it was decided that "tariffs should be based on means not medians". As far as I can tell, notes from the meeting are not publicly available, and no rationale or theoretical justification for this decision has been provided. With few notable exceptions (Devlin et al., 2019; Dewitt et al., 2017; Dewitt and Torrance, 2020), the problem of preference aggregation has in fact been largely ignored in the literature on health state valuation, whereas in the welfare economics and social choice theoretical literature (see above), it has been a critical topic of debate for decades (i.e. under which conditions is what type of aggregation appropriate?). Chapters 3 and 4 will discuss this topic in more detail.

DISCUSSION

In order to design a health valuation study, one should ideally be able refer to first principles, i.e. go back to a normative theory and derive an internally consistent approach, or at least get some clear guidance on whom to elicit values from, which elicitation method to use, and how to aggregate individual preferences, etc (Peasgood et al., 2021). Unfortunately, no such principles presently exist. My review of the normative foundations of health state valuation shows that there is no coherent normative theory that could justify the choices made in practice. Extra-welfarism, the framework most commonly referenced in the literature, appears underdetermined: it does not prescribe any particular approach or method, but leaves much room for interpretation and discretion. Other frameworks, such as communitarianism, are not well articulated, or, in the case if welfarism, not in line with current practices.

Furthermore, the overview of the (normative) choices that need to be made when designing a health valuation study (What is being valued? Whose values are being elicited? How are values elicited?) has further shown that there are many additional degrees of freedom not explicitly addressed by the theoretical health economics literature. The ways in which these choices are made in practice are often not well documented or justified.

In fact, the NICE reference case – defined by using the EQ-5D instrument, eliciting preference from the general public, using TTO (+/- DCE), and aggregating preferences using the unweighted arithmetic mean – seems merely one of many plausibly defensible approaches. In the absence of a well-articulated normative framework, one could even question whether the reference case was indeed the result of a long and thorough process of deliberation, or rather the result of a number of historical contingencies. That this framework has been used for many years and has been broadly adopted by many other HTA agencies should not be taken as a sign of its normative superiority. Alternative approaches may be equally defensible and may even be considered more appropriate in certain contexts.

Overall, health state valuation seems to be under-theorised. The absence of any gold standard for valuing health makes it difficult to assess the validity of a particular valuation method and poses a challenge to defending the legitimacy of any particular approach at all. Ways to better inform the normative debate in a systematic manner are discussed below.

A meta-theoretical framework

As argued above, I take the position that the normative foundations of health state valuation are not well articulated. There seems to be no conclusive justification for current practices. My review identified four types of arguments used to justify a particular approach to health state valuation.

1. Ontological arguments: Starting from a certain assumption about the nature of what is to be valued, the 'it', a number of implications can be derived. If, for example, one takes the view that health policy should maximise personal utility, the preferences of those most affected by a policy decision should be prioritised, and an instrument that captures all relevant dimensions, not just health or HRQoL, should be used. Psychometric testing could then potentially be used to assess to what extent the approach is successful in capturing the target construct.

2. Epistemological arguments: If the underlying concept *is* well defined, one can ask whose assessment is most accurate or best informed? Some argue, for example, that patients suffering from a particular condition or ailment

may have more information about what a particular health state feels like than a professional in the field or the general public, but their assessment may be biased due to adaptation to their condition. Members of the general public, on the other hand, may have less information, but their assessment may be more objective. These arguments refer to epistemological bases for making relevant judgements. Cognitive burden, respondent engagement and framing effects also fall into this category.

3. Deontological/procedural arguments: Irrespective of the nature of *what* is to be valued, one may consider the properties of the method for deriving a value set itself. The frequently mentioned 'taxpayer approach' an example. It is maintained that, since the taxpayer is the one who pays for it, it is only fair that they should have a say in how the money is spent in the health care sector. Note that there is no particular concept assumed to underlie the allocation, but, rather, it appeals to a notion of a fair or democratic decision-making process. Arguments could also be made for the use of patient preferences, e.g. arguing that since patients are the ones who are affected by a policy decision, their preferences should be given due consideration.

4. Practical considerations: Besides normative arguments, there are, of course, many practical aspects that must be considered when implementing a health valuation method in the real world. These include, for example, the availability of data, resources, potential respondents, or administrative feasibility, etc.

These four categories are not mutually exclusive. If, for example, one believes that a value set should reflect health, as a primary good, one can also argue that it should be valued by a citizen jury, using a deliberative approach. In most cases, assumptions about the ‘it’, or any other aspects, will not yield a unique methodological approach but rather a set of constraints.

This meta-framework is a work in progress, and does not provide guidance on how to choose between different approaches. It may only serve as a starting point for a more systematic discussion of the normative foundations of health state valuation. However, it can potentially help highlight main areas of agreement and disagreement and/or identify future research needs.

Ways forward

There are so many conflicting theories and paradigms – and reasonable people can disagree on any number of them – that it does not seem feasible to reach consensus on a single approach. Even if it was agreed that, say, wellbeing should be the ultimate goal of health policy, there would still be many different theories of wellbeing to choose from.

One way forward may be to elicit meta-preferences, e.g. preferences about whose preferences should count or how decisions should be made. An interesting contribution in this respect was recently made by Powell et al. (Powell et al., 2022), who found that members of the UK general public did not seem to support the use of their own preferences to inform health care resource allocation. This creates a paradox, which seems to counter the idea of the ‘taxpayer approach’ that is often used to justify the use of general public preferences in health state valuation. Similarly, Dewitt et al. (2017) saw meta-preferences as a way to reconcile different views on how individual preferences should be aggregated. This ‘meta-approach’ does, however, suffer the same limitations: In order to derive criteria (e.g. meta preferences) for selecting a particular approach for valuing health, one first needs to determine a method for deriving these criteria, which in turn requires criteria, etc. This is a fundamental problem of circularity, inherent to any decision-making framework (Triantaphyllou and Mann, 1989).

Recognising the problem of 'reasonable disagreement' and the inability to reach a consensus on what the goals and methods of policy making in the real world should be, Daniels (2010) proposed a potentially less ambitious goal: to develop a framework for legitimate policy making, centred on the premise that "establishing a fair process for priority setting is easier than agreeing on principles". Although it was originally developed for the allocation of health care resources in general, it could also be applied to the valuation of health states. The framework, called 'Accounting for Reasonableness' seeks to establish a fair deliberative process comprising four conditions (Daniels, 2010):

1. Decisions and their rationales must be publicly accessible.
2. The rationales should provide reasonable explanations.
3. Revision and appeals must be possible.
4. There is public oversight to ensure that 1-3 are met.

'Accounting for Reasonableness' provides a framework for evaluating the legitimacy of the process of choosing a particular approach. If applied to the NICE reference case, it may not be immediately clear whether decisions and their rationale are accessible to the (lay) public. In particular, the emphasis on mathematical modelling and psychometric testing seems problematic in this regard, because it may give the impression that health valuation is essentially a technical exercise. This conceals or at least distracts from the important underlying normative choices and assumptions. Greater transparency would undoubtedly be useful more generally. Researchers working in the field of health state valuation should be more explicit about the assumptions and values that underlie the value sets they develop.

The main lesson to draw from the discussion above, I think, is the benefit of a more pluralistic approach. Attempting a coherent normative foundation for health state valuation seems futile, as do attempts to define a single best approach to value health. Referring back to the very beginning of this chapter, I would like to point out that health valuation is part of a larger HTA process, which involves deliberation and reflection on various aspects of the intervention under consideration (Charlton, 2022; NICE, 2020). Therefore, it would not seem unreasonable to expand the scope of these discussions to also include multiple perspectives when valuing health benefits. Rather than insisting on a single reference case for the sake of consistency, it would be possible to assess the sensitivity of the results to different assumptions and approaches, reflecting the plurality of views and preferences that exist in society. To me, this seems more in line with policy decision making in democratic societies, where different views are considered and reflected upon (Baker et al., 2021).

What is considered the most appropriate approach to value health may well depend on the context of the decision, the type of intervention, and the population being affected, etc. However, even if policy makers do primarily consider the preferences of the general public the most important perspective, it does not necessarily follow that they should be the only perspective that is considered. In fact, it may be desirable to assess the value of the health benefits from the patients' and potentially other perspectives, as well. Carers, doctors, nurses, health care providers, and even policy makers themselves may have some claims by virtue of their expertise, experience, position, or trust (Culyer, 1989). Similarly, while EQ-5D provides a consistent measurement across different disease areas, more narrow instruments, that capture specific aspects of a particular disease, or wider instruments, that capture outcomes beyond health, might augment the available value information and provide a more complete picture of the (health) benefits derived from a particular intervention.

In some cases, the assessment will be the same or very similar, irrespective of the approach. In other cases, however, the results may differ. Either way, a more pluralistic approach may give policy makers additional useful information. In the absence of any compelling normative justification for only using one particular approach, policy makers should at least be open to acknowledging the value of other approaches. Consistency alone does, in any case, not seem to be a sufficient reason to ignore and dismiss the potentially large benefits as measured from other perspectives, by other instruments, or using other methods.

Limitations

This review suffers from major limitations that need to be considered when interpreting the presented conclusions and recommendations. First, this chapter is neither a comprehensive review of the health valuation literature – which would be an impossible task in the context of this thesis –, nor are the issues discussed in a systematic way. The review is based on a very selective literature search, which may have missed relevant articles, and thereby omitting important concepts, paradigms, and theories. The chapter also does not attempt to provide an objective or neutral analysis. The narrative form of the review is, to some extent, reflective of the author's personal views. Secondly, for the purpose of this review, the QALY model was taken as given. Yet, it is by no means the only framework that can be used to value health. There are many other frameworks (e.g. willingness to pay, value of a statistical life, subjective health, wellbeing) based on different theories, paradigms and assumptions. However, in the context of HTA in the UK, the QALY is – despite its many limitations – the most dominant, and therefore relevant, framework for this thesis, and it is for this reason I have focused this review on it.

Closing remarks

None of the issues with the normative foundations of health economics discussed above are new. They have been known and debated for decades. In a paper by Culyer from 1989 entitled ‘the normative economics of health care finance and provision’, he already noted the ambiguity inherent to extra-welfarism, and the open questions around how to determine what should be valued, whose values should count, and how to decide who should decide these questions (Culyer, 1989). Culyer concluded: “The heady atmosphere of grand designs has to be replaced by the mundane, but ultimately more fruitful, ground of systematically applied economics.” Looking back, one could argue that the mundane application of economic methods went too far. Now might be the time to pause, reflect and reconsider some of those choices that have been made 20 odd years ago, and to consciously decide whether they are (still) appropriate in light of new insights and developments in the field This may also open up the discussion to a wider audience, taking into account a broader, more diverse range of voices and perspectives.

REFERENCES

Ackoff RL, Sasieni MW. *Fundamentals of Operations Research.* New York: John Wiley & Sons; 1968.

Adler MD, Posner EA. Rethinking cost-benefit analysis. *The Yale Law Journal.* 1999;109: 165.

Al Sayah F, Jin X, Johnson JA. Selection of patient-reported outcome measures (PROMs) for use in health systems. *Journal of Patient-Reported Outcomes.* 2021;5: 1–5.

Alava MH, Wailoo A, Pudney S, Gray L, Manca A. Mapping clinical outcomes to generic preference-based outcome measures: Development and comparison of methods. *Health Technology Assessment.* (Winchester, England) 2020;24: 1.

Al-Janabi H, N Flynn T, Coast J. Development of a self-report measure of capability wellbeing for adults: The ICECAP-a. *Quality of Life Research.* 2012;21: 167–76.

Arrow KJ. *Collected Papers of Kenneth J. Arrow, Volume 1: Social Choice and Justice.* Cambridge, Massachusetts: Harvard University Press; 1983.

Baker R, Mason H, McHugh N, Donaldson C. Public values and plurality in health priority setting: What to do when people disagree and why we should care about reasons as well as choices. *Social Science & Medicine.* 2021;277: 113892.

Barberà S, Hammond P, Seidl C. *Handbook of Utility Theory: Volume 2 Extensions.* Springer Science & Business Media; 2004.

Belton V, Stewart T. *Multiple criteria decision analysis: An integrated approach.* Berlin/Heidelberg, Germany: Springer Science & Business Media; 2002.

Bentham J. *An Introduction to the Principles of Morals and Legislation.* London: Many Editions; 1789.

Bergson A. A Reformulation of Certain Aspects of Welfare Economics. *The Quarterly Journal of Economics.* 1938;52(2): 310-334.

Bleichrodt H, Pinto JL, Abellan-Perpiñan JM. A consistency test of the time trade-off. *Journal of Health Economics.* 2003;22: 1037–52.

Bleichrodt H, Quiggin J. Life-cycle preferences over consumption and health: When is cost-effectiveness analysis equivalent to cost–benefit analysis? *Journal of Health Economics.* 1999;18: 681–708.

Brazier J. *Current state of the art in preference-based measures of health and avenues for further research.* HEDS Discussion Paper 05/05; 2005.

Brazier J, Akehurst R, Brennan A, Dolan P, Claxton K, McCabe C, et al. Should patients have a greater role in valuing health states? *Applied Health Economics and Health Policy.* 2005;4: 201–8.

Brazier J, Ara R, Rowen D, Chevrou-Severac H. A review of generic preference-based measures for use in cost-effectiveness models. *Pharmacoeconomics*. 2017a;35: 21–31.

Brazier J, Deverill M. A checklist for judging preference-based measures of health related quality of life: Learning from psychometrics. *Health Economics.* 1999;8: 41–51.

Brazier J, Peasgood T, Mukuria C, Marten O, Kreimeier S, Luo N, et al. The EQ health and wellbeing: Overview of the development of a measure of health and wellbeing and key results. *Value in Health*; 2022.

Brazier J, Ratcliffe J, Saloman J, Tsuchiya A. *Measuring and Valuing Health Benefits for Economic Evaluation*. Oxford: Oxford University Press; 2017b.

Brazier J, Tsuchiya A. Improving cross-sector comparisons: Going beyond the health-related QALY. *Applied Health Economics and Health Policy.* 2015;13: 557–65.

Brouwer WB, Culyer AJ, Exel NJA van, Rutten FF. Welfarism vs. Extra-welfarism. *Journal of Health Economics.* 2008;27: 325–38.

Brouwer WB, Koopmanschap MA. On the economic foundations of CEA. Ladies and gentlemen, take your positions! *Journal of Health Economics.* 2000;19: 439–59.

Buchanan JM. Positive economics, welfare economics, and political economy. *The Journal of Law and Economics.* 1959;2: 124–38.

Buckingham K, Devlin N. A theoretical framework for TTO valuations of health. *Health Economics.* 2006;15: 1149–54.

Burström K, Teni FS, Gerdtham U-G, Leidl R, Helgesson G, Rolfson O, et al. Experience-based swedish TTO and VAS value sets for EQ-5D-5L health states. *Pharmacoeconomics.* 2020;38: 839–56.

Bush JW, Fanshel S, Chen MM. Analysis of a tuberculin testing program using a health status index. *Socio-Economic Planning Sciences.* 1972;6: 49–68.

Charlton V. The normative grounds for NICE decision-making: A narrative cross-disciplinary review of empirical studies. *Health Economics, Policy and Law.* 2022: 1–27.

Coast J. Is economic evaluation in touch with society's health values? *The British Medical Journal.* 2004;329: 1233–6.

Colander D. Edgeworth's hedonimeter and the quest to measure utility: retrospectives. *Journal of Economic Perspectives.* 2007;21: 215–25.

Cookson R, Claxton K, et al. *The humble economist: Tony Culyer on health, health care and social decision making.* Monographs; 2012.

Cookson R, Skarda I, Cotton-Barratt O, Adler M, Asaria M, Ord T. Quality adjusted life years based on health and consumption: A summary wellbeing measure for cross-sectoral economic evaluation. *Health Economics.* 2021;30: 70–85.

Craig BM, Reeve BB, Brown PM, Cella D, Hays RD, Lipscomb J, et al. US valuation of health outcomes measured using the PROMIS-29. *Value in Health.* 2014;17: 846–53.

Culyer AJ. The ideas and influence of Alan Williams: Be reasonable—do it my way. *Proceedings of a Conference to Celebrate the Work of Alan Williams.* London: Office of Health Economics; 2007. p.57–74.

Culyer AJ. The normative economics of health care finance and provision. *Oxford Review of Economic Policy.* 1989;5: 34–58.

Daniels N. Capabilities, opportunity, and health. *Measuring Justice: Primary Goods and Capabilities.* 2010: 131–49.

Davies C, Wetherell M, Barnett E. *Citizens at the Centre.* Bristol: Policy Press; 2006.

Devlin NJ, Brooks R. EQ-5D and the EuroQol group: Past, present and future. *Applied Health Economics and Health Policy.* 2017;15: 127–37.

Devlin NJ, Shah KK, Mulhern BJ, Pantiri K, Hout B van. A new method for valuing health: Directly eliciting personal utility functions. *The European Journal of Health Economics.* 2019;20: 257–70.

Devlin N, Roudijk B, Ludwig K. *Value sets for EQ-5D-5L: A compendium, comparative review & user guide.* 2022.

Dewitt B, Davis A, Fischhoff B, Hanmer J. An approach to reconciling competing ethical principles in aggregating heterogeneous health preferences. *Medical Decision Making.* 2017;37: 647–56.

Dewitt B, Torrance GW. Incorporating mortality in health utility measures. *Medical Decision Making.* 2020;40: 862–72.

Dolan P. The nature of individual preferences: A prologue to Johannesson, Jonsson and Karlsson. *Health Economics.* 1997;6: 91–3.

Dolan P, Green C. Using the person trade-off approach to examine differences between individual and social values. *Health Economics.* 1998;7: 307–12.

Dolan P, Stalmeier P. The validity of time trade-off values in calculating QALYs: Constant proportional time trade-off versus the proportional heuristic. *Journal of Health Economics.* 2003;22: 445–58.

Driver J. *The History of Utilitarianism.* Stanford Encyclopaedia of Philosophy; 2010.

Elster J, Przeworski A. *Deliberative Democracy.* Cambridge: Cambridge University Press; 1998.

Fanshel S, Bush JW. A health-status index and its application to health-services outcomes. *Operations Research.* 1970;18: 1021–66.

Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Medical Care.* 2002;40: 113–28.

Finch AP, Brazier JE, Mukuria C. What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. *The European Journal of Health Economics.* 2018;19: 557–70.

Fischhoff B. Value elicitation: Is there anything in there? *American Psychologist.* 1991;46: 835.

Gandhi M, Rand K, Luo N. Valuation of health states considered to be worse than death—an analysis of composite time trade-off data from 5 EQ-5D-5L valuation studies. *Value in Health.* 2019;22: 370–6.

Gansen F, Klinger J. Reasoning in the valuation of health-related quality of life: A qualitative content analysis of deliberations in a pilot study. *Health Expectations.* 2020;23: 405–13.

Garber AM, Phelps CE. Economic foundations of cost-effectiveness analysis. *Journal of Health Economics.* 1997;16: 1–31.

Gold M. Panel on cost-effectiveness in health and medicine. *Medical Care.* 1996. DS197–9.

Goodwin E, Green C. A systematic review of the literature on the development of condition-specific preference-based measures of health. *Applied Health Economics and Health Policy.* 2016;14: 161–83.

Group M. *The measurement and valuation of health: Final report on the modelling of valuation tariffs.* York: Centre for Health Economics, University of York; 1995.

Harsanyi JC. Assessing other people's utilities. *Risk, Decision and Rationality.* 1988:127–38.

Hausman DM. *Valuing Health: Well-Being, Freedom, and Suffering.* Oxford: Oxford University Press; 2015.

Hausman DM. Valuing health: A new proposal. *Health Economics.* 2010;19: 280–96.

Hays R, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. *Quality of Life Research.* 1993;2: 441–9.

Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *Quality of Life Research.* 2009;18: 873–80.

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. *Quality of Life Research.* 2020;29: 1465–82.

Hernández-Alava M, Pudney S, Wailoo A. *Quality review of a proposed EQ-5D-5L value set for England.* EEPRU Report: [Online], 2018.

Himmler S, Soekhai V, Exel J van, Brouwer W. What works better for preference elicitation among older people? Cognitive burden of discrete choice experiment and case 2 best-worst scaling in an online setting. *Journal of Choice Modelling.* 2021;38: 100265.

Horsman J, Furlong W, Feeny D, Torrance G. The health utilities index (HUI): Concepts, measurement properties and applications. *Health and Quality of Life Outcomes.* 2003;1: 1–13.

Huber M, Knottnerus JA, Green L, Van Der Horst H, Jadad AR, Kromhout D, et al. How should we define health? *The British Medical Journal.* 2011; 343.

Jonker MF, Donkers B, Bekker-Grob EW de, Stolk EA. Effect of level overlap and color coding on attribute non-attendance in discrete choice experiments. *Value in Health.* 2018;21: 767–71.

Kaldor N. Welfare propositions of economics and interpersonal comparisons of utility. *The Economic Journal.* 1939;49: 549–52.

Karimi M, Brazier J. Health, health-related quality of life, and quality of life: What is the difference? *Pharmacoeconomics*. 2016;34: 645–9.

Karimi M, Brazier J, Paisley S. Effect of reflection and deliberation on health state values: A mixed-methods study. *Value in Health.* 2019;22: 1311–7.

Keeney RL, Raiffa H, Meyer RF. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* Cambridge: Cambridge University Press; 1993.

Kennedy-Martin M, Slaap B, Herdman M, Reenen M van, Kennedy-Martin T, Greiner W, et al. Which multi-attribute utility instruments are recommended for use in cost-utility analysis? A review of national health technology assessment (HTA) guidelines. *The European Journal of Health Economics.* 2020;21: 1245–57.

Kind P. In memoriam: Alan Williams (1928-2005). *Value in Health.* 2005;8: 615.

Klarman HE, Rosenthal GD. Cost effectiveness analysis applied to the treatment of chronic renal disease. *Medical Care.* 1968;6: 48–54.

Knight J, Johnson J. Aggregation and deliberation: On the possibility of democratic legitimacy. *Political Theory.* 1994;22: 277–96.

Leidl R, Reitmeir P. A value set for the EQ-5D based on experienced health states: Development and testing for the german population. *Pharmacoeconomics*. 2011;29: 521–34.

Longworth L, Yang Y, Young T, Mulhern B, Hernández Alava M, Mukuria C, et al. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: A systematic review, statistical modelling and survey. *Health Technology Assessment.* 2014.

Lorgelly PK, Doble B, Rowen D, Brazier J. Condition-specific or generic preference-based measures in oncology? A comparison of the EORTC-8D and the EQ-5D-3L. *Quality of Life Research.* 2017;26: 1163–76.

Louviere JJ, Hensher DA. On the design and analysis of simulated choice or allocation experiments in travel choice modelling. *Transportation Research Record.* 1982;890: 11–7.

Ludwig K, Ramos-Goñi JM, Oppe M, Kreimeier S, Greiner W. To what extent do patient preferences differ from general population preferences? *Value in Health.* 2021;24: 1343–9.

Lugnér AK, Krabbe PF. An overview of the time trade-off method: Concept, foundation, and the evaluation of distorting factors in putting a value on health. *Expert Review of Pharmacoeconomics & Outcomes Research.* 2020;20: 331–42.

MacKillop E, Sheard S. Quantifying life: Understanding the history of quality-adjusted life-years (QALYs). *Social Science & Medicine.* 2018;211: 359–66.

Matza LS, Stewart KD, Lloyd AJ, Rowen D, Brazier JE. Vignette-based utilities: Usefulness, limitations, and methodological recommendations. *Value in Health.* 2021;24: 812–21.

McCabe C, Claxton K, Culyer AJ. The NICE cost-effectiveness threshold: What it is and what that means. *Pharmacoeconomics.* 2008;26: 733–44.

McFadden D. Econometric models of probabilistic choice. *Structural Analysis of Discrete Data with Econometric Applications.* 1981; 198272.

McNamara S, Holmes J, Stevely AK, Tsuchiya A. How averse are the UK general public to inequalities in health between socioeconomic groups? A systematic review. *The European Journal of Health Economics.* 2020;21: 275–85.

McNamee P, Seymour J. Comparing generic preference-based health-related quality-of-life measures: Advancing the research agenda. *Expert Review of Pharmacoeconomics & Outcomes Research.* 2005;5: 567–81.

Menzel P, Dolan P, Richardson J, Olsen JA. The role of adaptation to disability and disease in health state valuation: A preliminary normative analysis. *Social Science & Medicine.* 2002;55: 2149–58.

Miyamoto JM, Wakker PP, Bleichrodt H, Peters HJ. The zero-condition: A simplifying assumption in QALY measurement and multiattribute utility. *Management Science.* 1998;44: 839–49.

Mooney G. "Communitarian claims" as an ethical basis for allocating health care resources. *Social Science & Medicine.* 1998;47: 1171–80.

Mukuria C, Rowen D, Harnan S, Rawdin A, Wong R, Ara R, et al. An updated systematic review of studies mapping (or cross-walking) measures of health-related quality of life to generic preference-based measures to generate utility values. *Applied Health Economics and Health Policy.* 2019;17: 295–313.

Musgrave RA. *The Theory of Public Finance: A Study in Public Economy.* New York City: McGraw-Hill; 1959.

National Institute for Health and Care Excellence (NICE), NICE health technology evaluations: The manual-process and methods. 2022. https://www.nice.org.uk/process/pmg36 [Accessed 20th January 2023].

National Institute for Health and Care Excellence (NICE), *CHTE methods review. Health-related quality of life. Task and finish group report.* 2020. https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/nice-guidance/chte-methods-consultation/Health-related-quality-of-life-task-and-finish-group-report.docx [Accessed 8th February 2023].

National Institute for Health and Care Excellence (NICE), *Quality-adjusted Life Year.* [NICE Glossary]. 2016. https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisals/Brouwer-advice-EQ-5D-5L-valuation.pdf.

Netten A, Burge P, Malley J, Potoglou D, Towers A-M, Brazier J, et al. Outcomes of social care for adults: Developing a preference-weighted measure. *Health Technology Assessment.* 2012;16: 1–166.

Nord E. The person trade-off approach to valuing health care programs. *Medical Decision Making.* 1995 Jul-Sep;15(3): 201-8.

Nord E. The validity of a visual analogue scale in determining social utility weights for health states. *The International Journal of Health Planning and Management.* 1991;6: 234–42.

Nussbaum M, Sen A. *The Quality of Life.* Oxford: Oxford University Press; 1993.

Parkin D, Devlin N. Is there a case for using visual analogue scale valuations in cost-utility analysis? *Health Economics.* 2006;15: 653–64.

Payakachat N, Ali MM, Tilford JM. Can the EQ-5D detect meaningful change? A systematic review. *Pharmacoeconomics.* 2015;33: 1137–54.

Peasgood T, Mukuria C, Carlton J, Connell J, Devlin N, Jones K, et al. What is the best approach to adopt for identifying the domains for a new measure of health, social care and carer-related quality of life to measure quality-adjusted life years? Application to the development of the EQ-HWB? *The European Journal of Health Economics.* 2021;22: 1067–81.

Pliskin JS, Shepard DS, Weinstein MC. Utility functions for life years and health status. *Operations Research.* 1980;28: 206–24.

Powell PA, Karimi M, Rowen D, Devlin N, Hout B van, Brazier JE. Hypothetical versus experienced health state valuation: A qualitative study of adult general public views and preferences. *Quality of Life Research.* 2022: 1–11.

Rabin R, Gudex C, Selai C, Herdman M. From translation to version management: A history and review of methods for the cultural adaptation of the EuroQol five-dimensional questionnaire. *Value in Health.* 2014;17: 70–6.

Robbins L. Economics and political economy. *The American Economic Review.* 1981;71: 1–10.

Robbins L. The nature and significance of economic science. *The Philosophy of Economics: An Anthology.* 1932;1: 73–99.

Roberts JT, Hite A, Chorev N (eds). *The Globalization and Development Reader: Perspectives on Development and Global Change.* Hoboken, NJ: Wiley-Blackwell; 2014.

Robinson LA, Hammitt JK, Chang AY, Resch S. Understanding and improving the one and three times GDP per capita cost-effectiveness thresholds. *Health Policy and Planning.* 2017;32: 141–5.

Ronen GM. Reflections on the usefulness of the term "health-related quality of life". *Developmental Medicine & Child Neurology.* 2017;59: 1105–6.

Rosser R, Kind P. A scale of valuations of states of illness: Is there a social consensus? *International Journal of Epidemiology.* 1978;7: 347–58.

Roudijk B, Janssen B, Olsen JA. *How do EQ-5D-5L value sets differ? Value sets for EQ-5D-5L: A compendium, comparative review & user guide.* Springer International Publishing Cham; 2022, p. 235–58.

Round J, Hawton A. Statistical alchemy: Conceptual validity and mapping to generate health state utility values. *PharmacoEconomics-Open.* 2017;1: 233–9.

Rowen D, Brazier J, Ara R, Azzabi Zouraq I. The role of condition-specific preference-based measures in health technology assessment. *Pharmacoeconomics.* 2017;35: 33–41.

Rowen D, Rivero-Arias O, Devlin N, Ratcliffe J. Review of valuation methods of preference-based measures of health for economic evaluation in child and adolescent populations: Where are we now and where are we going? *Pharmacoeconomics.* 2020;38: 325–40.

Russell LB, Gold MR, Siegel JE, Daniels N, Weinstein MC. The role of cost-effectiveness analysis in health and medicine. *JAMA.* 1996;276: 1172–7.

Sampson C. Meeting round-up: 7th EuroQol academy meeting. *The Academic Health Economist's blog*. 2023. https://aheblog.com/2023/03/13/meeting-round-up-7th-euroqol-academy/. [Accessed 13th March 2023]

Sampson C. NICE and the EQ-5D-5L: Ten years trouble. *PharmacoEconomics-Open.* 2022;6: 5–8.

Schneider P. The QALY is ableist: On the unethical implications of health states worse than dead. *Quality of Life Research.* 2022;31: 1545–52.

Sen A. *Collective Choice and Social Welfare.* Cambridge, Massachusetts: Harvard University Press; 2018.

Sen A. *Development as Freedom.* Oxford: Oxford University Press; 1999.

Sen A. Rationality and social choice. *The American Economic Review.* 1995;85: 1.

Shah KK, Tsuchiya A, Wailoo AJ. Valuing health at the end of life: A stated preference discrete choice experiment. *Social Science & Medicine.* 2015;124: 48–56.

Shiroiwa T, Ikeda S, Noto S, Igarashi A, Fukuda T, Saito S, et al. Comparison of value set based on DCE and/or TTO data: Scoring for EQ-5D-5L health states in Japan. *Value in Health.* 2016;19: 648–54.

Simmons SJ. Health: A concept analysis. *International Journal of Nursing Studies.* 1989;26: 155–61.

Slovic P. The construction of preference. *American Psychologist.* 1995;50: 364.

Soekhai V, Bekker-Grob EW de, Ellis AR, Vass CM. Discrete choice experiments in health economics: Past, present and future. *Pharmacoeconomics.* 2019a;37: 201–26.

Soekhai V, Whichello C, Levitan B, Veldwijk J, Pinto CA, Donkers B, et al. Methods for exploring and eliciting patient preferences in the medical product lifecycle: A literature review. *Drug Discovery Today.* 2019b;24: 1324–31.

Sun S, Chen J, Kind P, Xu L, Zhang Y, Burström K. Experience-based VAS values for EQ-5D-3L health states in a national general population health survey in china. *Quality of Life Research.* 2015;24: 693–703.

Sunstein CR. Incompletely theorized agreements. *Harvard Law Review.* 1994;108: 1733.

Thokala P, Devlin N, Marsh K, Baltussen R, Boysen M, Kalo Z, et al. Multiple criteria decision analysis for health care decision making—an introduction: Report 1 of the ISPOR MCDA emerging good practices task force. *Value in Health.* 2016;19: 1–13.

Thorndike EL. Valuations of certain pains, deprivations, and frustrations. *The Pedagogical Seminary and Journal of Genetic Psychology.* 1937;51: 227–39.

Torrance GW. A generalized cost-effectiveness model for the evaluation of health programs. Faculty of Business McMaster University Research Series. 1970;101.

Torrance GW, Boyle MH, Horwood SP. Application of multi-attribute utility theory to measure social preferences for health states. *Operations Research.* 1982;30: 1043–69.

Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions: Health utilities index. *Pharmacoeconomics.* 1995;7: 503–20.

Triantaphyllou E, Mann SH. An examination of the effectiveness of multi-dimensional decision-making methods: A decision-making paradox. *Decision Support Systems.* 1989;5: 303–12.

Tsuchiya A, Watson V. Re-thinking "the different perspectives that can be used when eliciting preferences in health." *Health Economics.* 2017;26: 103–7.

Vanberg VJ. Individual choice and social welfare: Theoretical foundations of political economy. *Freiburger Diskussionspapiere zur Ordnungsökonomik.* 2018.

Versteegh M, Brouwer W. Patient and general public preferences for health states: A call to reconsider current guidelines. *Social Science & Medicine.* 2016;165: 66–74.

Versteegh MM, Leunis A, Uyl-de Groot CA, Stolk EA. Condition-specific preference-based measures: Benefit or burden? *Value in Health.* 2012;15: 504–13.

Wang L, Poder TG. A systematic review of SF-6D health state valuation studies. *Journal of Medical Economics.* 2023 Mar 27(just-accepted):1-31.

Weinstein MC, Torrance G, McGuire A. QALYs: The basics. *Value in Health.* 2009.

Whitehurst DG, Bryan S. Another study showing that two preference-based measures of health-related quality of life (EQ-5D and SF-6D) are not interchangeable. But why should we expect them to be? *Value in Health.* 2011;14: 531–8.

Williams A. Inequalities in health and intergenerational equity. *Ethical Theory and Moral Practice.* 1999;2: 47–55.

Williams A. Economics of coronary artery bypass grafting. *The British Medical Journal (Clinical Research Edition).* 1985;291: 326–9.

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.

World Health Organisation (WHO), Constitution of the World Health Organization. 1948. https://apps.who.int/gb/bd/pdf_files/BD_49th-en.pdf [Accessed 8th May 2023]

Part I

Normative issues in the valuation of health

‘When I use a word,’ Humpty Dumpty said in rather a scornful tone, ‘it means just what I choose it to mean — neither more nor less.’

’The question is,’ said Alice, ‘whether you can make words mean so many different things.’

’The question is,’ said Humpty Dumpty, ‘which is to be master — that’s all.’

– Lewis Caroll, Alice in Wonderland

Chapter 2

Social tariffs and democratic choice

The following chapter examines the implications of using a social tariff to value health states in terms of QALYs, for democratic participation. It suggests that social tariffs should not only be seen as a technical problem, but should also be recognised as an important instrument of democracy. Possible implications for the aggregation of individual preferences and the selection of individuals whose preferences should count are discussed, as well as alternative tariff specifications and decision rules.

This chapter has been published in an identical form as:

Schneider PP. Social tariffs and democratic choice—Do population-based health state values reflect the will of the people?. Health Economics. 2021 Jan;30(1):104-12. https://doi.org/10.1002/hec.4179

Received: 19 August 2019 | Revised: 22 April 2020 | Accepted: 18 August 2020

DOI: 10.1002/hec.4179

RESEARCH ARTICLE

Health Economics WILEY

Social tariffs and democratic choice—Do population-based health state values reflect the will of the people?

Paul Peter Schneider

ScHARR, University of Sheffield, Sheffield, UK

Correspondence
Paul Peter Schneider, School of Health and Related Research, University of Sheffield, 30 Regent St, Sheffield S1 4DA, UK.
Email: p.schneider@sheffield.ac.uk

Funding information
Wellcome Trust

Abstract

In economic evaluations of health technologies, health outcomes are commonly measured in terms of quality-adjusted life years (QALYs). QALYs are the product of time and health-related quality of life. Health-related quality of life, in turn, is determined by a social tariff, which is supposed to reflect the public's preference over health states. This study argues that, because of the tariff's role in the societal decision-making process, it should not be understood as merely an operational (statistical) definition of health, but as a major instrument of democratic participation. I outline what implications this might have for both the method used to aggregate individual preferences, and the set of individuals whose preferences should count. Alternative tariff specifications and decision rules are explored, and future research directions are proposed.

KEYWORDS
conceptual model, decision-making, democracy, health state, normative theory, QALY, social choice, tariff, valuation

1 | INTRODUCTION

Societal decisions, on whether or not certain health programs should be publicly provided, are often informed by economic evaluations: The (additional) costs and health benefits of, say a new drug as compared against alternatives courses of action (e.g., another drug). The results are often summarized into an incremental cost-effectiveness ratio (ICER; Dakin et al., 2015). In England, as in many other countries, health effects are measured in quality-adjusted life years (QALYs), which are the product of length and health-related quality of life. The measurement of health-related quality of life, in turn, consists of a two components: a descriptive system of health states and a social tariff which maps these states to preference values.

The currently preferred instrument for valuing health outcomes in England is the UK social EQ-5D 3L tariff (MVH Group 1995; NICE, 2013). It is based on the preferences of (around 3000) members of the general public. When the tariff is applied in economic evaluations, it is supposed to incorporate societal (instead of patients') preferences into health policy decisions regarding the allocation of (publicly funded) health care resources (Whitehead & Ali, 2010). Therefore, I argue that the tariff should not be understood as merely an operational (statistical) definition of health, but as a major instruments of collective choice. As such, tariffs do not only have to adhere to scientific standards, but also need to reflect the norms and democratic principles of society as a whole.

Despite the considerable impact on health policy decision-making, the implied value judgments of social tariffs have received very little attention, and research into their conceptual and normative basis has been scarce (N. Devlin, Shah, & Buckingham, 2017; Dewitt, Davis, Fischhoff, & Hanmer, 2017). In this study, I make a first attempt to examine the role of the tariff within the wider decision-making framework from a collective choice perspective (section 2). I go on to highlight the (im)possibility of aggregating individual health state preferences into a societal preference (section 3) and outline further implications for health state valuation studies (section 4), before I propose future research directions (section 5).

2 | THE HEALTH POLICY DECISION-MAKING FRAMEWORK

In the following, I will provide a basic framework for economic evaluations, incorporating health state values from the general population. For clarity, some simplifications will be made: any uncertainty and discounting are being discarded; and we assume that allocative efficiency is the only relevant criterion for societal decision-making, ignoring any other consideration that may influence health policy in the real world (e.g., outcomes beyond health, approval regulations, or equity concerns). Moreover, it should be noted that the QALY is built on strong assumptions itself, including, among others, the measurability of interpersonally comparable, cardinal preferences over hypothetical health states (Carr-Hill, 1989; Dolan, 2000; Dolan, Shaw, Tsuchiya, & Williams, 2005; Fleurbaey & Hammond, 2004; Lipscomb, Drummond, Fryback, Gold, & Revicki, 2009). Challenging these assumptions is outside the scope of this study, and the function of the social tariff is only investigated within this given context.

2.1 | Basic notations and concepts

Suppose society consists of n individuals, whose preference functions over m (mutually exclusive) health states, given by $H=\{h_1,h_2,\cdots,h_m\}$, are denoted $p_1,p_2,\cdots,p_n$, with $p: H\rightarrow u, (u\in\mathbb{R}, u\leq 1)$. Note that preference values over health states are measured on a ratio scale, in relation to the preference for "full health," denoted h^*, with $p'(h_j)=\frac{p(h_j)}{p(h^*)}$. The societal value of health states is captured by the social tariff $t(.)$, which is an aggregate function of individual preference functions, given by $t(h_j)=f(p'_1(h_j),p'_2(h_j),\cdots,p'_n(h_j))$. Let $S=\{s_{11},s_{12},\cdots,s_{21},\cdots,s_{nm}\}$then denote an $n\times m$matrix containing individuals' 'Health States Times', that is, the amount of time that individual i spent in state j. If we assume additive separability and zero time preference, the number of QALYs (as valued by society) accrued by all members of society $Q^t=\{q_1^t,q_2^t,\cdots,q_n^t\}$is determined by the products of individuals' Health States Times and the corresponding societal valuation, given by $Q^t=t(H)\times S$. The total number of QALYs in society can be evaluated by the following formula:

$$\sum_{i=1}^{n} q_i^t=\sum_{i=1}^{n}\sum_{j=1}^{m} t\left(h_j\right)\times s_{ij}$$

2.2 | The role of the tariff in societal decision-making

The role of the social tariff in societal decision-making is outlined below, and Figure 1 provides a schematic overview of the framework (superscripts are used to link the text with the figure). The aim of the health system is assumed to be the maximization of QALYs, subject to a fixed budget constraint. The marginal opportunity costs of 1 QALY are further assumed to be θ—that is a marginal decrease in the health care spending by θ results in a 1 QALY loss. The societal decision$^{(12)}$ over some health program a thus depends on its ICER$^{(10)}$, compared to its most cost-effective alternative $\overline{a}$. While the incremental costs$^{(8)}$, given by $\Delta c_a=[c\mid a]-[c\mid\overline{a}]$, can, in principle, be directly observed in a study$^{(6)}$, the incremental QALYs$^{(9)}$ ΔQ_a^t are not only determined by the incremental health outcomes$^{(7)}$ $\Delta S_a=([S\mid a]-[S\mid\overline{a}])$, but also by the social tariff$^{(5)}$ $t(.)$, with $\Delta Q_a^t=t(H)\times\Delta S_a$. To derive t, however, first individual health state preferences$^{(3)}$ have to be elicited, for example using the time-trade-of method$^{(2)}$. Preferences are then aggregated$^{(4)}$, as specified by $f(.)$, before the tariff can be used to translate health outcomes into QALYs.

Finally, program a should be adopted if its ICER is smaller or equal to θ. The societal decision function$^{(11)}$ $W(.)$ can be defined as follows:

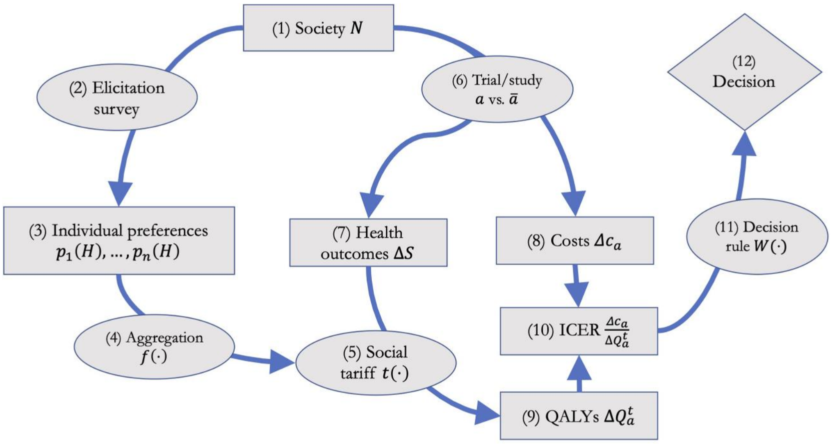
(1) Society N
(2) Elicitation survey
(6) Trial/study a vs. $\bar{a}$
(12) Decision
(3) Individual preferences $p_1(H), \ldots, p_n(H)$
(7) Health outcomes ΔS
(8) Costs Δc_a
(11) Decision rule $W(\cdot)$
(10) ICER $\frac{\Delta c_a}{\Delta Q_a^t}$
(4) Aggregation $f(\cdot)$
(5) Social tariff $t(\cdot)$
(9) QALYs ΔQ_a^t

FIGURE 1 The role of the social tariff within the wider decision-making framework. *Ovals represent functions and rectangles the inputs/outputs.* ICER, incremental cost-effectiveness ratio; QALYs, quality-adjusted life years [Colour figure can be viewed at wileyonlinelibrary.com]

$$W(a) = \begin{cases} 0 \; if \; \dfrac{\Delta c_a}{\Delta Q_a^t} > \theta \\[2ex] 1 \;\; if \;\; \dfrac{\Delta c_a}{\Delta Q_a^t} \leq \theta \end{cases}$$

The overview (see Figure 1) illustrates the central role of the social tariff $t(.)$ in the decision-making framework: the tariff specifies the societal value of the time individuals spend in any health state other than full health, and thereby, it determines to some, potentially great extent whether or not a health program is considered cost-effective. Depending on the distribution of preferences, the method of aggregation $f(.)$ can, thus, also have significant impact on societal decision-making.

2.3 | The social tariff as an instrument of collective choice

Before I go on to discuss the aggregation of individual preference functions, it will be useful to briefly consider what the resulting societal preference values represent. First of all, it should be noted that the current social tariff framework is fundamentally incompatible with the notion of utility maximization. This is because, even though the tariff is based on individual (health state) preferences, it is not individual i's own valuation of their own (actual or potential) health state (s) that informs societal decisions. Instead, a change in individual i's health from state j to state k is valued by the aggregate preference of society. Since individual i's preference will generally not be identical to the societal preference, $p'_i(h_j) \not\equiv t(h_j)$, it follows that maximizing societal QALYs is not the same as maximizing health-related utilities: $\sum_{i=1}^{n}\sum_{j=1}^{m} t(h_j) \times s_{ij} \not\equiv \sum_{i=1}^{n}\sum_{j=1}^{m} p'_i(h_j) \times s_{ij}$.

A more convincing interpretation of the QALY can be derived from 'extra-welfarism', which offers an alternative approach for the evaluation of health policies beyond utilities (Brouwer, Culyer, van Exel, & Rutten, 2008; Coast, Smith, & Lorgelly, 2008; Cookson, 2005; Culyer, 1989). Here, health is not primarily recognized as a source of utility, but it has a social value in itself. In fact, this is how the QALY seems to be generally understood: as an operational definition of *health*. Hence, it seems inadequate to define the social tariff as a statistical summary function of individual (health-related) utilities. Instead, it should be understood more broadly as a mechanism, through which society collectively derives an interpersonally comparable index of value for different sets of health functionings (Cookson, 2005).

3 | AGGREGATING INDIVIDUALS' HEALTH STATE PREFERENCES

3.1 | Problem statement

With only few exceptions (e.g., Shaw et al., 2010), health state valuation studies have used the arithmetic mean to aggregate individual preferences into a societal preference (Xie, Gaebel, Perampaladas, Doble, & Pullenayegum, 2014; MVH, 1995). If the tariff would reflect individuals' own, self-assessed (health-related) utilities, the use of the mean could potentially be justified by utilitarian welfare maximization through potential pareto improvements. But, as argued above, the current framework is incompatible with this interpretation of the QALY (N. Devlin et al., 2017). Within the 'extra-welfarist' approach, however, there does not seem to be a normative basis for selecting the mean over any of the (infinitely) many other possible aggregation functions (Roberts, 1980). In particular, it cannot be assumed that there is an objectively true value for each health state. Differences between individuals' health state valuations can, therefore, not be regarded as measurement errors, which cancel out when taking the average. Rather, differences have to be understood as genuine disagreements. If all individuals had similar preferences, however, the choice of the aggregation method would be trivial. Yet, empirical studies show that health state preferences differ considerably (Xie et al., 2014; also see Figure 2), and the societal preference is thus intimately dependent on the method of aggregation—if the method is changed, the outcome might differ. This raises the question: how should preferences be aggregated?

The (im)possibility of aggregating individual preferences into a social preference has been extensively discussed in social choice and welfare economic literature. Various welfare functions and voting rules have been axiomatically examined and their attractions and drawbacks have been described (Arrow, 1951; Brandt, Conitzer, Endriss, Lang, & Procaccia, 2016; Fleurbaey & Hammond, 2004; Sen, 2018). Seminal findings suggest that no method can be assumed to be unequivocally superior, or unanimously accepted. The decision which method to use always requires making value judgments. This means, to be able to say one method is better than another, it first needs to be decided what values should be incorporated. However, since this question has not yet been addressed in the context of population-based health state valuations, it is unclear what properties these functions should have. Currently, it is not even obvious what types of aggregation functions are admissible at all. In a recent discussion study, N. Devlin et al. (2017) suggested that a reasonable starting point for conceptualizing a social tariff would be the fundamental principle of the democratic system within which the health system operates: the majority rule. As an example, they consider the most common measures of central tendency (mode, mean, and median), but do not derive at a conclusive solution. In the following, I expand on their analysis and show that none of the three measures can appropriately reflect the majority view.

3.2 | Measures of central tendency and the majority rule

3.2.1 | The arithmetic mean

The arithmetic mean is commonly used to aggregate preferences in health state valuation studies (Xie et al., 2014; MVH, 1995), and it has convenient properties: it is easy to compute and to predict using regression models, and, unlike the median or mode, it is consistent with $[f(p_1(h_1), p_n(h_1))] - [f(p_1(h_2), p_n(h_2))] = f([p_1(h_1) - p_1(h_2)], [p_n(h_1) - p_n(h_2)])$. However, it takes into account the preference intensity of individuals, and thus does not reflect the majority view: the mean gives more weight to individual values that are distant from the average, which makes it sensitive to individuals with extreme preferences and outliers. This clearly conflicts with the democratic principle of 'one man (or woman), one vote'. As an example, consider Figure 2. The histogram shows the distribution of 735 individual preferences values for the EQ-5D 3L health state, "11,131" (no problems with mobility, self-care, usual activities and no anxiety or depression, but extreme pain or discomfort). Even though 58% of the individuals would prefer a higher value, the average is 0.24, because it is "pulled down" by individuals with more extreme negative utility values.

3.2.2 | The mode

Selecting the most frequent value from a complex distribution of cardinal preference values seems to be meaningless. The frequency of values mainly depends on the accuracy of the measurement and the extent of up- and down-rounding. In our example (Figure 2), 1 is by far the most frequent value ($n = 72$). However, 90% ($n = 663$) prefer a lower value. Overall, in the MVH (1995) data, all health states have a mode value of either 1, 0, or −1.

3.2.3 | The median

At first glance, the median provides a promising alternative: according to the Median Voter Theorem (Black, 1948), a majority will select the outcome most preferred by the median voter (given single peaked preferences).

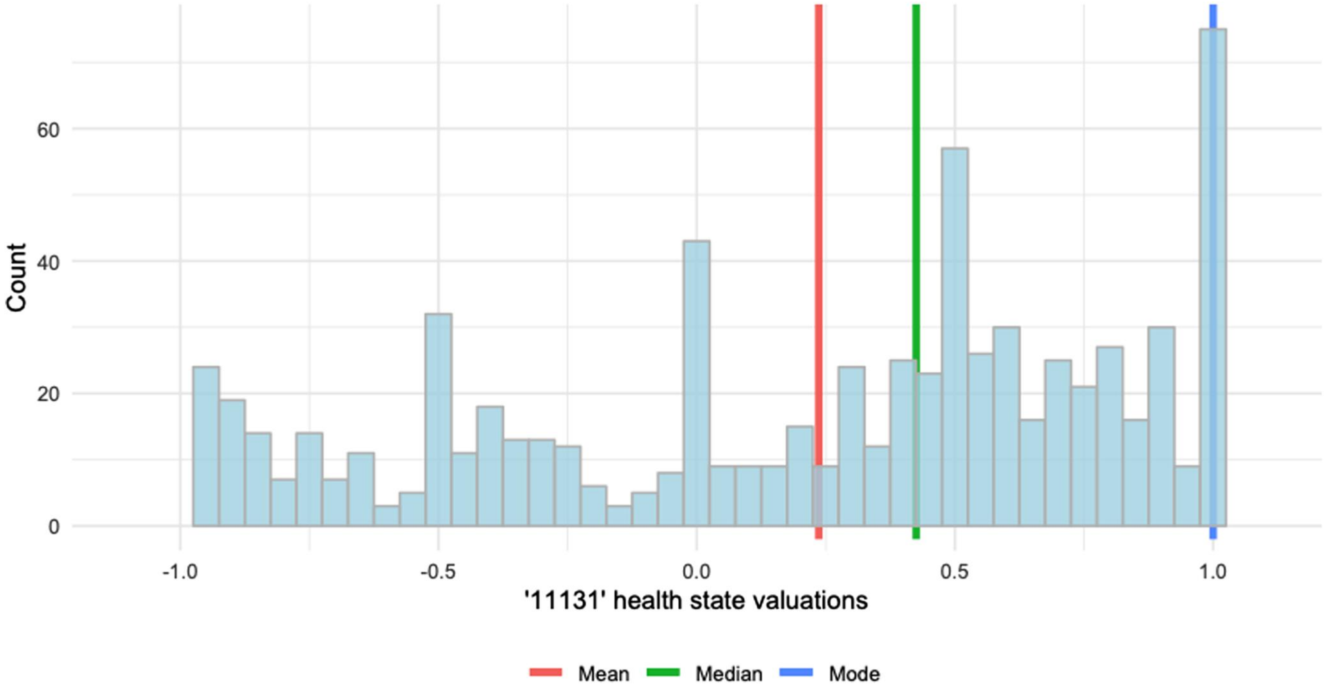

FIGURE 2 Distribution of individual preferences values ($n = 735$) for EQ-5D 3L health state, "11,131", and the corresponding mean, mode, and median. Source: data from MVH (1995)

Correspondingly, in our example, there is no majority for a value that is higher (preferred by 49.5%) or lower (preferred by 48.3%) than the median (=0.43). From this one might conclude that this is the value that the majority supports. However, the Median Voter Theorem only applies to voting on one dimension. For multiple dimensions, there is not necessarily a stable majority, and societal preferences might be intransitive (McKelvey, 1976). For the valuation of multiple health states, this means that although median values would reflect the majority view for each state individually, combining the median values of multiple health states into a social tariff might not represent the majority view globally. Moreover, the interpretation of median preferences is further complicated by the fact that the difference between the medians for two health states is not equal to the median difference. This can lead to paradoxical results, as the following example may illustrate.

Suppose individuals $x1$, $x2$, and $x3$ have preferences over health states h_1 and h_2: $x1$ prefers h_1 ($p_{x1}(h_1) = 0.65$) over h_2 (0.44); $x2$ also prefers h_1 (0.94) over h_2 (0.83); and only $x3$ prefers h_2 (0.98) over h_1 (0.34). One could thus conclude that a majority of individuals prefers h_1. However, the median values for the two health states are 0.65 and 0.83 (the geometric medians are 0.68 and 0.72), which would indicate that the group prefers h_2. See Figure 3 for a visual illustration.

3.3 | Constructing a democratic decision rule

None of the three measures of central tendency discussed above are able to incorporate the majority rule into the social tariff, let alone into decision-making. In the following, I will thus outline an alternative approach: a reformulation of the social tariff as a majority voting system over health programs (see Figure 4). Even if the proposed method is unlikely to be considered a viable alternative to the current system in the near future, it might serve to illustrate the conception of the social tariff as an instrument of democratic participation.

As noted above, the incremental societal QALYs of program a are given by ΔQ_a^t whereby the superscript indicates that incremental Health State Times ΔS are valued using the societal tariff $t(.)$. Alternatively, QALY estimates could be derived from individuals' health state preference functions directly, with $\Delta Q_a^{p_i} = p'_i(H) \times \Delta S$. The societal health effects of program a would then be evaluated by all n individuals separately (i.e., how many QALYs does program a generate in society from the perspective of individual i?). Imposing the societal efficiency decision rule $W(.)$ on everyone, individual i's decision function is given by

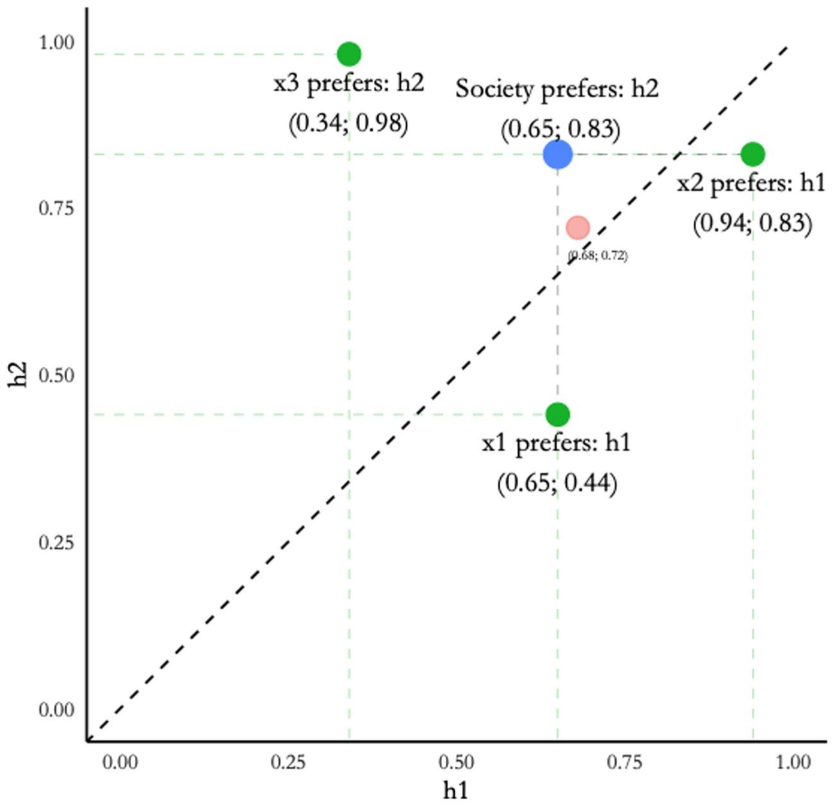

FIGURE 3 The "median health state paradox." Even though a majority of individuals (green dots) prefers health state *h*1 over state *h*2, based on median (blue) or geometric median (red) health state values, the group prefers *h*2

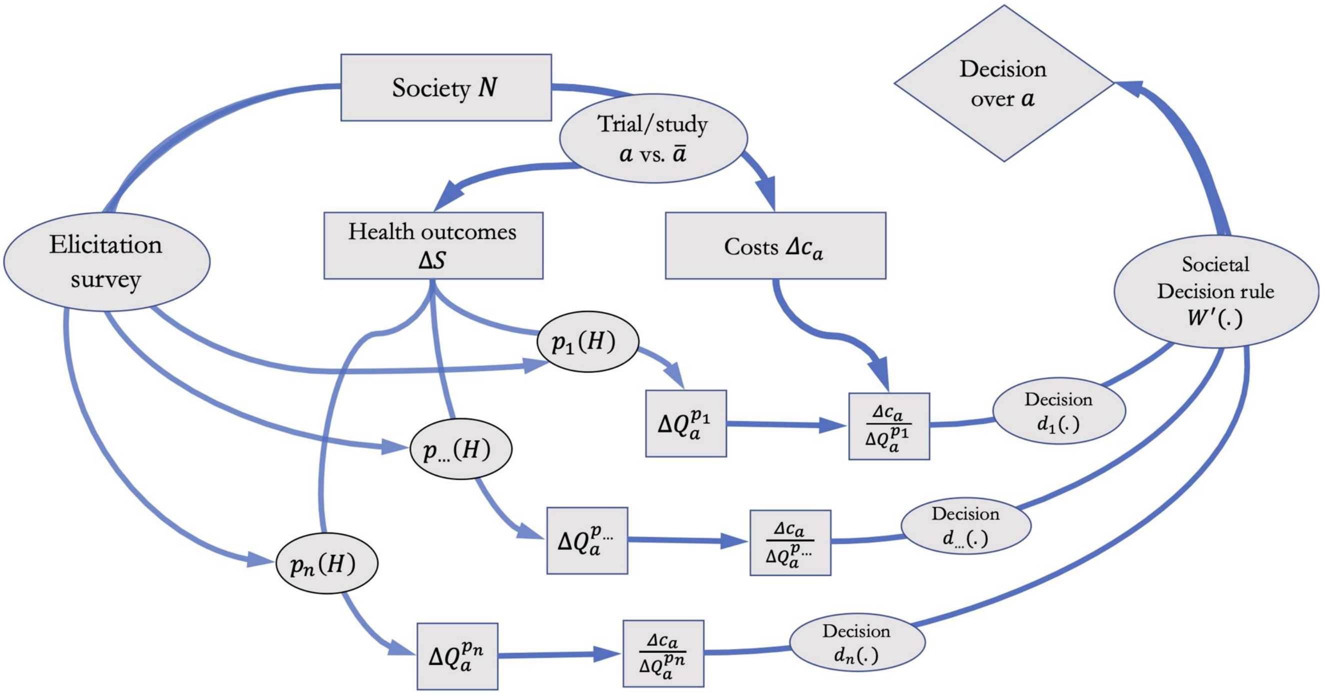

FIGURE 4 A democratic reformulating of the social tariff as a majority voting system [Colour figure can be viewed at wileyonlinelibrary.com]

$$d_i(a) = \begin{cases} 0 \ if \ \dfrac{\Delta c_c}{\Delta Q_a^{p_i}} > \theta \\ \\ 1 \ if \ \dfrac{\Delta c_a}{\Delta Q_a^{p_i}} \leq \theta \end{cases}$$

Subsequently, the societal decision rule could be reformulated as a majority voting system: individual decisions $d_1(a), d_2(a),..., d_n(a)$ could be summed up, and a should be adopted by society, if a majority of individuals 'voted' for it. The modified societal decision function W' is given below.

$$W^{'}(a) = \begin{cases} 0 \text{ if } \frac{n}{2} \leq \sum_{i=1}^{n} d_i(a) \\ 1 \text{ if } \frac{n}{2} > \sum_{i=1}^{n} d_i(a) \end{cases}$$

If more than two health programs are evaluated at the same time, majority voting has important limitations, and alternative voting rules should be considered (e.g., Brandt et al., 2016 provide a contemporary overview).

It should be stressed that the proposed change would only affect the level and the method of aggregation, while the source (the general population) and the objects (hypothetical health states) of preferences remain the same. Conceptually, however, this method offers a clear advantage over the current system: it would give all individuals equal weight in the decision. Furthermore, it would also be more transparent, in terms of how many individuals do and do not support a given policy decision. Thereby, the voting system might not only be more democratic, but also easier to understand than an average societal health state tariff. Nevertheless, one might rightly object that the informational demands of this system would be significant. Detailed primary data on the health outcomes, as well as individuals' health state preference functions would be required. Moreover, it should be emphasized that majority voting is insensitive toward individuals' preference intensities. The principle of "one person, one vote" means that any preference for program a counts as one (and any preference against as 0), regardless of how strong or weak the preference is—minuscule benefits to a small majority may thus outweigh any losses to a large minority. Finally, if the threshold of θ is assumed to be based on the marginal opportunity costs, each individual could, in principle, have a different threshold, depending on how many QALYs are currently being generated in the health care system from their perspective (i.e., according to their individual preference function).

4 | DEMOCRATIC REPRESENTATIVENESS

From a democratic perspective, it is not sufficient to address the question *how* preferences should be aggregated. It also needs to be determined *whose* preferences should count. Even if *the public* is to be accepted as the source of preferences, it has not been established what practical implications this may have for health state valuation studies. In the following, I make some recommendations for aspects that should be considered.

First, the surveyed group of individuals should be representative of the society for which a decision is to be made. This means, participants should be selected randomly. After the data are collected, all reasonable efforts should be made to retain representativeness throughout subsequent analyses. This implies that incomplete cases should not be excluded, nor should missing values be ignored. The exclusion of 399 (12%) participants from the in the MVH study (1994) because of missing values appears disconcerting in this regard. Missingness is unlikely to be (completely) random, and appropriate imputation methods should be considered (Rubin, 1976). Moreover, seemingly irrational preferences—for example, assigning the same value to all health states (Lamers, Stalmeier, Krabbe, & Busschbach, 2006) —should also not automatically be removed. Preferences might be consistent with some underlying beliefs, and researchers should not presume to make judgments about them (N. Devlin et al., 2017).

Second, democratic representativeness also commands that only those individuals are considered in the tariff, who are members of the very society, for which decisions are to be made. Health preferences vary across different regions and cultures (Gerlinger et al., 2019). NICE's decision to use a UK-wide, instead of an English tariff, to value health outcomes seems problematic in this regard, as it might well be the case that the four UK countries also have distinct preference profiles. One could take this a step further and argue that local authorities should also consider the use of local tariffs to evaluate local health programs. However, eliciting preferences and constructing social tariffs takes time and resources. The derivation of local, more accurate QALY estimates might thus only be worthwhile, if health preferences and subsequent policy decisions differ significantly between local communities. However, due to the scale of health care budgets, wrong decisions, based on biased estimates, could have significant opportunity costs.

Finally, it seems self-evident that an individual's participation in collective, democratic decisions needs to be intentional and deliberate. First and foremost, this means that participants in health state valuation studies need to be informed about the (potential) purpose of the survey (Israel, 2015). Using participants' stated preferences to inform policy decisions without obtaining informed consent for doing so does not only violate the autonomy of the participants, but it also seems utterly undemocratic. Given the potential impact their responses may have on health policy decisions, some individuals may want to give their answers more thought, and some may also prefer to abstain from participating. Notwithstanding, informing participants about the purpose of health valuation study may also invoke strategic behavior. Even though it seems unlikely that participants are able to foresee the effects their responses will have on any particular decisions, they may try to exaggerate their preferences in order to tilt the social tariff in the desired direction. Hausman (2010) further proposed that societal decisions should not be based on individuals' "private" health state values at all. Instead, public deliberations would be required to derive an adequate information basis for economic evaluations. I would argue that, at the very least, participants in health state valuation studies should be given the opportunity to reflect on their responses and to seek additional information about the health states they are not familiar with (N. J. Devlin, Shah, Mulhern, Pantiri, & van Hout, 2019; Gansen, Klinger, & Rogowski, 2019).

5 | HOW TO MOVE FORWARD

I have outlined research gaps related to the use of social tariffs in health economic evaluations. Considering their significance for health policy decision-making, further conceptual work is warranted to establish a sound and coherent theoretical foundation for social health state values. Before more appropriate theories and methods can be developed, it will be the responsibility of the decision makers to determine what social value sets are supposed to represent (e.g., utilities? indices of health?) and how they are to be derived (N. Devlin et al., 2017). I use the term "decision maker" here to include not only politicians and civil servants, but also members of the general public. Health economists can support the search for more appropriate preference aggregation methods and social welfare function by translating normative value judgments into corresponding decision rules. To this end, Dewitt et al. (2017) proposed a deliberative approach for eliciting meta-preferences from decision makers—that is survey how do they think preferences should be aggregated. In a first step, relevant ethical norms and societal values are identified from decision makers. Potential social tariffs are then constructed and subsequently presented to the participants. The preferences over the aggregation procedures (i.e., their meta-preferences) are then elicited in an iterative process.

6 | CONCLUSION

Under the assumption that the social tariff represents a major instrument of democratic participation, this study raises several critical questions and challenges the conceptual foundation of the current framework. Although the practical implications are still to be determined, a democratic (re)interpretation of the social tariff would undoubtedly have important consequences for population-based health state valuations. A new line of research is proposed to establish a conceptual basis for social tariffs from a democratic perspective.

ACKNOWLEDGMENTS

I am grateful to Aki Tsuchiya and Simon McNamara of the University of Sheffield, and to the participants of the 2019 HESG meeting in Norwich, for providing very helpful comments on a previous version of this manuscript. The usual disclaimers apply. This work was supported by funding from Wellcome [108903/B/15/Z] and the University of Sheffield. Paul Peter Schneider has received funding from the EuroQol foundation.

CONFLICT OF INTEREST

The author has declared that he has no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

ORCID

Paul Peter Schneider https://orcid.org/0000-0003-3552-1087

REFERENCES

Arrow, K. J. (1951). *Social choice and individual values*. New York, NY: John Wiley and Sons, Inc.

Black, D. (1948). On the rationale of group decision-making. *Journal of Political Economy*, *56*(1), 23–34.

Brouwer, W. B., Culyer, A. J., van Exel, N. J. A., & Rutten, F. F. (2008). Welfarism vs. extra-welfarism. *Journal of Health Economics*, *27*(2), 325–338.

Brandt, F., Conitzer, V., Endriss, U., Lang, J., & Procaccia, A. D. (2016). *Handbook of computational social choice*, New York, NY: Cambridge University Press.

Carr-Hill, R. A. (1989). Assumptions of the QALY procedure. *Social Science & Medicine*, *29*(3), 469–477.

Coast, J., Smith, R. D., & Lorgelly, P. (2008). Welfarism, extra-welfarism and capability: The spread of ideas in health economics. *Social Science & Medicine*, *67*(7), 1190–1198.

Cookson, R. (2005). QALYs and the capability approach. *Health Economics*, *14*(8), 817–829.

Culyer, A. J. (1989). The normative economics of health care finance and provision. *Oxford Review of Economic Policy*, *5*(1), 34–58.

Dakin, H., Devlin, N., Feng, Y., Rice, N., O'Neill, P., & Parkin, D. (2015). The influence of cost-effectiveness and other factors on nice decisions. *Health Economics*, *24*(10), 1256–1271.

Devlin, N., Shah, K. K., & Buckingham, K. (2017). *What is the normative basis for selecting the measure of 'average'preferences for use in social choices*. London, UK: Office of Health Economics OHE Research Paper.

Devlin, N. J., Shah, K. K., Mulhern, B. J., Pantiri, K., & van Hout, B. (2019). A new method for valuing health: Directly eliciting personal utility functions. *The European Journal of Health Economics*, *20*(2), 257–270.

Dewitt, B., Davis, A., Fischhoff, B., & Hanmer, J. (2017). An approach to reconciling competing ethical principles in aggregating heterogeneous health preferences. *Medical Decision Making*, *37*(6), 647–656.

Dolan, P. (2000). The measurement of health-related quality of life for use in resource allocation decisions in health care. *Handbook of Health Economics*, *1*, 1723–1760.

Dolan, P., Shaw, R., Tsuchiya, A., & Williams, A. (2005). QALY maximisation and people's preferences: A methodological review of the literature. *Health Economics*, *14*(2), 197–208.

Fleurbaey, M., & Hammond, P. J. (2004). Interpersonally comparable utility. In M. Fleurbaey & P. J. Hammond (Eds.), *Handbook of utility theory* (pp. 1179–1285). Boston, MA: Springer.

Gansen, F., Klinger, J., & Rogowski, W. (2019). MCDA-based deliberation to value health states: Lessons learned from a pilot study. *Health and Quality of Life Outcomes*, *17*(1), 112.

Gerlinger, C., Bamber, L., Leverkus, F., Schwenke, C., Haberland, C., Schmidt, G., & Endrikat, J. (2019). Comparing the EQ-5D-5L utility index based on value sets of different countries: Impact on the interpretation of clinical study results. *BMC Research Notes*, *12*(1), 18.

Hausman, D. M. (2010). Valuing health: A new proposal. *Health Economics*, *19*(3), 280–296.

Israel, M. (2015). Informed consent. In M. Israel (Ed.), *Research ethics and integrity for social scientists: Beyond regulatory compliance* (pp. 79–101). London, UK: SAGE Publications Ltd.

Lamers, L. M., Stalmeier, P. F., Krabbe, P. F., & Busschbach, J. J. (2006). Inconsistencies in TTO and VAS values for EQ-5D health states. *Medical Decision Making*, *26*(2), 173–181.

Lipscomb, J., Drummond, M., Fryback, D., Gold, M., & Revicki, D. (2009). Retaining, and enhancing, the QALY. *Value in Health*, *12*, S18–S26.

McKelvey, R. D. (1976). Intransitivities in multidimensional voting models and some implications for agenda control. *Journal of Economic Theory*, *12*(3), 472–482.

MVH Group. (1994). *The measurement and valuation of health: First report on the main survey*, York, UK: Centre for Health Economics, University of York.

MVH Group. (1995). *The measurement and valuation of health: Final report on the modelling of valuation tariffs*, York, UK: Centre for Health Economics, University of York.

National Institute for Health and Care Excellence (NICE) (2013). *Guide to the methods of technology appraisal 2013*. Retrieved from https://www.nice.org.uk/guidance/pmg9/resources/guide-to-the-methods-of-technology-appraisal-2013-pdf-2007975843781

Roberts, K. W. (1980). Interpersonal comparability and social choice theory. *The Review of Economic Studies*, 421–439.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Sen, A. (2018). *Collective choice and social welfare*, Cambridge, MA: Harvard University Press.

Shaw, J. W., Pickard, A. S., Yu, S., Chen, S., Iannacchione, V. G., Johnson, J. A., & Coons, S. J. (2010). A median model for predicting United States population-based EQ-5D health state preferences. *Value in Health*, *13*(2), 278–288.

Whitehead, S. J., & Ali, S. (2010). Health outcomes in economic evaluation: The QALY and utilities. *British Medical Bulletin*, *96*(1), 5–21.

Xie, F., Gaebel, K., Perampaladas, K., Doble, B., & Pullenayegum, E. (2014). Comparing EQ-5D valuation studies: A systematic review and methodological reporting checklist. *Medical Decision Making*, *34*(1), 8–20.

How to cite this article: Schneider PP. Social tariffs and democratic choice—Do population-based health state values reflect the will of the people? *Health Economics*. 2021;30:104–112. https://doi.org/10.1002/hec.4179

Conclusion to Chapter 2

In this chapter, I highlighted the importance of considering the role of social tariffs in health policy decision-making from a democratic perspective. I argued that, in order to reflect the majority opinion, a reformulation of the social tariff as a ‘majority voting system’ over health programs may be necessary. To this end, I outlined a theoretically attractive, but practically challenging alternative approach, which gives all individuals equal weight in the decision-making process. Additionally, I discussed the need to ensure democratic representativeness in health state valuation studies, and made recommendations for aspects that should be considered. Overall, this chapter suggests that further research is needed to develop a sound and coherent theoretical foundation for (social) health valuation. By translating normative value judgments into corresponding decision rules, health economists may help to establish more appropriate methods of preference aggregation and social value functions.

Chapter 3

Fair interpersonal utility comparison in the valuation of health

In the following chapter, I propose a new method for aggregating individual health state preferences into a social value set, called Utility Normalisation with Post-hoc Anchoring (UNPAc). In the standard aggregation function (the raw arithmetic mean), people's preferences for survival time (compared to health-related quality of life) determine the weight they have in the estimation. I argue that this influence should be considered unfair, and, inspired by relative utilitarianism, I propose UNPAc as an alternative solution. The potential impact of the new method and its feasibility is demonstrated using real world data (MVH) and a simulation study. The results indicate that UNPAc is able to reduce the association between individuals' relative preference for survival time and their influences on the overall outcome, and that this can yield a different set of social values. However, the differences may not always be considered relevant. Substantial unexplained variation in individuals' influences on the social value set remain.

Fair interpersonal utility comparison in the valuation of health: a relative utilitarian preference aggregation method

ABSTRACT

Background

The time trade-off method is widely used to elicit individual preferences over health states. The resulting utility values, measured on a scale anchored at full health (=1) and dead (=0), are commonly aggregated across individuals, in order to derive a social value set. In this paper, we argue that using dead, instead of the worst health state as a lower anchor is problematic, because individuals with a wider range of utility values have – simply for arithmetic reasons – more influence in the estimation of the social value set.

Methods

Inspired by relative utilitarianism, we propose an alternative aggregation procedure, the 'UNPAc method', which aims to equalise individuals' influences. We conducted a simulation study to demonstrate the potential impact of the new method on aggregate social health state values, and tested its practical feasibility in an EQ-5D 3L data set from the UK.

Results

For the simulation study we find that, in the standard approach, an agent's influence on the relative social value set increases linearly with their range, while in the UNPAc approach, there is practically no association between and range and influence. Both methods yield different sets of social values. When applied to the real-world EQ-5D-3L data, social value sets also differ between the two method, but, we do not find evidence for a clear, positive relationship between agents' utility ranges and their influences.

Conclusion

Our findings suggest that the new relative utilitarian approach can eliminate differences in influence that are due to differences in individuals' utility ranges. However, the differences in the resulting social value may not always be considered relevant, and substantial, currently unexplained variation in individuals' influences on the social value set remain.

Highlights

- Standard health valuation methods do not give every individual equal weight in the estimation of social value sets. This is because preferences for length of life are conflated with preferences for quality of life.
- We propose a new preference aggregation method (UNPAc) to resolve this problem
- Our simulation study demonstrates the effectiveness of the new method under certain circumstances
- When applied to actual EQ-5D-3L data, however, results are more ambiguous: differences between the standard and the new approach might not always be considered relevant

Introduction

The measurement and valuation of health-related quality of life (HRQoL) is an integral component of health economic evaluation [1–3]. Methods for eliciting health preferences and estimating social value sets for health descriptive systems, such as the EQ-5D-3L, have evolved significantly over the last two decades. However, while much attention has been paid to the psychometric and statistical properties, little consideration was given to their normative underpinnings. In fact, one might even get the impression that the elicitation of preferences and the estimation of a social tariff is treated mainly as a measurement problem. Yet, this disregards the fact that 'measuring' social preferences requires making strong value judgements [4].

The problem which we will be concerned with in this paper is that when different individuals are asked to state their preferences over a number of health states, there can be considerable disagreement between them. For example, is mild pain better or worse than severe mobility impairment? To determine the *social* value of a health state, which is supposed to represent the preference of the group of individuals as a whole, some form of compromise must be reached [4, 5]. By far the most commonly used approach to solve this problem is to aggregated utilities across individuals, by taking the arithmetic average [6].

In previous work we have already argued that aggregating health state utilities across individuals by taking the average may yield unfair social outcomes [Anonymous]. The present study aims to extend our theoretical work by presenting an alternative aggregation procedures that equalises the influences individuals have on the estimation of (relative) social value sets. We test the alternative method in a simulation study, and assess its feasibility, using a real world EQ-5D-3L data set from the UK.

Theoretical framework

Interpersonally (in)comparable health state utilities

Social preferences cannot be measured (purely) objectively. They are not physical quantities, like blood pressure or body weight. The commonly used scale, anchored at full health, set to one, and dead, set to zero, does thus not measure health state preferences in some natural units of utility (or HRQoL). Instead, it is presumably imposed on normative grounds, as a matter of fairness. In contrast to, for example, willingness to pay, the 1-0 scale used in time trade-off or standard gamble method, equalises individuals' preference functions, in the sense that one year in full health (and one year 'being dead') is assumed to be of equal value for all individuals [7, 8]. Thereby, utility values are supposedly made fully interpersonally comparable. At closer inspection, however, this assumption seems implausible and may lead to unfair outcomes [9].

While 'full health' and 'being dead' might intuitively seem to be obvious and valid anchor points, at closer inspection, the situation becomes more difficult: 'full health' refers to a state with a certain HRQoL in which a person spends some amount of time. Dead, however, refers to a 'non-state' in which survival time is absent.

This means, a utility scale that is anchored at full health and dead actually conflates the evaluations of two different criteria: HRQoL and survival time. The former reflects individuals' assessment for how much better or worse a given health state i is compared to another health state j – we refer to this as the relative value of a health state. The latter reflects the value of one year in state i (or j) compared to being dead - we refer to this as the absolute value. A comparison with 'being dead' is, however, qualitatively different comparisons with other health states, as it does not only involve the evaluation of HRQoL, but also the value of any non-health components of life.

We would argue that these two types of preferences should be considered – and aggregated – separately. Conflating both into one score before aggregating utilities across individuals affects individuals' ranges of utility values, that is the utility difference between the best (full health) and the worst health state (pit state). All else being equal, individuals with a a higher rate of substitution between units of HRQoL and units of survival time (RSQS) – that is an individuals' willingness to trade survival time for gains in HRQoL – will have a wider utility range. Individuals with a lower RSQS, on the other hand, will have a more narrow range.

This is problematic, because individuals with a wider range of utility values will – simply for arithmetic reasons – get more, potentially disproportionate influence in the estimation of the social tariff: all else being equal, the wider the utility range, the more influence an individual has on the relative social value of a given health state. This means, if (and only if) individuals' preferences over health states differ with respect to their utility ranges, average health state values will reflect the preferences of individuals with wide ranges to a greater extent than of individuals with a narrow range.

One might think that this effect is acceptable or even desirable, because individuals with a wide utility range apparently care more about HRQoL than others – maybe they should have more say in the decision how much more or less valuable health state i is compared state j? We would oppose this position for two reasons.

Firstly, the argument above implies that individuals who prioritise survival time over HRQoL (i.e. individuals with a lower RSQS) then get more say over the value of state i (and j) compared to being dead in return. Yet, this is not the case. Following the rules of simple arithmetic, individuals with a very low utility range (low RSQS) have the same influence as individuals with a very wide range (high RSQS), while individuals with an average utility range have the least influence (this will be illustrated in the simulation study below).

Secondly, only because a person is not willing to trade much survival time for gains in HRQoL, does not mean that they actually do not care about their health. Imagine a single-parent. They might be very reluctant go give up any life time, in order to be able to continue caring for their children. Yet, their unwillingness to trade survival time should not be mistaken for indifference. They may well be willing to spend all their income on improving their HRQoL – just not their survival time. Another person, with the same level of income, but without any dependants, might not be willing to spend as much money to improve their health, but may trade large proportions of their remaining lifetime, just because they do not care so much about it. It thus appears questionable whether the single-parent in this example should principally have less influence on the relative social value of health states than the other person. For the aforementioned reasons, we think that anchoring the utility scale at full health and dead does not yield fully interpersonally comparable utility values. In the absence of any compelling reason to consider the preferences of individuals with a wider range of utility values more important, disparities in people's influences on the relative social health state values (that are due to utility range differences) should be considered illegitimate.

A novel approach: Utility Normalisation with Post-hoc Anchoring (UNPAc)

To resolve the inequitable distributions of influence based on utility ranges, and inspired by concept of relative utilitarianism [10], we propose a new method for aggregating utility values across individuals: 'Utility Normalisation with Post-hoc Anchoring' (UNPAc). UNPAc consists of three simply steps (a more formal description of the method is provided in the Appendix):

1. The preferences of each individual are normalised between their best and their worst health state, so that all preference functions range from full health (=1) to their pit state (=0).

2. Utilities for each health state are aggregated across individuals (by taking the average) to derive a normalised social value set.

3. The normalised social value set is re-anchored using the average utility range (the average utility range is equal to 1 minus the average utility of the worst health state, and may also be interpreted as the average RSQQ)

In contrast to the standard approach, anchoring health state preferences between full health and the pit states effectively normalises individuals' preference functions: the utility difference between the most preferred and the least preferred health state then has the same value for everyone, irrespective of their preferences for survival time. Only after normalised utilities are aggregated into social values, they are re-scaled on to the QALY scale (anchored at full health and dead). This means, the UNPAc disentangles the aggregation of preferences for HRQoL and the aggregation of the RSQQ. Preferences for each are aggregated separately, and only then are the two combined to derive the final social value set.

It should be noted that the proposed UNPAc approach is informationally more demanding than the standard approach. To be able to normalise utilities on the individual-level, two conditions must be met: for each individual, one must know the utilities for their best and worst health state, and the difference between those two must be some non-zero value. Otherwise, a normalised preference function can not be constructed. However, in most cases this will be feasible, as health descriptive systems usually include one objectively best (perfect health) and worst (pit) health state, and most people are willing to trade survival time for gains in HRQoL.

Simulation study

We conducted a simulation study to demonstrate the impact of using the UNPAc, instead of the standard approach, for aggregating health state preferences. The primary objective was to assess the differences in the association between individuals' utility ranges and their influences on the relative and absolute social value set.

Methods

We simulated 10,000 agents. 4,500 (45%) agents were assigned to group G_M, and 5,500 (55%) were assigned to group G_P. Agents had preferences over 4 health states: FH = Full health; PM = Physical and mental health problems; P = Physical health problems; M = Mental health problems. All agents considered FH the best (= 1), and PM the worst state, but agents in G_M preferred M over P, while agents in G_P preferred P over M.

To construct cardinal preference functions, we first randomly generated values for PM from a truncated normal distributions. For G_M agents, the mean (SD) was set to -0.9 (0.3), and for G_P, it was set to -0.25 (0.3). For both groups, the minimum and maximum was set to -1 and 1.

Utility values for states P and M were generated in two steps: we first sampled, for each agent separately, two values from a uniform random distribution, spanning from 1 to the agent's utility for PM. Secondly, the two values were ordered according to the group assignment, with $G_M : u(FH) > u(M) > u(P) > u(PM)$; and $G_P : u(FH) > u(P) > u(M) > u(MP)$. This procedure ensured that agents in both groups had, on average, the same relative preference intensities for either P or M. The only systematic difference was that that agents in G_M had a higher RSQS than agents in G_P, i.e their ranges of utility values were wider, i.e. the utility of PM was lower.

Social values sets were computed using the standard (mean aggregation) and the UNPAc approach. We compared the resulting absolute and relative value sets, and assessed differences in the relationship between individuals' utility ranges and their influences on the social value set, which was the primary outcome of interest.

An individual agent's influence was measured as the sum of the absolute differences between a social value set with and a set without their preferences taken into account. This means, we first estimated a reference value set including the preferences of all n agents. We then computed n test value sets, each of which ignored the preferences of one of those agents.

The association between agents' utility ranges and their influences was investigated by means of plotting the data and fitting smooth loess curves to it.

We also conducted a sensitivity analysis to verify that in a scenario where agents' utility ranges are not associated with specific preference profiles, UNPAc is equivalent to the standard approach. For this, Groups G_P and G_M were simulated to have, on average, the same utility ranges (= -0.575).

Results

Table 1 shows the average absolute and relative health state values separately for agents with the groups G_M and G_P. G_M agents prefer M over P, while G_P agents prefer P over M. While the absolute utility difference between M and P is greater in the G_M group (0.56 vs 0.43), the relative utility difference is the same (0.33 vs 0.33). Additional descriptive statistics are provided in the appendix (see S1).

After both groups were merged into one (society), we applied the standard and the UNPAc aggregation method. The resulting social value sets are provided in Table 2. It shows that the standard approach yielded a social value set in which the absolute and relative value of M was larger than the value of P (0.28 vs 0.26 and 0.51 vs 0.49). This means, the social values reflect the preference

Table 1: Mean absolute (relative) state utilities within groups

State	G_M (n = 4,500; 45%)	G_P (n = 5,500; 55%)
FH	1.00 (1.00)	1.00 (1.00)
M	0.42 (0.66)	0.16 (0.33)
P	-0.14 (0.34)	0.59 (0.67)
PM	-0.72 (0.00)	-0.24 (0.00)

profile of G_M agents, even though they are a minority (45%) and despite the fact that their relative preference for M over P was exactly the same as the other group's preference for P over M.

The social value set derived through the UNPAc method showed different results. The values for FH and PM remained the same, but the preference order for M and P was reversed. The absolute and relative values were 0.24 vs 0.30 and 0.48 vs 0.52, respectively.

Table 2: Comparison between the absolute (relative) social value sets based on the standard approach and the UNPAc

State	Standard approach		UNPAc approach		Difference	
	rank	value	rank	value	rank	value
FH	1	1.00 (1.00)	1	1.00 (1.00)	0	0 (0)
M	2	0.28 (0.51)	3	0.24 (0.48)	1	-0.04 (-0.03)
P	3	0.26 (0.49)	2	0.30 (0.52)	-1	+0.04 (+0.03)
PM	4	-0.46 (0.00)	4	-0.46 (0.00)	0	0 (0)

Figure 1 illustrates the mechanism behind these divergent results. It shows the influences of individual agents on the relative and absolute value set as a function of their utility ranges, with smooth loess curves fitted to the data.

In the standard approach, an agent's influence on the relative social value set increases linearly with their range, while in the UNPAc approach, there is practically no association between and range and influence. In both approaches, the association between agents' utility ranges and their influences on the absolute value set, shows a V-shaped pattern, i.e. agents with a very wide and very narrow have more influence, while agents with an average range have the least influence.

The sensitivity analysis confirmed that when agents with different preference profiles do not differ with respect to their utility ranges, the UNPAc approach provides results that are equivalent to the standard approach (see Table S2 in the Appendix).

Empirical application

In order to test the feasibility of using the proposed alternative UNPAc utility aggregation method in practice, we applied it to an EQ-5D-3L health state preference data set from the UK. As for the simulation study, the primary outcome of interest was the association between individuals' utility ranges and their influences on the (relative) social value set. However, we also compared the resulting value sets, to assess whether or not differences could be considered relevant in practice.

Methods

The EQ-5D-3L Instrument

The EQ-5D 3L instrument is a generic preference-based instrument for the measurement and valuation of health. It consists of 243 health state, defined along five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), each of which has three severity levels (no problems, some problems, and severe problems). Health states can accordingly be referred to by a 5-digit code. '23111', for example, denotes a state with some mobility problems, severe problems with self-care, but no problems with usual activities, pain/discomfort, or anxiety/depression. Accordingly, '11111' denotes full health; and '33333' denotes the (objectively) worst health state.

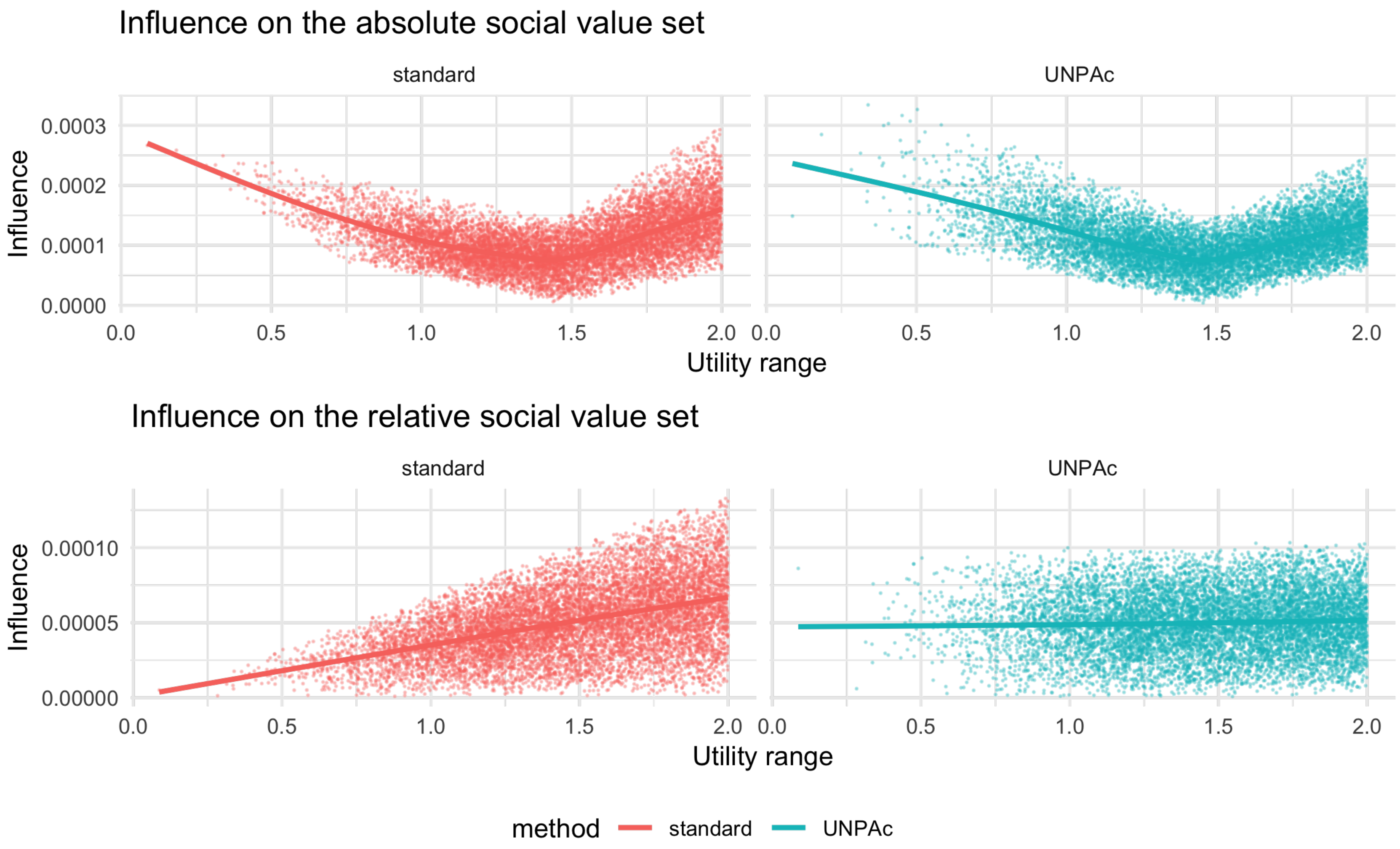

Figure 1: Association between agents' utility value ranges and their influences on the relative (bottom) and absolute (top) value sets in the standard (left) and alternative 'Utility Normalisation and post-hoc anchoring' (right) approach. Each dot represent one agent. Smooth loess curves are fitted to the data to illustrate the association.

Setting and data

We used EQ-5D-3L data from the 1993 Measurement and Valuation of Health (MVH) Study [11–13]. We reproduced the methods of the original study as far as possible, but had to make two adjustments, in order to apply the normalisation methods.

Firstly, to avoid zero-division, we had to exclude 12 of the 2,997 individuals included in the original study, because they had a utility range of 0, i.e. they assigned a utility of 1 to all states. Secondly, we added full health to the data set and set it equal to one for all individuals.

Statistical modelling

In contrast to the simulation study, the data collected in real world valuation tasks does not allow computing social value sets directly. Instead, a statistical model has to be fit to the data, which can then be used to predict social health state values.

To estimate the standard (average) social value set, we used the main effects OLS model proposed by Dolan [11]. It includes twelve variables, two dummy variables for each dimension (e.g. MO2 and MO3 for mobility), representing moves to some and severe problems, a constant (α), representing any move away from full health, and N3, which is an additional utility decrement for having severe problems on at least one dimension.

To estimate the UNPAc social value set, we first normalised each participants health state preferences between their best and their worst health state. Secondly, we fitted the standard OLS model to the normalised preference data, and then re-anchored the model using the mean RSQS/utility range.

The remaining steps of the analysis are identical to the simulation study. We estimated the relative and absolute social values for all 243 health states using the standard and the UNPAc approach, and then compared the results. The associations between participants' utility ranges and their influences on the social value sets were assessed by plotting the data and fitting smooth loess curves to it.

Results

A total of 2,985 participants were included in the analysis.

Table 3 shows a comparison between the standard OLS model and the alternative UNPAc model. For most parameters, the differences were small. However, in the UNPAc model, the beta for AD3 was 6% higher compared to the standard model, and the beta for N3 was 8% lower, respectively.

Table 2. Main effects model - standard and UNPAc parameter estimates

	Standard	UNPAc		UNPAc/Standard	
Variable	b (95%CI)	Normalised b (95%CI)	Rescaled b (95%CI)	Ratio	Difference
Intercept	0.05 (0.04; 0.06)	0.03 (0.03; 0.04)	0.05 (0.04; 0.06)	1.01	0.000
MO2	0.07 (0.06; 0.08)	0.04 (0.04; 0.05)	0.07 (0.06; 0.08)	0.99	-0.001
MO3	0.24 (0.23; 0.25)	0.15 (0.14; 0.16)	0.24 (0.23; 0.25)	1.01	0.002
SC2	0.12 (0.11; 0.13)	0.07 (0.06; 0.08)	0.11 (0.10; 0.12)	0.96	-0.005
SC3	0.11 (0.09; 0.12)	0.07 (0.06; 0.08)	0.11 (0.10; 0.12)	1.05	0.005
UA2	0.04 (0.02; 0.05)	0.02 (0.02; 0.03)	0.04 (0.03; 0.05)	1.04	0.002
UA3	0.06 (0.04; 0.07)	0.03 (0.03; 0.04)	0.06 (0.04; 0.07)	1.01	0.001
PD2	0.13 (0.12; 0.14)	0.08 (0.07; 0.08)	0.12 (0.11; 0.13)	0.96	-0.006
PD3	0.26 (0.24; 0.27)	0.16 (0.15; 0.17)	0.26 (0.25; 0.27)	1.00	0.000
AD2	0.08 (0.07; 0.10)	0.05 (0.04; 0.06)	0.08 (0.07; 0.09)	0.97	-0.003
AD3	0.16 (0.15; 0.18)	0.11 (0.10; 0.11)	0.17 (0.16; 0.18)	1.06	0.010
N3	0.28 (0.27; 0.30)	0.16 (0.15; 0.17)	0.26 (0.25; 0.28)	0.92	-0.022
R^2	0.51	0.59	0.51		
Observations	38,805	38,805	38,805		

These differences translate into differences in social health state values. Figure 2 shows a comparison of the social values for all 243 EQ-5D-3L health states. Detailed results for both value sets, and a state-by-state comparison, can be found in the appendix. Here, we limit ourselves to some general observations: The UNPAc method generally yielded higher social values than the standard approach. The mean difference between standard and UNPAc social values was 0.022, with a standard deviation of 0.008. Differences tend to be smaller for high and low ranked states, and larger for intermediate states. The maximum absolute difference of 0.036 (standard: 0.003 vs. UNPAc: 0.040) was observed for state '22132'. Differences in social values also resulted in rank differences.

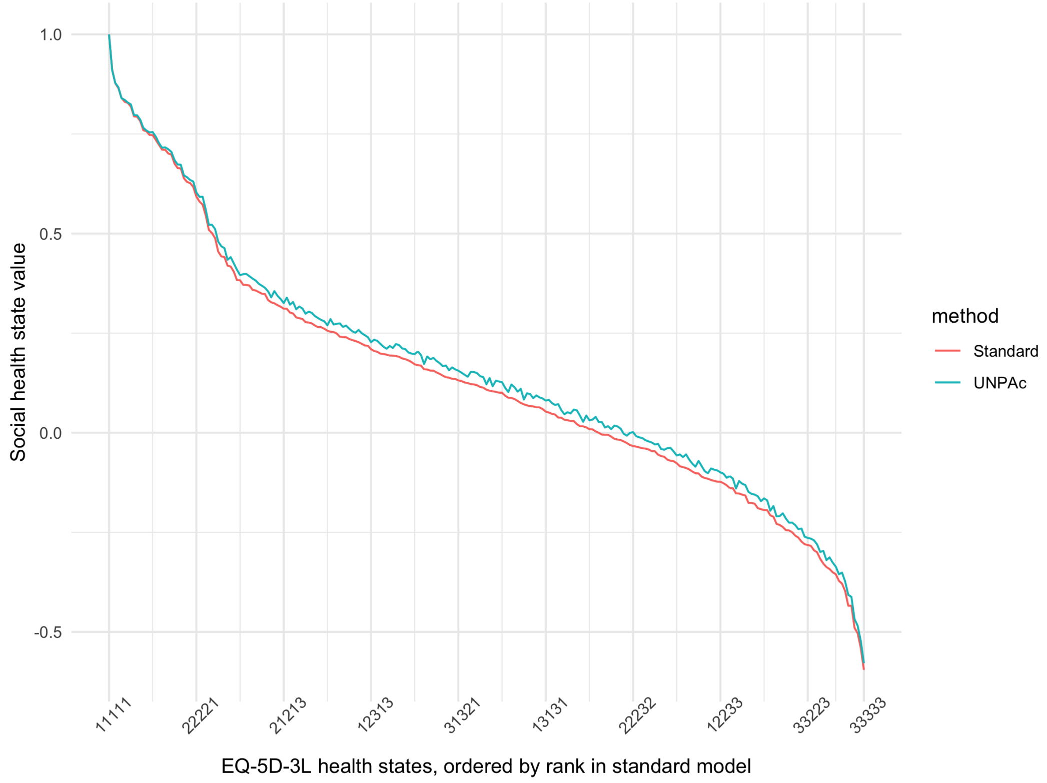

Figure 2: Comparison of the social values for all 243 EQ-5D-3L health states, based on the standard and the UNPAc utility aggregation methods. Health states are ordered by desirability, from the most (FH) to the least preferred state, according to the standard model.

A total of 156 (64.2%) health states had a different rank. The highest rank change of +6 (134→140) occurred for health state '31213'.

Utility ranges and influence

Individuals' utility ranges varied between 0.01 and 1.98, with an average (SD) of 1.62 (0.36) and a median (IQR) of 1.73 (0.55). Only 134 (4%) of the participants did not consider any health state to be worse than dead and had a utility range of less than one. Furthermore, only 2,205 (74%) participants assigned the lowest utility value to the objectively worst health state '33333'. This means, for 780 (26%) participants, at least one health state had a lower value.

A visual illustration of the association between agents' utility ranges and their influences on the absolute and relative social value sets in the standard and the UNPAc approach is given in figure 3. It shows individuals' influences on the relative and absolute value set as a function of their utility ranges, with smooth loess curves fitted to the data.

In contrast to the results derived from the simulation study, we neither found evidence for a clear, positive linear relationship between agents' utility ranges and their influences on the relative social value set in the standard approach, nor for any systematic difference in the shape of the association between the standard and the UNPAc approach. Instead, all four subfigures showed a rather similar, flat U-shaped association, indicating that agents with low (< 1) and agents with high (> 1.75) utility ranges had more influence than agents with more average ranges. For the absolute value set, individuals with narrow ranges furthermore seemed to have more influence than those with very wide ranges. However, the number of individuals with narrow ranges was low ($n = 134$; 4%) and thus patterns may be affected by few outliers.

Data and code availability

The entire R source code for both, the simulation study and the empirical analysis, are provided on a data repository, where it is available for reuse and adaptation. The principal investigators of the 1993 MVH study also kindly gave us permission to publish the relevant data alongside our code [Anonymous].

Discussion

In this paper, we have outlined potential flaws in the procedure that is commonly used to aggregate health state utilities across individuals into social value sets for use in economic evaluations. Our main theoretical argument

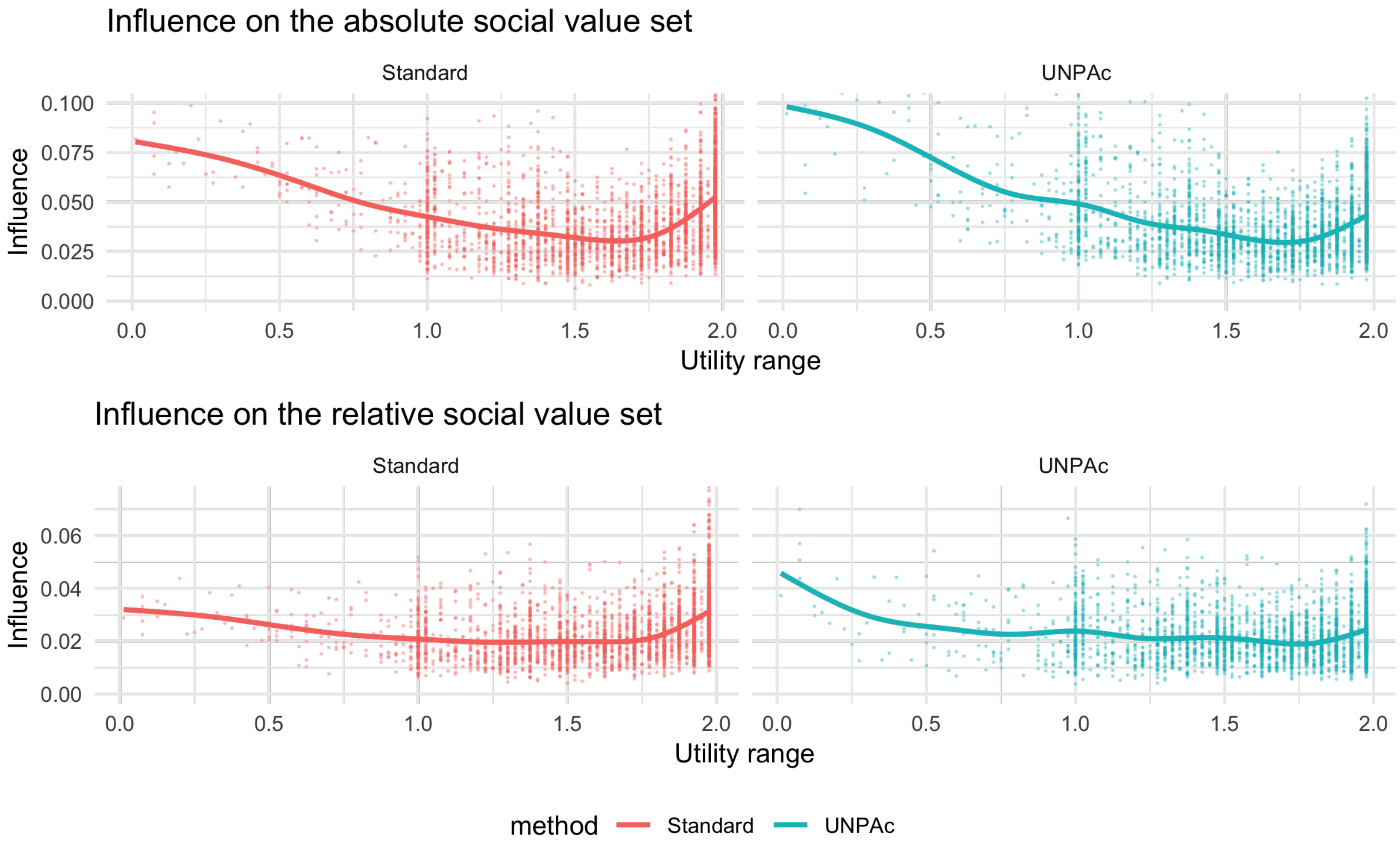

Figure 3: Association between agents' utility value ranges and their influences on the relative (bottom) and absolute (top) EQ-5D-3L social value sets in the standard (left) and alternative UNPAc (right) approach. Each dot represent one agent. Smooth loess curves are fitted to the data to illustrate the association.

is that the standard method fails to account for differences in the range of utility values, and may thus lead to a very unequal – and potentially unjust – distribution of influence on social value sets. Social value sets might then predominantly represent the preferences of those individuals who are more willing to trade units of survival time for gains in units of HRQoL. This does not seem to be an intentional choice, or to be founded in theory, but rather an accidental result of the fixed 1-0 full health - dead utility scale.

Inspired by the concept of relative utilitarianism, we proposed an alternative utility aggregation method, the UNPAc approach, to resolve the problem of the inequitable distribution of influence [10]. The method equalises individuals' influences in the estimation of the relative social tariff, by normalising everyone's health state preferences between their best and worst state.

We demonstrated the feasibility of the UNPAc approach and demonstrated its potential impact in a simulated data set. When applying the approach to the data set of the well-known MVH study [11–13], however, results were not fully conclusive: although we found evidence of an unequal distribution of influence on the UK EQ-5D-3L social value set, the relationship between utility ranges and influence was more complex than in the simulation study (and then we expected): in contrast to the simulated data set, in the empirical data, not only individuals with a very wide range of utility values had more influence, but also individuals with very low range. In this situation, the UNPAc method did not – and should not be expected to – fully equalise the influence across individuals, because it only eliminates those inequalities that are due to differences in the utility ranges. There are, of course, inequalities due to other factors. These may include, for example, the specific subset of health states an individuals is asked to value, or the respective preference profile. The group of individuals with very low utility ranges but high influence in the MVH data set was small (n = 134; 4%), and the observed high influence could well be due to unusual preference profiles of these particular individuals.

Overall, the UNPAc provided a different social EQ-5D 3L value set for the UK than the standard approach. Even though the differences between the two models were small, there are two important consideration that should be taken into account. Firstly, the TTO full health - dead scale only ranges from one to zero, and so the magnitude of the differences must be interpreting on that scale. Secondly, the EQ-5D 3L instrument pre-specifies a large proportion of the relationships between its 243 health states: State '21111', for example, dominates states like '22111', '21211', etc, each of which, in turn, dominates states like '22222', or '32222', etc. This inherent structure constraints the range of health state utility differences that alternative valuation and aggregation methods can be be expected to yield. For the 5L version of the EQ-5D instrument, differences between methods will be even more constrained [14]. Nevertheless, when used to value health outcomes in terms of QALYs for health economic evaluations, even these small differences might become relevant.

An important limitation of the proposed UNPAc approach are its informational requirements. To be able to apply the method, it is essential that the utility values for the best and the worst health state are known for each individual. Yet, in the context of the EQ-5D 3L descriptive system, it does not seem necessary to elicit the preferences for all health states: full health has, by definition, the highest value (=1), and for the worst state, it seems plausible to assume that it is '33333', which is objectively dominated by all other states. However, it should be noted that 780 (26%) participants in the MVH study actually assigned the lowest TTO value to one of the other 11 health states that they valued (own analysis). While the min-max normalisation can still be applied, this rather stunning finding calls into question the quality of the preference data obtained on the individual level.

Even though several open questions remain, we think our theoretical framework provides a promising approach to generate more equitable social value estimates for use in health economic evaluations. Moreover, we think that the framework might also be useful for addressing conceptual problems in other areas. In particular, we think it may help with the development of social values for EQ-5D-Y health states in children [15]. Studies have shown that TTO values are higher for health states experienced by a 10-year-old child compared to health states experienced by an adult. At the same time, values derived through the VAS valuation techniques tend to be lower for states experienced by children [16]. These results are difficult to interpret and might seem paradoxical. Within our framework, however, they can be easily integrated. TTO values can be higher in children, even though their HRQoL is judged to be lower. This can be the case if the rate of substitution between quality and quantity of life is also higher. This means, people might be willing to trade fewer units of survival time, to gain a unit of HRQoL in in children than they are in adults. To harmonise the valuation systems for children and adults, it might be necessary to either apply a common social scaling factor, or to increase the value of survival time in children, so that one year in full health is worth more than

More research is needed to better understand the practical and normative implications of the standard and other utility aggregation methods. The UNPAc and other methods should be tested empirically in other data sets and in other descriptive systems. Ideally, studies should seek to investigate the relationship between individual-level preference functions and the social value set [**devlin**]. Nevertheless, to be able to ultimately decide which properties an aggregation method should or should not have, it seems crucial to establish a more sound theoretical foundation for the valuation of health in general. The field may benefit from a closer consideration of findings from Welfare Economics, Social Choice Theory and parts of philosophy, in which the problem of interpersonal utility comparisons has been addressed more rigorously [9, 17, 18].

References

[1] Karimi, M. and Brazier, J., "Health, health-related quality of life, and quality of life: What is the difference?" *Pharmacoeconomics*, vol. 34, no. 7, pp. 645–649, 2016.

[2] Brazier, J., "Valuing health states for use in cost-effectiveness analysis," *Pharmacoeconomics*, vol. 26, no. 9, pp. 769–779, 2008.

[3] Drummond, M. F., Sculpher, M. J., Claxton, K., Stoddart, G. L., and Torrance, G. W., *Methods for the economic evaluation of health care programmes.* Oxford university press, 2015.

[4] Devlin, N., Shah, K., Buckingham, K., *et al.*, "What is the normative basis for selecting the measure of 'average'preferences for use in social choices," *OHE research paper. Office of Health Economics, London*, 2017.

[5] Schneider, P., "Social tariffs and democratic choice–do population-based health state values reflect the will of the people?," 2019.

[6] Xie, F., Gaebel, K., Perampaladas, K., Doble, B., and Pullenayegum, E., "Comparing eq-5d valuation studies: A systematic review and methodological reporting checklist," *Medical Decision Making*, vol. 34, no. 1, pp. 8–20, 2014.

[7] Attema, A. E., Edelaar-Peeters, Y., Versteegh, M. M., and Stolk, E. A., "Time trade-off: One methodology, different methods," *The European Journal of Health Economics*, vol. 14, no. 1, pp. 53–64, 2013.

[8] Miyamoto, J. M., Wakker, P. P., Bleichrodt, H., and Peters, H. J., "The zero-condition: A simplifying assumption in qaly measurement and multiattribute utility," *Management Science*, vol. 44, no. 6, pp. 839–849, 1998.

[9] Hausman, D. M., "The impossibility of interpersonal utility comparisons," *Mind*, vol. 104, no. 415, pp. 473–490, 1995.

[10] Isbell, J., *Absolute games, in" contributions to the theory of games, iv"(aw tucker and rd luce, eds.)* 1959.

[11] Dolan, P., "Modeling valuations for euroqol health states," *Medical care*, pp. 1095–1108, 1997.

[12] Williams, A. H., Gudex, C., Kind, P., and Dolan, P., "The measurement and valuation of health study," Tech. Rep., 1993, UK Data Service. SN: 3444. Available at: `http://doi.org/10.5255/UKDA-SN-3444-1`.

[13] MVH Group, "The measurement and valuation of health: Final report on the modelling of valuation tariffs," *Centre for Health Economics, University of York*, 1995, Available at: `https://www.york.ac.uk/media/che/documents/reports/MVHFinalReport.pdf`.

[14] Devlin, N., Shah, K., Feng, Y., Mulhern, B., and Van Hout, B., "Valuing health-related quality of life: An eq-5d-5l value set for england. 2016," *URL: www. ohe. org/publications/valuing-healthrelated-quality-life-eq-5d-5l-value-set-england (accessed 10 June 2016)*, 2017.

[15] Kreimeier, S. and Greiner, W., "Eq-5d-y as a health-related quality of life instrument for children and adolescents: The instrument's characteristics, development, current use, and challenges of developing its value set," *Value in Health*, vol. 22, no. 1, pp. 31–37, 2019.

[16] Rowen, D., Rivero-Arias, O., Devlin, N., and Ratcliffe, J., "Review of valuation methods of preference-based measures of health for economic evaluation in child and adolescent populations: Where are we now and where are we going?" *PharmacoEconomics*, pp. 1–16, 2020.

[17] d'Aspremont, C. and Gevers, L., "Social welfare functionals and interpersonal comparability," *Handbook of social choice and welfare*, vol. 1, pp. 459–541, 2002.

[18] Fleurbaey, M. and Hammond, P. J., "Interpersonally comparable utility," in *Handbook of utility theory*, Springer, 2004, pp. 1179–1285.

Appendix

S1. The conventional and the UNPAc utility aggregation procedure

Let $H \in \{h_1, h_2, \ldots, h_k\}$ denote a set of k health states, for which social values are to be determined for a group of n individuals. Individual j's health state preferences, denoted $u_j(H) \in \{u_j(h_1), u_j(h_2), \ldots, u_j(h_k)\}$, are measured on the TTO scale, anchored at full health (=1) and dead (=0). By definition, full health has a utility of 1, which is everyone's highest utility value ($u(h_{full}) = 1 = \max u(H)$), while the lowest value can take different values with $1 > \min u_j(H) \geq -1$ (we have to define $\min u_j(H) < 1$ to avoid division by zero). Finally, j's utility range is given by $r_j = \max u_j(H) - \min u_j(H)$. Note that 'being dead' is not considered a health state.

The Conventional Aggregation Method (CAM)

The Conventional utility Aggregation Method (CAM) $F(.)$ defines the social value of any health state simply as the average utility, as shown below.

$$F(H) = \frac{\sum_{j=1}^{n} u_j(H)}{n} \implies F(h_i) = \frac{\sum_{j=1}^{n} u_j(h_i)}{n}$$

Multi-step utility aggregation procedure (UNPAc)

The UNPAc - Step-by-step instructions

1. **Min-Max Normalisation**

   For each individual $j \in n$, utilities are normalised between the best (full health = 1) and worst health state, so that everyone's utilities range from 1 (best) to 0 (worst).

$$u'_j(H) = \frac{u_j(H) - \min u_j(H)}{\max u_j(H) - \min u_j(H)} \tag{1}$$

2. **Relative social tariff**

   Normalised utility values are aggregated across individuals to derive the *relative* social tariff.

$$S'(H) = \frac{\sum_{j=1}^{n} u'_j(H)}{n} \tag{2}$$

3. **Scaling factor**

   Utility ranges ($r_j = \max u_j(H) - \min u_j(H)$) are aggregated across individuals to derive the scaling factor.

$$R = \frac{\sum_{j=1}^{n} r_j}{n} \tag{3}$$

4. **Rescaling**

   The relative social tariff is re-scaled to the original full-health-dead scale, using the scaling factor R, and re-anchored at full health with $1 - R$.

$$S(H) = R * S'(H) + 1 - R \tag{4}$$

Notations:

H: set of all health states $h_1, h_2, \ldots, h_k$
$u'_j(H)$: individual j's normalised preferences
$S'(.)$ relative social tariff

$u_j(H)$: individual j's health state preferences
$\min_j(H)$ / $\max_j(H)$: j's lowest/highest utility
$S(.)$ final alternative social tariff.

Table S2: Sensitivity Analysis: comparison between the absolute (relative) social value sets based on the standard approach and the UNPAc in the absence of preference heterogeneity

State	Standard approach		UNPAc approach		Difference	
	rank	value	rank	value	rank	value
FH	1	1.00 (1.00)	1	1.00 (1.00)	0	+0.00 (+0)
M	2	0.26 (0.52)	3	0.27 (0.52)	1	-0.00 (-0)
P	3	0.21 (0.48)	2	0.22 (0.49)	-1	+0.00 (+0)
PM	4	-0.52 (0.00)	4	-0.52 (0.00)	0	-0.00 (+0)

Table S3: Full UK social EQ-5D 3L value sets for the CAM and the UNPAc.

	Health	CAM	UNPAc	Difference		Health	CAM	UNPAc	Difference
Rank	state	Value	Value	Abs. (rank)	Rank	state	Value	Value	Abs. (rank)
1	11111	1.000	1.000	0.00 (-)	62	21322	0.287	0.317	0.03 (-1)
2	11211	0.912	0.910	-0.00 (-)	63	23112	0.286	0.312	0.03 (-1)
3	21111	0.877	0.878	0.00 (-)	64	23311	0.278	0.299	0.02 (+2)
4	11112	0.865	0.867	0.00 (-)	65	11131	0.277	0.304	0.03 (-1)
5	21211	0.841	0.840	-0.00 (-)	66	13221	0.274	0.301	0.03 (-1)
6	12111	0.831	0.836	0.00 (-)	67	31112	0.269	0.293	0.02 (-)
7	11212	0.828	0.829	0.00 (-)	68	13312	0.265	0.288	0.02 (-)
8	11121	0.819	0.824	0.01 (-)	69	12213	0.265	0.283	0.02 (+1)
9	12211	0.794	0.797	0.00 (-)	70	31311	0.261	0.280	0.02 (+1)
10	21112	0.793	0.797	0.00 (-)	71	21313	0.256	0.270	0.01 (+4)
11	11221	0.782	0.786	0.00 (-)	72	22321	0.253	0.285	0.03 (-3)
12	22111	0.759	0.765	0.01 (-)	73	11223	0.253	0.272	0.02 (+1)
13	21212	0.757	0.759	0.00 (-)	74	23212	0.249	0.274	0.02 (-1)
14	21121	0.747	0.754	0.01 (+1)	75	12322	0.241	0.274	0.03 (-3)
15	12112	0.747	0.755	0.01 (-1)	76	11231	0.240	0.266	0.03 (+1)
16	11122	0.735	0.743	0.01 (-)	77	23121	0.240	0.269	0.03 (-1)
17	22211	0.723	0.727	0.00 (-)	78	32111	0.235	0.261	0.03 (-)
18	21221	0.711	0.716	0.00 (+1)	79	31212	0.233	0.255	0.02 (+1)
19	12212	0.710	0.716	0.01 (-1)	80	22113	0.230	0.251	0.02 (+1)
20	12121	0.701	0.712	0.01 (-)	81	13122	0.227	0.258	0.03 (-2)
21	11222	0.698	0.705	0.01 (-)	82	31121	0.223	0.250	0.03 (-)
22	22112	0.676	0.684	0.01 (-)	83	13321	0.219	0.245	0.03 (-)
23	12221	0.664	0.673	0.01 (-)	84	21123	0.218	0.240	0.02 (-)
24	21122	0.664	0.673	0.01 (-)	85	12313	0.210	0.227	0.02 (+2)
25	22212	0.639	0.646	0.01 (-)	86	21131	0.205	0.234	0.03 (-1)
26	22121	0.630	0.641	0.01 (-)	87	23221	0.203	0.231	0.03 (-1)
27	21222	0.627	0.635	0.01 (-)	88	32211	0.199	0.223	0.02 (-)
28	12122	0.617	0.631	0.01 (-)	89	11323	0.198	0.216	0.02 (+3)
29	22221	0.593	0.603	0.01 (-)	90	13113	0.196	0.211	0.01 (+5)
30	12222	0.580	0.592	0.01 (+1)	91	23312	0.194	0.218	0.02 (-)
31	11311	0.572	0.592	0.02 (-1)	92	22213	0.194	0.213	0.02 (+1)
32	22122	0.546	0.560	0.01 (-)	93	11132	0.193	0.223	0.03 (-4)
33	22222	0.509	0.522	0.01 (+1)	94	13222	0.191	0.220	0.03 (-4)
34	21311	0.501	0.522	0.02 (-1)	95	31221	0.187	0.212	0.03 (-1)
35	11312	0.489	0.511	0.02 (-)	96	11331	0.185	0.210	0.03 (-)
36	12311	0.455	0.480	0.02 (-)	97	21223	0.182	0.201	0.02 (+1)
37	11321	0.443	0.468	0.03 (-)	98	31312	0.177	0.199	0.02 (+1)
38	13111	0.441	0.463	0.02 (-)	99	12123	0.172	0.197	0.03 (+1)
39	11113	0.419	0.434	0.01 (+1)	100	22322	0.170	0.204	0.03 (-3)
40	21312	0.417	0.441	0.02 (-1)	101	21231	0.168	0.195	0.03 (-)
41	13211	0.404	0.425	0.02 (-)	102	13213	0.159	0.173	0.01 (+5)
42	22311	0.383	0.409	0.03 (-)	103	12131	0.159	0.191	0.03 (-1)
43	11213	0.383	0.396	0.01 (+2)	104	11232	0.156	0.184	0.03 (-)
44	21321	0.371	0.398	0.03 (-)	105	23122	0.156	0.188	0.03 (-2)
45	12312	0.371	0.399	0.03 (-2)	106	32112	0.151	0.180	0.03 (-1)
46	23111	0.370	0.393	0.02 (-)	107	23321	0.148	0.175	0.03 (-1)
47	11322	0.359	0.387	0.03 (-)	108	32311	0.143	0.167	0.02 (+1)
48	13112	0.357	0.382	0.03 (-)	109	31122	0.139	0.169	0.03 (-1)
49	31111	0.353	0.374	0.02 (-)	110	22313	0.138	0.157	0.02 (+2)
50	13311	0.349	0.369	0.02 (-)	111	13322	0.135	0.164	0.03 (-1)
51	21113	0.348	0.364	0.02 (-)	112	12223	0.135	0.159	0.02 (-1)
52	23211	0.333	0.355	0.02 (+1)	113	31321	0.131	0.156	0.02 (-)
53	11313	0.327	0.340	0.01 (+2)	114	33111	0.130	0.151	0.02 (+2)
54	12321	0.325	0.356	0.03 (-2)	115	21323	0.126	0.145	0.02 (+3)
55	13212	0.320	0.344	0.02 (-1)	116	23113	0.125	0.141	0.02 (+4)
56	31211	0.316	0.336	0.02 (+1)	117	12231	0.122	0.153	0.03 (-3)
57	21213	0.311	0.325	0.01 (+2)	118	21132	0.121	0.152	0.03 (-3)
58	13121	0.311	0.339	0.03 (-2)	119	23222	0.119	0.149	0.03 (-2)
59	12113	0.302	0.321	0.02 (+1)	120	32212	0.115	0.142	0.03 (-1)
60	22312	0.299	0.328	0.03 (-2)	121	21331	0.113	0.139	0.03 (-)
61	11123	0.290	0.310	0.02 (+2)	122	31113	0.108	0.122	0.01 (+4)

Table S3: Full UK social EQ-5D 3L value sets for the CAM and the UNPAc (continued).

	Health	CAM	UNPAc	Difference		Health	CAM	UNPAc	Difference
Rank	state	Value	Value	Abs. (rank)	Rank	state	Value	Value	Abs. (rank)
123	32121	0.106	0.137	0.03 (-1)	184	33122	-0.084	-0.054	0.03 (-)
124	13313	0.104	0.117	0.01 (+4)	185	12133	-0.086	-0.061	0.03 (+1)
125	31222	0.103	0.130	0.03 (-2)	186	22332	-0.088	-0.054	0.03 (-3)
126	11332	0.101	0.129	0.03 (-2)	187	33321	-0.092	-0.067	0.02 (-)
127	22123	0.100	0.127	0.03 (-2)	188	23323	-0.097	-0.078	0.02 (+1)
128	33211	0.093	0.113	0.02 (+2)	189	32313	-0.102	-0.085	0.02 (+2)
129	23213	0.088	0.102	0.01 (+4)	190	23132	-0.102	-0.071	0.03 (-2)
130	22131	0.087	0.121	0.03 (-3)	191	23331	-0.110	-0.084	0.03 (-1)
131	21232	0.085	0.114	0.03 (-2)	192	31323	-0.114	-0.097	0.02 (+3)
132	12323	0.080	0.103	0.02 (-)	193	33113	-0.115	-0.102	0.01 (+4)
133	12132	0.075	0.110	0.04 (-2)	194	31132	-0.119	-0.090	0.03 (-2)
134	31213	0.071	0.083	0.01 (+6)	195	33222	-0.121	-0.093	0.03 (-2)
135	32221	0.069	0.099	0.03 (-1)	196	13332	-0.123	-0.095	0.03 (-2)
136	12331	0.067	0.097	0.03 (-1)	197	12233	-0.123	-0.099	0.02 (-1)
137	13123	0.066	0.087	0.02 (+1)	198	31331	-0.127	-0.103	0.02 (-)
138	23322	0.064	0.094	0.03 (-2)	199	21333	-0.132	-0.113	0.02 (+1)
139	22223	0.064	0.089	0.03 (-2)	200	23232	-0.139	-0.109	0.03 (-1)
140	32312	0.059	0.086	0.03 (-1)	201	32123	-0.140	-0.115	0.02 (-)
141	13131	0.053	0.081	0.03 (+1)	202	33213	-0.152	-0.140	0.01 (+3)
142	22231	0.051	0.083	0.03 (-1)	203	32131	-0.153	-0.121	0.03 (-1)
143	31322	0.047	0.075	0.03 (-)	204	31232	-0.155	-0.128	0.03 (-1)
144	33112	0.046	0.070	0.02 (+1)	205	22133	-0.158	-0.131	0.03 (-1)
145	12232	0.038	0.072	0.03 (-1)	206	33322	-0.176	-0.149	0.03 (-)
146	33311	0.038	0.057	0.02 (+1)	207	32223	-0.176	-0.153	0.02 (-)
147	23313	0.033	0.046	0.01 (+4)	208	12333	-0.178	-0.155	0.02 (-)
148	11133	0.032	0.051	0.02 (+1)	209	32231	-0.189	-0.160	0.03 (-)
149	13223	0.029	0.049	0.02 (+1)	210	13133	-0.192	-0.172	0.02 (+2)
150	21332	0.029	0.058	0.03 (-4)	211	23332	-0.194	-0.165	0.03 (-1)
151	32122	0.022	0.056	0.03 (-3)	212	22233	-0.194	-0.170	0.02 (-1)
152	13231	0.016	0.042	0.03 (+1)	213	33313	-0.207	-0.196	0.01 (+1)
153	31313	0.016	0.027	0.01 (+4)	214	31332	-0.211	-0.184	0.03 (-1)
154	32321	0.014	0.043	0.03 (-2)	215	13233	-0.229	-0.210	0.02 (+2)
155	33212	0.009	0.031	0.02 (+1)	216	32323	-0.232	-0.209	0.02 (-)
156	22323	0.008	0.033	0.02 (-1)	217	32132	-0.236	-0.202	0.03 (-2)
157	22132	0.003	0.040	0.04 (-3)	218	32331	-0.245	-0.215	0.03 (-)
158	33121	-0.000	0.027	0.03 (+1)	219	33123	-0.245	-0.226	0.02 (+1)
159	22331	-0.005	0.027	0.03 (-1)	220	22333	-0.250	-0.226	0.02 (-1)
160	11233	-0.005	0.013	0.02 (+3)	221	33131	-0.258	-0.232	0.03 (-)
161	23123	-0.005	0.017	0.02 (-)	222	23133	-0.263	-0.242	0.02 (+1)
162	32113	-0.010	0.009	0.02 (+3)	223	32232	-0.273	-0.241	0.03 (-1)
163	32222	-0.015	0.018	0.03 (-3)	224	31133	-0.280	-0.261	0.02 (-)
164	12332	-0.017	0.016	0.03 (-2)	225	33223	-0.282	-0.264	0.02 (-)
165	23131	-0.018	0.010	0.03 (-1)	226	13333	-0.284	-0.266	0.02 (-)
166	31123	-0.022	-0.003	0.02 (+2)	227	33231	-0.295	-0.270	0.02 (-)
167	13323	-0.026	-0.007	0.02 (+2)	228	23233	-0.300	-0.280	0.02 (-)
168	13132	-0.031	-0.000	0.03 (-1)	229	31233	-0.316	-0.299	0.02 (+1)
169	22232	-0.033	0.002	0.03 (-3)	230	32332	-0.328	-0.297	0.03 (-1)
170	31131	-0.035	-0.009	0.03 (-)	231	33323	-0.337	-0.320	0.02 (+1)
171	33221	-0.037	-0.012	0.03 (-)	232	33132	-0.342	-0.313	0.03 (-1)
172	13331	-0.039	-0.013	0.03 (-)	233	33331	-0.350	-0.326	0.02 (-)
173	21133	-0.040	-0.019	0.02 (-)	234	23333	-0.355	-0.336	0.02 (-)
174	23223	-0.042	-0.022	0.02 (-)	235	31333	-0.372	-0.355	0.02 (+1)
175	33312	-0.046	-0.024	0.02 (-)	236	33232	-0.379	-0.351	0.03 (-1)
176	32213	-0.046	-0.029	0.02 (+1)	237	32133	-0.398	-0.374	0.02 (-)
177	23231	-0.055	-0.028	0.03 (-1)	238	33332	-0.434	-0.407	0.03 (-)
178	31223	-0.058	-0.041	0.02 (+2)	239	32233	-0.434	-0.412	0.02 (-)
179	11333	-0.060	-0.043	0.02 (+2)	240	32333	-0.490	-0.468	0.02 (-)
180	13232	-0.068	-0.039	0.03 (-1)	241	33133	-0.503	-0.484	0.02 (-)
181	32322	-0.070	-0.038	0.03 (-3)	242	33233	-0.540	-0.522	0.02 (-)
182	31231	-0.071	-0.047	0.02 (-)	243	33333	-0.595	-0.578	0.02 (-)
183	21233	-0.077	-0.057	0.02 (+2)					

Conclusion to Chapter 3

Health state utility values must have boundaries - they need to be scaled or normalised in some fashion to allow for interpersonal comparability. A key insight from this chapter is that the kind of normalisation that is currently applied, the full health (=1) to dead (=0) scale, is arbitrary. There is no sound, compelling reason why the utilities should be anchored at full health and dead. Moreover, the results clearly show that the type of normalisation that is used to specify health state utilities can have considerable impact on the aggregate social value set. The proposed UNPAc method employs a min- max normalisation, such that individuals' utility values lie on a 0-1 scale, before aggregating health state utilities into a social value set. Only in a second step is the full health - dead anchor re-introduced, to map the resulting values on to the QALY scale. Whether or not this type of aggregation produces more appropriate social values is an open, normative, question. It critically depends on the role of 'dead' in the elicitation process. The next chapter discusses this issue in more detail.

Chapter 4

Setting Dead at Zero? On the contingency of the utility unit scale

BACKGROUND

In the QALY model, health state utilities are commonly measured on a scale that is anchored at full health, set to a value of one, and dead, set to zero. Even though this practice is adopted almost universally, a theoretical basis for it appears to be missing (Sampson et al., 2018). In their literature review, Roudijk et al. (2018) report that most authors do not justify setting dead to zero, and even those who do, merely state that this is done 'by definition', 'by convention', or simply, 'for convenience'. Notwithstanding, some authors have proposed theoretical arguments for why dead has to be set to zero and not to any other value. In this brief report, I revisit and rebut the four arguments known to the authors, to demonstrate that anchoring the utility scale on a different state may well be permissible.

ARGUMENT 1: UTILITIES ON A RATIO SCALE, DEAD AS A NATURAL ZERO POINT

The first argument is taken from Roudijk et al. (2018) They state that, in order for the QALY model to satisfy basic principles of rationality, utilities must be measured on a ratio scale, for which dead is as a natural zero point.

Let's consider the first part: is dead a natural zero point? For physical quantities, such as mass or temperature, zero has an unambiguous meaning. It defines the point at which there is no mass or no molecular motion. Yet, the position of dead on the utility scale seems to have a different function. Most people consider certain states worse than dead, so dead does not mark the state with the lowest utility value. Instead, it divides the scale into states with positive (better than dead) and negative values (worse than dead). This seems incompatible with the notion of a natural zero, like in 0° Kelvin (= absence of molecular motion), but rather appears to be a zero with an arbitrary reference, like in 0° Celsius or 0° Fahrenheit (= freezing point of water or salty water).

Returning to the first part of the argument, must utilities be measured on a ratio scale? For an interval scale to be admissible as the basis for the QALY model, preferences must be invariant to positive affine transformations, i.e. $f(A) \sim f(B) \rightarrow f'(A) \sim f'(B)$, for any $f'(x) = a * f(x) + b$) (Von Neumann & Morgenstern, 1966, pp. 15-29; Fleurbaey & Hammond, 2004, pp. 1179). This is a basic rationality requirement, and, if Roudijk et al. (2018) were right in that it is violated, the interval scale may be inappropriate to measure health state utilities (see Table 1 for an overview of measurement scales and their properties). In the following, we reproduce the example given in their paper (with minor adaptations) and show that their reasoning is flawed.

Table 1: Measurement scales and their properties

Scale	Description	Examples	Invariance
Nominal	Qualitative classification	gender, colour	n.a.
Ordinal	Rank order, distances between ranks are not known or nor defined	Likert scale, quantile rankings	$u(x)^2, \log(u(x))$
Interval	Ranking with meaningful differences, ratios and the zero point have no meaning (20°C is 15°C warmer than 5°C, but it is not four-times warmer)	°C, °F, vNM utilities	3x+4
Ratio	most informative scale; order, differences, ratios, and the zero point are meaningful	Meters, gramm, °Kelvin	3x

The QALY model is represented mathematically as: $Q = u(h_i) * v(t_i)$, whereby $u(h_i)$ denotes the utility derived from state i, and $v(t_i)$ is a function of the time spent in that state – note that in the standard model v(t) = t.

Now, suppose Alice is indifferent between living 10 years in full health with $u(h_{full}) = 1$, followed by 10 years in state i, with $u(h_i) = 0.7$ (A), and living 15 years in full health, followed by five years in state j, with $u(h_j) = 0.4$ (B). Both options yield 17 QALYs, as shown in equations A.1 and B.1 below.

A.1: $1 * 10 + 0.7 * 10 = 17$

B.1: $1 * 15 + 0.4 * 5 = 17$

To show that a positive affine transformation (i.e. $f'(x) = a * f(x) + b$) does not preserve indifferences, Roudijk et al. (2018) shift the origin of $v(t)$, with $v'(t) = t + 2$, and derive the following result:

A.2: $1 * (10 + 2) + 0.7 * (10 + 2) = 20.4$

B.2: $1 * (15 + 2) + 0.4 * (5 + 2) = 19.8$

Alice now appears to prefer A over B, and Roudijk et al. conclude that an interval scale is inadmissible as a basis for the QALY. Therefore, they argue, utilities must lie on a ratio scale.

Yet, their algebra is flawed: note that in A, Alice spends zero time in state *j*. In the standard model, any state in which zero time is spent can be omitted from the equation (because $u(hj) * v(tj) = 0.4 * 0 = 0$). However, after the origin is shifted, this is no longer allowed, because now v 0 (0) = 2. State *j* must thus be considered in A, as must state *i* in B.

When the positive affine transformation is applied consistently, Alice's transformed utility function is represented by the following equations:

A.3: $1 * (10 + 2) + 0.7 * (10 + 2) + 0.4 * (0 + 2) = 21.2$

B.3: $1 * (15 + 2) + 0.4 * (5 + 2) + 0.7 * (0 + 2) = 21.2$

Alice's indifference is well preserved, and a compelling reason for rejecting the interval scale as a basis for the QALY model cannot be found.

ARGUMENT 2: DEAD AND ZERO TIME INDIFFERENCE

Another interesting mathematical argument for setting dead to zero has been derived from Miyamoto et al.'s (1998) seminal work on 'the zero-condition': it is maintained that all health states are equally preferred when their duration is 0, i.e. $u(h_i) * v(0) = u(h_j) * v(0)$ for any states *i* and *j*. Taking this a step further, Roudijk et al. (2018) claim that indifference should also hold for a choice between being in state i for a duration of 0 (followed by death) and being dead for some time t, i.e. $u(h_{dead}) * v(t) = u(h_i) * v(0)$. For this equation to hold true – for any duration t and any state i – the utility of dead must be zero: $u(h_{dead}) = 0$.

While this may sound logical at first, the premise of the argument appears dubious. The presented alternatives are not mutually exclusive. In fact, they are identical: being dead for some time *t* involves spending zero time in state *i*. Which also involves spending zero time in state *j*. One does not have to forgo one for the other. But preferences can only be meaningfully specified, if there is a choice involved. It thus seems impossible to postulate any indifferences here, and insights about the value of dead can not be derived.

ARGUMENT 3: STREAMS OF INFINITE NON-ZERO UTILITIES

The third argument comes from the third edition of Drummond et al's (2005. p. 176) standard textbook 'Methods for the Economic Evaluation of Health Care Programmes' (not included in the 4th edition):

> "[I]f any score other than zero were used for death, [...] the (nonzero) death score would be assigned to the state of death for each year off into the future for as long as the dead lasted (that is, forever). Thus, the analyses would have streams of numeric outcomes going to infinity - not a pretty picture. Accordingly, zero is the only practical score that can be used for death."

First of all, it should be noted that an infinite stream of zero utility values is also undesirable, because, strictly speaking, the product of zero and infinity is not defined. However, the concerns about infinite streams of non-zero utilities are unfounded. Future utilities are usually discounted, which causes non-zero utilities to tend towards zero. This prevents any infinite values from occurring. But even without discounting, infinite streams of non-zero utilities from being dead are not a problem, because they occur in all alternatives. Since economic evaluations are only concerned with the differences between alternatives (the increments), those non-zero utilities cancel out and can be discarded.

ARGUMENT 4: THE ONTOLOGICAL ZERO

The fourth and final argument is less mathematical and more ontological. It states that once you are dead, you cannot experience utilities anymore and, thus, zero is the only plausible value for dead (Devlin et al., 2004).

The argument is problematic for two reasons. Firstly, it conflates disparate concepts of utilities: the notion of (absolute) experienced utility used here for the dead state might be incompatible with the (relative) decision utilities measured for any other health state. Secondly, the argument might also be objectionably paternalistic. There are many different ideas of death. When people value the dead state using a visual analogue scale (0-100), their responses, unsurprisingly, vary. For example, one participant in an EQ-5D health valuation study commented that dead can have two values, 100 for ’dead in heaven’, and 0, for ‘dead in hell’ (Sampson et al., 2018). It would be presumptuous to discard this and other beliefs about death, and to simply assume that everyone shares a supposedly scientific zero utility valuation of dead.

WHY DOES THE VALUE OF DEAD MATTER?

Social values for health states are commonly derived by averaging over the utilities from different individuals. For this operation to be permissible, units of utilities must be measured on the same scale and interpersonally comparable (Fleurbaey & Hammond, 2004, pp. 1179; Stevens, 1946). To use an analogy, it would be meaningless to take the average of a range of temperature measurements, some in degrees Kelvin, others in degrees Celsius, and still others in degrees Fahrenheit. For the valuation of health, a common unit scale is enforced by anchoring everyone’s utilities at two points; by convention, full health and dead. The distance between those two is assumed to be the same for everyone and comparable across individuals.

It is not within the scope of this paper to further discuss the intricate problem of interpersonal utility comparisons, but it should be noted that, depending on which anchor points are chosen, aggregate social value sets may differ (Devlin et al., 2017). Therefore, the anchor points matter, and should not be assigned arbitrarily. While the use of full health as the upper anchor point seems indisputable - it is a dominant state, which should be weakly preferred over all other states -, it seems to be a matter of debate whether dead is an (or the only) appropriate lower anchor point. At least the four arguments considered in this paper fail to provide an unequivocal basis for setting dead to zero.

The results of this paper do not imply that dead must not be used as an anchor point, yet, they suggest that dead may not have to be used as one. Relaxing this property of the QALY model may open up new possibilities to develop alternative value frameworks and re-consider the role of states worse than dead. In any case, an open, impartial discussion about appropriate utility scales may be more valuable than trying to (ex-post) justify pragmatic decisions, taken decades ago, in the early days of the field. Even if dead is to be kept as an 'absolute zero', this work may have highlighted a weakness in the conceptual foundations of health economics, and, hopefully, it sparks more interest in this topic.

REFERENCES

Barrie S. QALYs, euthanasia and the puzzle of death. Journal of Medical Ethics. 2015 Aug;41(8):635-38.

Devlin NJ, Hansen P, Selai C. Understanding health state valuations: a qualitative analysis of respondents' comments. Quality of Life Research. 2004 Nov;13(7):1265-77.

Devlin NJ, Shah KK, Buckingham K, et al. What is the normative basis for selecting the measure of 'average'preferences for use in social choices? OHE research paper. Office of Health Economics, London. 2017.

Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the economic evaluation of health care programmes. 3rd ed. Oxford University Press; 2005. p. 176.

Fleurbaey M, Hammond PJ. Interpersonally comparable utility. In: Handbook of utility theory. 2004. p. 1179-85. Stevens SS. On the theory of scales of measurement. 1946.

Miyamoto JM, Wakker PP, Bleichrodt H, Peters HJM. The zero-condition: a simplifying assumption in QALY measurement and multiattribute utility. Management Science. 1998 Jun;44(6):839-49.

Roudijk B, Donders AR, Stalmeier PF. Setting dead at zero: applying scale properties to the QALY model. Medical Decision Making. 2018 Aug;38(6):627-34.

Sampson CJ, Devlin NJ, Parkin DW. Drop dead: is anchoring at 'dead' a theoretical requirement in health state valuation? 35th EuroQol Group Scientific Plenary. 2018.

Von Neumann J, Morgenstern O. Theory of games and economic behavior. 1966. p. 15-29.

Chapter 5

The QALY is ableist

The following chapter investigates the ethical implications of using health states worse than dead, and, as a consequence, negative QALYs in cost-effectiveness analysis. I argue that this practice is problematic, as it implies that the life of individuals in such states is not worth living, without considering the individual patient's perspective. It is demonstrated how negative QALY can result in a systematic underestimation of the value of life-extending treatments. Finally, it is concluded that states worse than dead should no longer be used, and a non-negative value should be placed on all human lives.

This chapter has been published in an identical form as:

Schneider P. The QALY is ableist: on the unethical implications of health states worse than dead. Quality of Life Research. 2022 May;31(5):1545-52. https://doi.org/10.1007/s11136-021-03052-4

Quality of Life Research (2022) 31:1545–1552
https://doi.org/10.1007/s11136-021-03052-4

The QALY is ableist: on the unethical implications of health states worse than dead

Paul Schneider[1]

Accepted: 23 November 2021 / Published online: 9 December 2021

Abstract

Introduction A long-standing criticism of the QALY has been that it would discriminate against people in poor health: extending the lives of individuals with underlying health conditions gains fewer QALYs than extending the lives of 'more healthy' individuals. Proponents of the QALY counter that this only reflects the general public's preferences and constitutes an efficient allocation of resources. A pivotal issue that has thus far been overlooked is that there can also be negative QALYs.
Methods and results Negative QALYs are assigned to the times spent in any health state that is considered to be worse than dead. In a health economic evaluation, extending the lives of people who live in such states reduces the overall population health; it counts as a loss. The problem with this assessment is that the QALY is not based on the perspectives of individual patients—who usually consider their lives to be well worth living—but it reflects the preferences of the general public. While it may be generally legitimate to use those preferences to inform decisions about the allocation of health care resources, when it comes to states worse than dead, the implications are deeply problematic. In this paper, I discuss the (un)ethical aspects of states worse than dead and demonstrate how their use in economic evaluation leads to a systematic underestimation of the value of life-extending treatments.
Conclusion States worse than dead should thus no longer be used, and a non-negative value should be placed on all human lives.

Keywords Bioethics · Health economics · Health valuation · QALY · States worse than dead · Utilities

Introduction

The concept of Quality-Adjusted Life Years (QALYs) is being widely used to inform societal decisions about the allocation of health care resources [1]. By some, it is even considered the 'gold standard' for measuring and valuing health in economic evaluations [2]. However, it is not without limitations: a long-standing line of critique has been that the QALY discriminates against people with disabilities and those in poor health [3]: all else being equal, extending the lives of individuals with disabilities or underlying health conditions gains fewer QALYs than extending the lives of 'more healthy' individuals. Several authors have argued that this is unjust, and that all life years should be of equal value [4–6].

Proponents of the QALY framework counter that since most people state that they are willing to give up some of their remaining lifetime for improvements in their health-related quality of life (HRQoL), it is only rational that one additional life year in poor health is of lower value than one additional life year in perfect health. Discrimination based on individuals' HRQoL is then necessary in order to allocate resources most efficiently [7–10].

One pivotal issue that has thus far not been considered in this debate is that HRQoL can not only be low, but also negative: 'health states worse than dead' (SWD) get assigned negative values. Extending the live of a person who lives in a SWD generates negative QALYs.

While largely overlooked by previous research, the implications of SWDs are significant. Their use in health economic evaluations implies value judgements that, at closer inspection, appear to be ableist and unethical. Furthermore, they lead to the systematic underestimation of the value of

✉ Paul Schneider
p.schneider@sheffield.ac.uk

[1] School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, 30 Regent St, Sheffield S1 4DA, UK

life-extending treatments in almost any patient group. In this paper, I thus argue that the concept of SWD should be abandoned.

The sophistication and complexity of health economic evaluations can make it difficult to examine their implicit value judgements [11]. The remainder of this paper thus begins with a background section, in which some key concepts are revisited ("Background" section). In "Motivating example—Step I" section, a very simple motivating example is provided, which is used to develop some intuition for the ethical implications of SWD. The subsequent section ("SWD and the conflict between individual and social preferences" section) is a brief digression to clarify potential misconceptions about social value sets. Only then I expand the example from "Motivating example—Step I" section, to demonstrate the, perhaps somewhat intuitive, effects of SWD on the group-level ("Motivating example—continued" section). In "Discussion and further considerations" section, I discuss the implications of and solutions for the issues raised.

Background

The valuation of health

The QALY is defined as the arithmetic product of survival time and HRQoL. HRQoL, in turn, is determined by the health state an individual is living in. This means, 'measuring' QALYs usually involves two components: firstly, a set of health states; and secondly, numeric scores that reflect their respective desirability. These values are often also referred to as utilities, social values, preference-, (health-related) quality of life-, or QALY-weights. Customarily they are supposed to reflect the preferences of the general public [12].

There are many different ways to classify health states (such as EQ-5D, SF-6D, or HUI), and various methods to derive numeric score/social values for them (such as time trade-off (TTO), standard gamble, or discrete choice experiments) [13]. The arguments of this paper are relevant to all of them, but for simplicity, I will only refer to EQ-5D-3L system and the TTO method, as those are currently used as the reference case in the UK [14].

In a TTO exercise, individual preferences for health states are elicited by identifying points of indifference between a longer life in poor health, and a shorter life in perfect health [15, 16]. Preferences are measured in terms of utility values on a scale that is anchored at perfect health, which is assigned a value of 1, and dead, which is assigned a value of 0. The social value of any health state is then constituted by the average utility [17, 18].

Negative utilities for SWD

If an individual states that they prefer immediate death over living any amount of time in state j, this state is considered to be worse than dead. The point of indifference is then derived from the number of life years in full health a person would be willing to give up to avoid living in that state for a certain number of years. If, for example, a person is indifferent between living 5 years in perfect health (followed by death), and living 10 years in perfect health, followed by 10 years in some health state j (then followed by death), it is inferred that state j has a utility of − 0.5 ($5 \times 1 \sim 10 \times 1 + 10 \times j => j = -0.5$).

It may be interesting to note that negative utilities have different characteristics than their positive counterparts. Positive utilities are measured as a proportion of the utility for full health, with an upper limit of 1. Negative utilities are much harder to interpret and have no limit. Theoretically, they can take the value of minus infinity. In practice, this can cause problems, because very low negative values can have significant impact on the estimation of the average utility values. To limit their influence, negative utilities are usually constrained (rather arbitrarily) to a lower limit of − 1, either by choosing an experimental design that does not allow for lower values, or by rescaling lower negative values, after they are collected [19, 20].

Motivating example—Step I

Suppose Alice has a severe health condition called D, and, according to some social value set, her health state has a value of − 0.1. With the current standard treatment (alternative A), she will be able to live 10 years in her current state before she dies. Now, suppose a new treatment (alternative B) becomes available, which prolongs Alice's life by 10 more years, i.e. giving her 20 years in total, but it has no effect on HRQoL. Further suppose that the new treatment costs exactly the same as the old treatment—it does not incur any additional costs.

An economic evaluation that weighs the costs and the benefits of the two alternatives will come to the conclusion that, compared to the old treatment A, the new treatment B generates − 1 QALY at no cost (see below). This means, alternative B is not only not cost-effective, but it is dominated by A. Assuming a threshold of £ 20,000 per QALY, the new treatment would need to save more than £20,000, before it would be considered cost-effective [21]. Based on this economic evaluation, the recommendation would unmistakably be not to provide the new treatment to Alice.

$$\Delta Q_{\mathrm{B}} = \frac{c_{\mathrm{B}} - c_{\mathrm{A}}}{s_{\mathrm{B}} * q_{\mathrm{B}} - s_{\mathrm{A}} * q_{\mathrm{A}}} = \frac{0}{(-0.1) * 20 - (-0.1) * 10} = \frac{0}{-1} \rightarrow \text{dominated}$$

ΔQ is the incremental cost-effectiveness ratio; c is the costs; q is the HRQoL; s is the survival time; subscripts A and B indicate the respective alternatives.

The outcome of the economic appraisal seems striking. The new treatment would extend Alice's survival time by 10 years, it is available at no extra cost, and Alice might be desperate to receive the treatment, yet, society considers Alice's health state to be worse than dead. Based on this evaluation, the treatment is withhold from her.

It seems obvious that, in this simple example, the value judgement implicit in SWD is unethical. The negative HRQoL suggests that Alice's health state is worse than dead—but maybe not for her. As a matter of fact, Alice herself might well enjoy life [22]. Even if her health state causes severe suffering, there might be numerous other good reasons for her to seek life-extending treatment (faith, meaning, family, etc.). It should be self-evident that it is not for society to decide whether or not Alice's life is worth living. To do so would be a blatant violation of her autonomy [23–25]. If she is willing to receive the life-saving treatment, society seems to have no right to deny its provision.

Note that this only holds unequivocally if the new treatment is not more expensive than the old treatment. If the treatment were more costly, the question if, and if so, how much society should be willing to spend to save Alice is a separate issue. It might then be legitimate to decide that saving Alice is not the most efficient use of resources. Yet, given that society is willing to pay for the current treatment, it would be unethical to withhold the new treatment from her.

SWD and the conflict between individual and social preferences

Before we further expand the example, it will be useful to clarify some potential misconceptions about the type and the admissible domain of the preferences that underlie social value sets/HRQoL values and the QALY.

Generally, social value sets are based on the preferences of the general public [26]. In fact, most national HTA agencies make this explicitly the reference case for health economic evaluations—one notable exception is Sweden, which uses patient preferences (see below) [27]. In a publicly financed health care system, this seems desirable from a democratic perspective. Citizens—sometimes confused with 'taxpayers' (e.g. [28])—should have some say in decisions about the allocation of health care resources [18, 26]. It may thus be legitimate to use health states preferences of the general public to inform societal decision-making. When it comes to SWD, however, the preferences of the general public are (1) ill-informed, (2) misconstrued, and/or (3) irrelevant. In the following, I shall further elaborate on these three points.

1. Ill-informed: The preferences of the general public do not correspond to patients' evaluation of their own situations; they should not be confused with a measure of patients' self-assessed HRQoL.

Members of the general public usually have little or no experience with severe health problems. When asked to imagine living 10 years with impaired mobility, for example, they tend to focus on the immediate negative impact that the loss of mobility might have of their life now. Yet, they fail to consider all the other relevant aspects that do not change—or even improve. As a result members of the general public generally overestimate the impact of health impairments. They give significantly lower health state utilities than people who actually live in those health states [29].

The Swedish, experience-based value set demonstrates the difference very clearly. For this study, Burström et al. [30] asked about 45,000 individuals in Sweden to value the (EQ-5D-3L) health state they are currently in, using the TTO method. The experience-based value set they derived is strikingly different from value sets that are based on the preferences of the general public, in that it did not contain any SWD. With a value of 0.34, even the worst health state had a relatively high value.

For comparison, the UK social value set (which is based on the preferences of the general public) contains 84 SWD—that is 34.6% of all the 243 health states that the EQ-5D-3L system can describe [31, 32]. The proportion of SWD varies greatly between countries, ranging from 2% in Zimbabwe to 60% in Singapore [33, 34]. According to the UK social value set, about 1.5% of the adult population in England, that is approximately 840,000 individuals, are currently considered to be living in a SWD (own analysis, [35]). Among patients, the proportion is likely to be much higher.

On a side note, it should be mentioned that people's adaptation to poor health and disability are sometimes also viewed as problematic. It is argued that patients' utility values could be higher only because of lowered expectations, cognitive denial, or some other bias, that leads patients to underestimate how much they would benefit from improvements of their health states. It may then not be desirable to take patients' utilities at face value [36]. Nonetheless, in the context of SWD, this argument seems hardly plausible. If a patient thinks their life is worth living, it would be absurd to consider them factually mistaken, and to maintain that they are objectively better off if they were dead.

2. Misconstrued: Social value sets do not reflect the general public's preferences for the allocation of resources.

It could be argued that social value sets are not supposed to reflect how individuals experience certain health states, but to reflect social preferences for the allocation of health care resources [29]. If that is the case, social value sets are falsely constructed and clearly misspecified.

Participants in health valuation studies are not asked how they prefer resources to be allocated. This would require using a method like the person trade-off, for example, in which participants are asked to make choices about two groups of people, which differ in size and in their health states [37]. Instead, TTO or SG are used, which ask participants to imagine being in a particular health state themselves. Yet, this one type of preferences can not easily be translated into another. Some people may, for example, say that they would rather prefer to be dead, than to be confined to bed [38]. Yet, the very same people will probably consider their preferences being misrepresented, if they led to the evaluation that people who are confined to bed should not be offered life-extending treatments. They may rightly object that this is just not what they meant.

3. Irrelevant: Even if social value sets would accurately reflect the general public's preferences, in the context of SWD, those should be considered irrelevant.

It seems rather improbable that members of the general public in the UK, or anywhere else for that matter, would actually support the concept of SWD and their implicit value judgement—which we will discuss in more detail in the next section. However, even if some individuals wanted some other individuals to die earlier rather than later, those preferences should be deemed irrelevant for treatment reimbursement decisions.

While everyone has, of course, the right to consider their own life in a certain health state to be worse than dead and to refuse life-extending treatments, considering someone else's life in a certain health state worse than dead is morally a completely different issue. To then also prefer that life-extending treatments are withheld from certain (other) individuals, because one prefers them to be dead, would undoubtedly be reprehensible. It would constitute an objectionable preference [11, 23].

To clarify, this paper neither tries to argue that SWD do not exist, nor to promote treating people in poor health states, who do not want to be treated. The focus of this paper is on societal reimbursement decisions—i.e. should a given life-extending treatment be made available in the health care system, in case an individual seeks it. Whether or not their life is worth extending, and the treatment is actually taken, is for to the individual to decide. If they do not wish to prolong their lives in poor health states, they can, of course, refuse to take the treatment and/or choose to stop the treatment at any time [25]. The point I am trying to make is that whether the general population considers these health states better or worse than dead should be considered irrelevant in this context.

Of course, in some cases individuals are not able to express their own will (e.g. young children, unconscious patients, etc.), which often poses complex ethical challenges. Yet, these lie outside the scope of this paper and will not be discussed here. It should only be noted that in these situations, decisions ought to be made on the individual's behalf ('what would they have decided?')—Social health state values, which are based on the preferences of the general public, do not appear to be particularly helpful to inform such decisions.

In liberal societies, individual rights set boundaries for the realisation of preferences and constrain what can be done in pursuit of collective interests. This means, restrictions are imposed on the domain of preferences to protect individual rights. Certain types of preferences, say for sexism, racism, genocide, or tyranny, are being discarded as objectionable and ignored in societal decision-making: it just does not matter how many people prefer that health care is only provided to people of a certain ethnicity or how strong their preferences are. Such views are simply not taken into account. This means, even if some individuals preferred that some other individuals in SWD do not get access to life-extending treatments, their preferences should be considered objectionable and be discarded.

Motivating example—continued

Step II

The example given above may not seem particularly relevant, as QALYs are not evaluated on the individual-level. Treatment reimbursement decisions are, accordingly, also not made for single individuals, but only for groups. However, by incrementally expanding the simple example I will try to show that the intuition developed for the individual case also applies to the aggregate level. That is to say, if one accepts that it would be unethical to withhold the life-extending treatment from Alice in the example above, it follows that one also has to reject the use of SWD in health economic evaluations altogether.

To value health outcomes for a group, HRQoL values are aggregated, across many different individuals and over time. The resulting 'disease state utilities' usually reflect the average HRQoL of a group of patients with some disease. Commonly used disease states include, for example, 'pre-progression' and 'post-progression' in lung cancer; or 'mild', 'moderate', and 'severe' in COPD.

If now there was a group of individuals, of which all, like Alice, live in SWD, it seems obvious that the arguments made above still apply. That is to say, if one accepts that society should provide life-extending treatment for Alice—if the treatment is not more expensive than the current standard of care—society should obviously do the same for each member of the group.

Now we will take the scenario one step further and show that SWD can also have significant implications for individuals who live in states that are better than dead (SBD), and that they can affect decisions for new treatments that are more costly than the current treatment.

Step III

Suppose Bob, and Claire are a group of patients with some chronic disease *D*. They live in health states with HRQoL values of + 0.2, and + 0.4, respectively. The average HRQoL for disease *D* is then given by $\frac{0.2+0.4}{2} = 0.3$. With the current standard treatment, they are both expected to live 10 years before they die.

Further suppose that a new life-extending treatment *C* becomes available (again, with no effect on HRQoL), which prolongs the lives of patients with disease *D* by another 10 years, i.e. giving them 20 years in total. The treatment is if $£19,000$ more expensive than the standard treatment.

Still assuming a threshold of $£20,000$ per QALY, we can derive an incremental cost-effectiveness ratio (ICER) of $\frac{£38,000}{0.3\times10} = £12,667$ per additional QALY. Consequently, treatment *C* would be considered cost-effective.

Step IV

Suppose that Alice, still living in a state with a HRQoL of − 0.1, also has disease *D*, and that she joins the group of Bob and Claire. The average HRQoL for disease *D* is then given by $\frac{-0.1+0.2+0.4}{3} = 0.167$. Now, the ICER increases to $\frac{£38,000}{0.167\times10} = £22,754$ per QALY and treatment *C* suddenly is no longer cost-effective.

This evaluation should be considered unethical. The average utility value of 0.167 reflects a mixture of Bob's and Claire's positive, and Alice's negative HRQoL values. Thereby, the willingness to pay for an additional life year in that group is reduced proportional to Alice's negative HRQoL. The implications are significant: Treatment *C* is not provided to the patients with disease *D*, only because society prefers Alice to die sooner rather than later—the decision is made *as if* Alice's life were considered unworthy of living.

If society were indifferent whether Alice dies or lives, i.e. her health state had a value of 0, the treatment would become cost-effective. The average HRQoL of disease *D* would then increase to $\frac{0+0.2+0.4}{3} = 0.2$, and the ICER would drop under 20,000 again, with $\frac{£38,000}{0.2\times10} = £19,000$ per QALY. What this result suggests is health economic evaluations may systematically underestimate the value of any life-extending medical intervention.

Discussion and further considerations

This paper has demonstrated that when SWD are used to value changes in survival times, they imply unethical value judgements and discriminate against those people in poor health states. This holds true, regardless of whether SWD occur on the individual-level, where they are immediately visible, or on the group-level, where they may be hidden within an aggregate average. I thus argue that SWD should not be used in health economic evaluations. Extending a person's life should generate at least zero QALYs, and shortening should not gain any QALYs, respectively.

This position does not seem to be controversial: while there may be reasonable disagreement over the relative value of life years gained in one group compared to another, an additional life year should never be considered a loss for society in itself. Yet, as the examples in this paper have shown, this is exactly what SWD imply. It therefore seems striking how widely and uncritically SWD have been and are being used in health economic evaluations. It can only be attributed to the complexity of economic modelling, which may conceal the implicit value judgements, that there has not been an outrage from the general public, patient advocacy groups, and/or health economists.

Some may argue that it is not immediately clear if, and if so, to what extent the thesis of this paper applies to decision-making in the real world. HTA agencies surely will recognise that it would be deeply problematic to estimate the QALY gains from, say, providing feeding tubes for children with severe birth defects, or mechanical ventilation for patients with advanced amyotrophic lateral sclerosis. Life-extending treatments like these for people in severe health states are likely to be provided, even if they are clearly not cost-effective (according to the current QALY framework). This means, the arguments raised in this paper are mainly relevant to those cases where the unethical implications of the QALY framework are not obvious; where the QALY losses from extending the lives of people in SWD are concealed from the decision makers. SWD may then lead to an underestimation of the value of a life-extending treatment. People in SWD living for longer cause the average ICER estimate to be higher, without anyone noticing it, and, most importantly, (presumably) without anyone's intention for it to be the case.

I would like to stress that the ICER estimates of almost any life-extending treatment can potentially be affected by SWD. As mentioned above, SWD are not uncommon: 1.5% of the English adult population lives in a SWD. The prevalence among patients can be assumed to be much higher, but detailed information on SWD is scarce. In some rare instances, SWD can be spotted directly by inspecting the economic model. To give but one example, in NICE's 2019 appraisal of Nusinersen for treating spinal muscular atrophy, four of seven non-dead states had a negative value in the reference scenario for one of the subgroups—increased survival time in these states led to a lower QALY estimate [39]. However, most often, one will need to assess the disaggregated data on patients' self-reported health states to identify SWD in the underlying patient population, because even if aggregate utility scores are positive, they may well be affected by SWD: Scott et al. [40], for example, report a median utility score of 0.36 in a sample of 2073 patients awaiting total hip arthroplasty. Yet, they also found that 18.9% of the patients reported to live in state with a negative utility value. Unfortunately, this information is usually not disclosed separately, and so the magnitude of the effect remains largely unknown.

On the other hand, there does not seem to be any compelling reason to use SWD in health economic evaluations to value additional survival time in the first place. SWD neither reflect the preferences of individual patients, nor can they be considered to represent the general public's preferences for the allocation of health care resources—so why are we using them?

It should be noted that SWD can also give people in poor health states an advantage. Moving someone from a SWD to full health for, say one year, actually generates more QALYs than extending the life of someone living in full health by one year: in the UK EQ-5D-3L social value set, the former is worth 1.59 QALYs; the latter only 1 QALY. This means, for treatments that mainly effect HRQoL, the arguments presented in this paper may indeed not apply. However, the advantage SWD give to some people does not justify the disadvantage they give to others. For treatments that affect both, length and health-related quality of life, it may also be very difficult to determine what the overall effect of SWD is. I thus maintain that, if it cannot be ruled out that some person's gain in survival time is valued as a loss to society (or vice versa), SWD shall not be used in health economic evaluations.

I would like to emphasise that assigning a non-negative value to all human lifetime should be considered a *minimal* ethical constraint [41]. There are many other, compelling, more fundamental critiques of the QALY metric and its ethical implications. Some have argued, for example, that all human life should have a positive (and just a non-negative) value [42], or that all human life should be of equal value [9] (see below). Admittedly, these proposals are only concerned with methodological details, while the QALY appears to be accepted as a valid point of reference. Yet, the utilitarian QALY framework itself is not value-free, and could also be called into question [4, 7, 43–45]. However, the argument presented in this paper is deliberately presented within a narrowly defined QALY framework. Even if one accepts the QALY framework in general, I would argue that one has to reject the concept of SWD as unethical.

Moving forward

While I argue for abolishing the use of SWD in health economic evaluations, I do not intend to prescribe a particular approach on how to replace them. Within the QALY framework, there are primarily two options that should be considered.

Firstly, the QALY metric itself could be adjusted, to ensure that every person's lifetime has some positive, or at least non-negative, value. The Equal Value of Life ('EVL') approach, proposed by Nord et al. [9], could be used for this, or the Health Years in Total ('HYT') framework, proposed by Basu et al. [42]. The former assigns every additional life year a value of one QALY, while the latter also takes into account HRQoL changes that occur during additional life years. However, both approaches add something extra to the QALY, which is not derived from the social value set, but imposed rather post-hoc by the researcher or decision maker.

The second alternative may thus seem more attractive: preferences could either be elicited from patients/people living in the health states themselves, or a different perspective could be used when eliciting preferences from the general public. The person trade-off method may have some appeal in this context, as it seems to come closest to the type of decision that social value sets actually inform [26, 37]. Both approaches are likely to generate much higher and probably exclusively positive health state values [17, 18, 30].

The question, which approach is most appropriate, cannot be answered in isolation, but must be guided by a normative theory of the valuation of health. Any alternative approach may also come with a number of wider, potentially unintended implications, which need to be considered. In the current absence of a widely accepted, coherent theoretical framework, more conceptual research seems to be needed. In particular, this should include two different strands: firstly, there should be more engagement with fundamental questions about the ethical underpinning of the QALY framework; and, secondly, health economists should enter into a meaningful and sustained dialog with citizens, policy makers, and other stakeholders, to ensure that their methods reflect the norms and values of society. However, it is unlikely that all considerations a society considers to be relevant can ever be operationalised and integrated into a coherent, formal decision analytical framework. It therefore seems essential that the results of any health economic model are checked and qualitatively scrutinised. Health policy decision makers should critically assess the underlying assumptions and their ethical implications. Greater involvement of patients, patient representatives, and carers may help to ensure that their perspectives are accounted for in the decision-making process.

Acknowledgements I am grateful to Elly Stolk, Aki Tsuchiya, Donna Rowen, Simon McNamara, and the participants of the 2020 lolaHESG online conference for providing comments on an earlier version of this manuscript. I would also like to thank Robert Sugden for his encouragement to continue developing the ideas expressed in this paper. The usual disclaimer applies

Funding This work was supported by the Wellcome Trust DTC in Public Health Economics and Decision Science (108903/Z/19/Z) and the University of Sheffield.

Declarations

Conflict of interest PS has received funding from the EuroQol Group.

Ethical approval This article does not contain any studies with human participants.

References

1. MacKillop, E., & Sheard, S. (2018). Quantifying life: understanding the history of quality-adjusted life-years (QALYs). *Social Science & Medicine, 211*, 359–366.
2. Lipscomb, J., Drummond, M., Fryback, D., Gold, M., & Revicki, D. (2009). Retaining, and enhancing, the QALY. *Value in Health, 12*, S18–S26.
3. Harris, J. (1987). QALYfying the value of life. *Journal of Medical Ethics, 13*, 117–123.
4. Pearson, S. D. (2019). Why the coming debate over the QALY and disability will be different. *The Journal of Law, Medicine& Ethics, 47*, 304–307.
5. Singer, P., McKie, J., Kuhse, H., & Richardson, J. (1995). Double jeopardy and the use of QALYs in health care allocation. *Journal of Medical Ethics, 21*, 144–150.
6. Ubel, P., Nord, E., Prades, J., & Richardson, J. (2000). Improving value measurement in cost-effectiveness analysis. *Medical Care, 1*, 982–901.
7. Beckstead, N., & Ord, T. (2015). Bubbles under the wallpaper: Healthcare rationing and discrimination. In *Bioethics: An anthology* (pp. 406-412). Oxford: Blackwell.
8. Cubbon, J. (1991). The principle of QALY maximisation as the basis for allocating health care resources. *Journal of Medical Ethics, 17*, 181–184.
9. Nord, E., Pinto, J. L., Richardson, J., Menzel, P., & Ubel, P. (1999). Incorporating societal concerns for fairness in numerical valuations of health programmes. *Health Economics, 8*, 25–39.
10. Williams, A. (1987). Brief response: QALYfying the value of life. *Journal of Medical Ethics, 13*, 123.
11. Klonschinski, A. (2016). *The economics of resource allocation in health care: Cost-utility, social value, and fairness*. Milton Park: Routledge.
12. Whitehead, S. J., & Ali, S. (2010). Health outcomes in economic evaluation: The QALY and utilities. *British Medical Bulletin, 96*, 5–21.
13. Brazier, J., Ratcliffe, J., Saloman, J., & Tsuchiya, A. (2017). *Measuring and valuing health benefits for economic evaluation*. Oxford: Oxford University Press.
14. NICE. (2019). Position statement on use of the EQ-5D-5L value set for England (updated October 2019). Accessed September 9, 2020, from https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/technology-appraisal-guidance/eq-5d-5l.
15. Attema, A. E., Edelaar-Peeters, Y., Versteegh, M. M., & Stolk, E. A. (2013). Time trade-off: one methodology, different methods. *The European Journal of Health Economics, 14*, 53–64.
16. Torrance, G. W. (1976). Social preferences for health states: An empirical evaluation of three measurement techniques. *Socio-Economic Planning Sciences, 10*, 129–136.
17. Brazier, J., Rowen D., Karimi, M., Peasgood, T., Tsuchiya, A., & Ratcliffe, J (2018). Experience-based utility and own health state valuation for a health state classification system: why and how to do it. *The European Journal of Health Economics, 19*, 881–891.
18. Versteegh, M., & Brouwer, W. (2016). Patient and general public preferences for health states: A call to reconsider current guidelines. *Social Science & Medicine, 165*, 66–74.
19. De Charro, F., Busschbach, J., Essink-Bot, M.-L., van Hout, B., & Krabbe, P. (2005). *EQ-5D concepts and methods: A developmental history* (pp. 171–179). Dordrecht: Springer.
20. Tilling, C., Devlin, N., Tsuchiya, A., & Buckingham, K. (2010). Protocols for time tradeoff valuations of health states worse than dead: A literature review. *Medical Decision Making, 30*, 610–619.
21. McCabe, C., Claxton, K., & Culyer, A. J. (2008). The NICE cost-effectiveness threshold. *Pharmacoeconomics, 26*, 733–744.
22. Bernfort, L., Gerdle, B., Husberg, M., & Levin, L. -Å. (2018). People in states worse than dead according to the EQ-5D UK value set: Would they rather be dead? *Quality of Life Research, 27*, 1827–1833.
23. Chang, H. F. (2000). A liberal theory of social welfare: fairness, utility, and the Pareto principle. *The Yale Law Journal, 110*, 173–235.
24. Farsides, B., & Dunlop, R. J. (2001). Is there such a thing as a life not worth living? *BMJ, 322*, 1481–1483.
25. Harris, J. (2003). Consent and end of life decisions. *Journal of Medical Ethics, 29*, 10–15.
26. Weinstein, M. C., Torrance, G., & McGuire, A. (2009). QALYs: The basics. *Value in Health, 12*, S5–S9.
27. Kennedy-Martin, M., Slaap, B., Herdman, M., van Reenen, M., Kennedy-Martin, T., Greiner, W., Busschbach, J., & Boye, K. S. (2020). Which multi-attribute utility instruments are recommended for use in cost-utility analysis? A review of national health technology assessment (HTA) guidelines. *The European Journal of Health Economics, 21*, 1245–1257.
28. Kreimeier, S., Oppe, M., Ramos-Goñi, J. M., Cole, A., Devlin, N., Herdman, M., Mulhern, B., Shah, K. K., Stolk, E., Rivero-Arias, O., & Greiner, W. (2018). Valuation of EuroQol five-dimensional questionnaire, youth version (EQ-5D-Y) and EuroQol five-dimensional questionnaire, three-level version (EQ-5D-3L) health states:

The impact of wording and perspective. *Value in Health, 21*, 1291–1298.

29. Helgesson, G., Ernstsson, O., Åström, M., & Burström, K. (2020). Whom should we ask? A systematic literature review of the arguments regarding the most accurate source of information for valuation of health states. *Quality of Life Research, 29*, 1465–1482.
30. Burström, K., Sun, S., Gerdtham, U.-G., Henriksson, M., Johannesson, M., Levin, L.-Å., & Zethraeus, N. (2014). Swedish experience-based value sets for EQ-5D health states. *Quality of Life Research, 23*, 431–442.
31. Dolan, P. (1997). Modeling valuations for EuroQol health states. *Medical Care, 35*(11), 1095–1108.
32. MVH Group. (1995). *The measurement and valuation of health: Final report on the modelling of valuation tariffs*. York: Centre for Health Economics, University of York.
33. Jelsma, J., Hansen, K., De Weerdt, W., De Cock, P., & Kind, P. (2003). How do Zimbabweans value health states? *Population Health Metrics, 1*, 1–10.
34. Luo, N., Wang, P., Thumboo, J., Lim, Y.-W., & Vrijhoef, H. J. (2014). Valuation of EQ-5D-3L health states in Singapore: Modeling of time trade-off values for 80 empirically observed health states. *Pharmacoeconomics, 32*, 495–507.
35. University College London Department of Epidemiology and Public Health; National Centre for Social Research (NatCen). (2021). Health Survey for England, 2017. UK Data Service.
36. Menzel, P., Dolan, P., Richardson, J., & Olsen, J. A. (2002). The role of adaptation to disability and disease in health state valuation: A preliminary normative analysis. *Social Science & Medicine, 55*, 2149–2158.
37. Nord, E. (1995). The person-trade-off approach to valuing health care programs. *Medical Decision Making, 15*, 201–208.
38. Rubin, E. B., Buehler, A. E., & Halpern, S. D. (2016). States worse than death among hospitalized patients with serious illnesses. *JAMA Internal Medicine, 176*, 1557–1559.
39. Tappenden, P., Hamilton, J., Kaltenthaler, E., Hock, E., Rawdin, A., Mukuria, C., Clowes, M., Simonds, A., & Childs, A. (2018). *Nusinersen for treating spinal muscular atrophy: A single technology appraisal*. Sheffield: School of Health and Related Research (ScHARR).
40. Scott, C. E. H., MacDonald, D., & Howie, C. (2019). 'Worse than death' and waiting for a joint arthroplasty. *The Bone & Joint Journal, 101*, 941–950.
41. Franklin, D. (2017). Calibrating QALYs to respect equality of persons. *Utilitas, 29*, 65.
42. Basu, A., Carlson, J., & Veenstra, D. (2020). Health years in total: a new health objective function for cost-effectiveness analysis. *Value in Health, 23*, 96–103.
43. Anand, P. (1999). QALYs and the integration of claims in healthcare rationing. *Health Care Analysis, 7*, 239–253.
44. Long, S. (2015). Squashed dreams and rare breeds: Ableism and the arbiters of life and death. *Disability & Society, 30*, 1118–1122.
45. Broome, J. (1978). Trying to value a life. *Journal of Public Economics, 9*, 91–100.

Conclusion to Chapter 5

This chapters taps into a long-standing debate about whether or not the QALY is discriminatory, and it adds to the discussion by presenting a hitherto overlooked problem. It demonstrates that when states worse than dead are used to value changes in survival times, they unequivocally imply unethical value judgements and discriminate against those people in poor health states. I argue that extending a person's life should generate at least zero QALYs, and that a meaningful and sustained dialog with citizens, policy makers, and other stakeholders is needed to ensure that health economic models reflect the norms and values of society.

Part II

Practical Tools

Alice laughed. `There's no use trying,' she said: `one CAN'T believe impossible things.'

'I daresay you haven't had much practice,' said the Queen. 'When I was younger, I always did it for half an hour a day. Why, sometimes I've believed as many as six impossible things before breakfast.'

– Alice in Wonderland, Lewis Carroll

Chapter 6

The Online Elicitation of Personal Utility Functions (OPUF) tool

This chapter describes the Online Elicitation of Personal Utility Functions (OPUF) approach, a new method for valuing EQ-5D-5L health states. The aims of this study are to report on the development of the survey tool, and to test its feasibility in a small pilot study with 50 participants from the UK. The results show that OPUF can be used to obtain not just group-level, but also individual-level value sets.

This chapter was published in an identical form as: Schneider PP, van Hout B, Heisen M, Brazier J, Devlin N. The Online Elicitation of Personal Utility Functions (OPUF) tool: a new method for valuing health states. Wellcome Open Research. 2022 Jan 14;7:14. https://doi.org/10.12688/wellcomeopenres.17518.1

NB: Wellcome Open Research has an open peer review process. The reviewers' comments are not included here, but can be accessed under the url given above.

The paper in the chapter was written with four co-authors, Ben van Hour, Marieke Heisen, John Brazier and Nancy Devlin. PS, BvH, and MH conceptualised the study. PS developed and implemented the survey software, conducted the analysis, and wrote the first draft of the manuscript. BvH, MH, JB, and ND provided critical feedback to the survey and the analysis. BvH and JB supervised the project. All authors reviewed, edited, and approved the final version.

Check for updates

METHOD ARTICLE

The Online Elicitation of Personal Utility Functions (OPUF) tool: a new method for valuing health states [version 1; peer review: 2 approved, 1 approved with reservations]

Paul P. Schneider[1], Ben van Hout[1,2], Marike Heisen[3], John Brazier[1], Nancy Devlin[4]

[1]School of Health and Related Research, University of Sheffield, Sheffield, UK
[2]OPEN Health, York, UK
[3]Open Health, Rotterdam, The Netherlands
[4]School of Population and Global Health, University of Melbourne, Melbourne, Australia

v1 **First published:** 14 Jan 2022, **7**:14
https://doi.org/10.12688/wellcomeopenres.17518.1
Latest published: 14 Jan 2022, **7**:14
https://doi.org/10.12688/wellcomeopenres.17518.1

Abstract

Introduction

Standard valuation methods, such as TTO and DCE are inefficient. They require data from hundreds if not thousands of participants to generate value sets. Here, we present the Online elicitation of Personal Utility Functions (OPUF) tool; a new type of online survey for valuing EQ-5D-5L health states using more efficient, compositional elicitation methods, which even allow estimating value sets on the individual level. The aims of this study are to report on the development of the tool, and to test the feasibility of using it to obtain individual-level value sets for the EQ-5D-5L.

Methods

We applied an iterative design approach to adapt the PUF method, previously developed by Devlin et al., for use as a standalone online tool. Five rounds of qualitative interviews, and one quantitative pre-pilot were conducted to get feedback on the different tasks. After each round, the tool was refined and re-evaluated. The final version was piloted in a sample of 50 participants from the UK. A demo of the EQ-5D-5L OPUF survey is available at: https://eq5d5l.me

Results

On average, it took participants about seven minutes to complete the OPUF Tool. Based on the responses, we were able to construct a personal EQ-5D-5L value set for each of the 50 participants. These value sets predicted a participants' choices in a discrete choice experiment with an accuracy of 80%. Overall, the results revealed that health state preferences vary considerably on the individual-level.

Open Peer Review

Approval Status ✓ ? ✓

	1	2	3
version 1 14 Jan 2022	✓ view	? view	✓ view

1. **Katie Spencer**, University of Leeds, Leeds, UK
   Leeds Teaching Hospitals NHS Trust, Leeds, UK
2. **Ciaran O'Neill**, Queen's University Belfast, Belfast, UK
3. **Stefan A. Lipman**, Erasmus University Rotterdam, Rotterdam, The Netherlands

Any reports and responses or comments on the article can be found at the end of the article.

Nevertheless, we were able to estimate a group-level value set for all 50 participants with reasonable precision.

Discussion

We successfully piloted the OPUF Tool and showed that it can be used to derive a group-level as well as personal value sets for the EQ-5D-5L. Although the development of the online tool is still in an early stage, there are multiple potential avenues for further research.

Keywords

EQ-5D, Health valuation, multi-attribute value theory, multi-criteria decision analysis, online survey, personal utility function, preference elicitation, stated preferences

Corresponding author: Paul P. Schneider (p.schneider@sheffield.ac.uk)

Author roles: Schneider PP: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **van Hout B**: Conceptualization, Methodology, Project Administration, Supervision, Writing – Review & Editing; **Heisen M**: Conceptualization, Investigation, Methodology, Project Administration, Supervision, Writing – Review & Editing; **Brazier J**: Conceptualization, Methodology, Supervision, Writing – Review & Editing; **Devlin N**: Conceptualization, Methodology, Supervision, Validation, Writing – Review & Editing

Competing interests: BvH, JB, and ND are members of, and all authors have received research funding from the EuroQol Group.

Grant information: This work was supported by Wellcome [108903, https://doi.org/10.35802/108903].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Schneider PP, van Hout B, Heisen M *et al.* **The Online Elicitation of Personal Utility Functions (OPUF) tool: a new method for valuing health states [version 1; peer review: 2 approved, 1 approved with reservations]** Wellcome Open Research 2022, **7**:14 https://doi.org/10.12688/wellcomeopenres.17518.1

First published: 14 Jan 2022, **7**:14 https://doi.org/10.12688/wellcomeopenres.17518.1

1 Introduction

The valuation of health, in terms of quality-adjusted life years (QALYs), is an essential component in health economic evaluations. The QALY is generally derived from generic measures of health, which, in turn, consist of two components: firstly, a health descriptive system, which defines a number of mutually exclusive health states and, secondly, a set of (social) values, that reflect their respective desirability. These values are commonly based on individual preferences of members of the general public[1,2].

Methods for eliciting preferences belong to one of two types: they are either compositional or decompositional[3–5]. Standard health state valuation methods, such as time trade-off (TTO), standard gamble (SG), discrete choice experiments (DCE) and best-worst scaling (BWS) belong to the latter group. Their main disadvantage is that they are inefficient. The amount of information that is obtained from each participant is so small, that data from hundreds, if not thousands, of participants is required in order to estimate a social value set. Generating value sets for small subgroups will thus often not be feasible at all[6,7].

Compositional methods, on the other hand, are much more efficient – they even allow the estimation of value sets on the individual-level. Values can also directly be aggregated across individuals, without the need for complicated statistical models. Nevertheless, compositional methods have seldom been used in the valuation of health and, where they have been used, it is generally in combination with decompositional methods[8].

Recently, Devlin *et al.*[9] pioneered a new method for eliciting health state values, based entirely on compositional preference elicitation techniques. Their personal utility function (PUF) approach was successfully piloted in face-to-face interviews to derive personal (as well as a social) value sets for the EQ-5D-3L instrument[10]. The EQ-5D-3L is a generic measure of self-reported health, which is widely used in health economic evaluations (see below).

In this paper, we aim to expand on the previous PUF work in three ways. Firstly, we establish its theoretical foundations, namely multi-attribute value theory, and how it relates to the valuation of health states more generally (section 2). Secondly, we report on the development of a new, PUF-based online tool (OPUF) to obtain individual-level value sets for the EQ-5D-5L (section 3), and then pilot the tool in a small sample of participants (section 4). Finally, we discuss the main advantages, disadvantages, and potential challenges, and propose potential next steps in the development of the OPUF approach (section 5).

2 Theoretical framework

Preference-based measures of health are (implicitly or explicitly) built on multi-attribute value or utility theory (MAVT/MAUT). These frameworks provide the theoretical foundations for the application of compositional and decompositional preference elicitation methods[11–13]. Before we provide a brief introduction into MAVT/MAUT, it may useful, however, to highlight some relevant aspects of health descriptive systems, to demonstrate how closely they are linked to MAVT/MAUT.

2.1 Health descriptive systems

Most health descriptive systems, generic or condition-specific, share a similar structure, in the sense that health states are defined along a set of dimensions (e.g. pain, mobility, etc), of which each has a number of attributes, reflecting different levels of performance[1,14]. These levels usually have an inherent order, such that higher levels are preferred over lower level, or vice versa (e.g. some pain is better than severe pain). All possible combinations of attributes from different dimensions define the complete set of health states that a descriptive system can represent. Moreover, in most systems there is one best state, full health, which dominates all other states, and one worst state, which is dominated by all other states. For use in health economic evaluations, health descriptive systems need to be valued: utility values, anchored at full health (=1) and dead (=0), need to be assigned to all health states. These values are sometimes also referred to as social values, preference-based indices and health utilities , (health-related) quality of life-, or QALY-weights (we use these terms synonymously). As we will explain below, the structure of a health descriptive system is crucial for its valuation.

2.2 The EQ-5D-5L instrument

To give an example, and also to describe the instrument that is to be valued in this study using the OPUF, we briefly introduce the EQ-5D-5L[15]. This health descriptive system defines health states using five dimensions/criteria: mobility (MO), self-care (SC), usual activities (UA), pain or discomfort (PD), and anxiety or depression (AD). Each dimension has five performance levels: no, slight, moderate, severe, and extreme problems. However, the extreme level for dimensions MO, SC; and UA use the word 'unable' (e.g. unable to walk about). In total, the instrument describes 3,125 mutually exclusive health states. They can be referred to by a 5-digit code, representing the severity levels for the five dimensions. '11111' denotes full health; and '55555' denotes the objectively worst health state.

2.3 Multi-attribute value and utility theory

MAVT and MAUT are general (multi-criteria decision making) frameworks to analyse decision problems involving multiple alternatives and conflicting objectives. The difference between MAVT and MAUT is that the former deals with problems under certainty, while the latter also incorporates uncertainty. The general concept, however, is the same: the stated preferences of an individual, or a group of individuals, over a number of alternatives can be quantified as a value (or utility) function, which assigns a score to any alternative under consideration. The alternatives only have value in so far as they meet certain objectives. This makes it possible to learn a decision maker's partial preferences for these objectives, construct a preference function, and then use it to predict values for different alternatives[3,5].

The valuation of health states can be described with this framework[13]. The three general structural levels (alternatives, objectives, performances) can be mapped directly to corresponding concepts in health descriptive systems. Firstly, the alternatives under consideration, which are to be valued, correspond to health states. Secondly, the objectives against which alternatives are to be evaluated correspond to the different health dimensions (e.g. pain, mobility). Thirdly, the alternatives' performance levels, i.e. the extent to which the alternatives meet the objectives, correspond to the dimension levels of the different health states (e.g. some pain, impaired mobility, etc).

2.4 Value measurement theory

In the context of the QALY framework, constructing a value function for health states requires three components:

1. **Level ratings/scores**: also referred to as marginal value functions, reflect the preferences for different levels of performance on a given criterion. This specifies, for example, how much better *some* pain is compared to *severe* pain. The scale is defined by the best and worst possible level of performance. The units of measurement are arbitrary, but for convenience, values are usually normalised between 100 (best) and 0 (worst).

2. **Criteria/dimension weights**: they represent the relative importance of a given criterion, compared to all other criteria. More specifically, it is a measure of the relative (utility) gain associated with replacing the lowest level with the highest level of performance for this criterion (e.g. moving from extreme pain to no pain). A value of 100 is assigned to the most important criterion, and the weights of all other criteria are then defined relative to this yardstick: a value of 50, for example, means a criterion is half as important; a value of zero means a criterion is not important at all.

3. **Anchoring factor**: anchoring is an additional step, only required in the context of the QALY framework. It is necessary, because health state utilities need to be mapped on to a scale, which is anchored at full health, set to 1, and dead, set to 0. For this, an additional parameter needs to be elicited, that we will call anchoring factor[16]. It was operationalised as a person's maximum range of utility values, i.e. the difference between their highest and their lowest utility value. Alternatively, it can be understood as a person's (assumed) rate of substitution between units of quantity and units of quality of life.

All three components are combined into a (global) value function, using some pre-specified aggregation method. Most commonly, an additive aggregation function (weighted sum) is chosen. It is easy to interpret, as it only considers marginal changes. Since we want to anchor utility values on the QALY scale, we first need to normalise the additive function between 1 and 0 (i.e. divide both components by 100), and then rescale the function, using the anchoring factor *a*. Accordingly, an additive model with *m* criteria can be written as:

$$V(h) = 1 - a * \sum_{i=1}^{m} 1 - \frac{w_i p(h_i)}{1000}$$

whereby $V(h)$ is the value function which assigns a utility value to any health state h; a is the anchoring factor (=utility range); w_i is the weight of the ith dimension, h_i is the level of performance of state h on criterion i, and $p(h_i)$ then gives the marginal value of state h's performance level on dimension i. It should be noted that the anchoring factor is usually not explicitly considered as a separate criterion in the value function. Instead, it is used to rescale the dimension weights and level ratings (see section 'How to construct PUF's from participants' responses' below).

2.5 Decompositional and Compositional methods

As stated in the introduction, there are two types of preference elicitation methods: compositional and decompositional methods. We assume that readers will be familiar with decompositional methods, in the form of TTO, SG, DCE, or BWS. All of these methods require participants to evaluate entire health states. This means, they need to consider all the relevant criteria at the same time, and then assign cardinal values to these states. Subsequently, these values are decomposed, with the aim to work out the marginal contribution of each attribute to the overall utility score. Ultimately, this procedure provides a scoring system, with coefficients for the different dimensions and levels, which can be used to estimate the values for all health states.

Another aspect that should be noted is that, in practice, it is usually infeasible to elicit values for all health states from one individual. Therefore, a statistical model needs to be fitted to the values elicited from multiple individuals over a subset of the states[17,18]. Depending on the complexity of the health descriptive system, large numbers of participants may need to be surveyed to yield sufficient data points for the statistical model to converge and to produce robust estimations[6,7]. This makes it generally impossible to construct value functions for small groups or for single individuals.

The elicitation of preferences through compositional methods works the other way around. They start with the valuation of the individual components of health states: criteria weights, level ratings and the anchoring factor are elicited directly and in separate tasks. The three components are then combined, using a pre-specified aggregation function, to estimate the values for all health states.

There are several compositional preference elicitation techniques that can be used[4]. The most straight-forward methods involve asking participants to allocate points or rate the attributes directly, using a visual analogue scale (VAS), for example. Alternative methods include ranking techniques, Likert-type scales (AHP) or semantic categories (MACBETH)[19–21].

These techniques have been used extensively in multicriteria decision analysis (MCDA), including numerous applications in the context of health technology assessments[22–24]. Up until now, however, the application of compositional methods in health valuation studies has been scarce. One notable exception is the Health Utility Index (HUI 2, HUI 3)[8,25]. Based on a MAUT framework, value sets were derived by combining the (decompositional) SG method with a (compositional) visual analogue scale. Criteria weights and the anchoring factor were (simultaneously) derived through the former, while the latter provided the levels scores. However, the PUF approach appears to be the first that is entirely based on compositional preference elicitation techniques[9].

3 Development of the OPUF Tool

3.1 From PUF to OPUF

The PUF approach was developed by Devlin *et al.*[9] as a new method to derive personal value sets for the EQ-5D-3L[10]. It consists of a series tasks, organised in seven sections (A: warm-up, B: dimension ranking, C: dimension rating, D: level rating, E: paired comparison, F: position-of-dead, G: check for interactions). The approach was successfully piloted in 76 face-to-face interviews. The results showed that compositional methods can be used to derive EQ-5D-3L value set on the group, as well as on the individual level.

In recent years, the use of online data collection of stated preferences data has become more and more popular. The main reasons for this are presumably the speed and the often markedly reduced costs compared to interviewer administration. This may, in part, also explain the rise in the use of DCE, which, compared to TTO, are much easier to apply online[26,27].

The aim of the present study was to adapt and refine the PUF approach for use as a stand-alone online survey, and to test its use in valuing the EQ-5D-5L. With one exception (G: check for interactions) all tasks used in the original approach were implemented in the OPUF. We only added one additional task, the 'Dead-VAS', to be able to anchor the PUF of participants with a certain preference profile (see below). Nevertheless, the overall implementation of the OPUF differed significantly from the original. The original PUF approach was delivered in face-to-face interviews. Participants were encouraged to reflect on, explain, and revise their responses. Deliberation and the interaction with the interviewer were key components of the study, and interviews took up to 90 minutes. We believe this approach cannot easily be replicated in a stand-alone online tool. Participants may be less motivated to work through difficult exercises or to reflect on their preferences, without the presence of a human interviewer. We therefore decided to make the survey shorter, and focused on clear and intuitive presentation of the tasks. For this, we simplified some of the instructions and tried to design an easy-to-use web interface.

3.2 Development of the EQ-5D-5L OPUF Tool

The OPUF Tool was programmed in R Shiny – an extension of the R programming language for creating interactive user interface[28]. For the development, we used an iterative design approach. First, we experimented with various approaches for emulating the PUF tasks, that were applied in face-to-face interviews conducted online survey. This involved exploring the capabilities of R Shiny, and testing different input elements, such as numeric or text input fields, buttons, drop-down menus, and sliders. Since default templates did not always seem adequate, we developed several new input elements, including visual analogue scales (VAS), a level rating scale, and a colour-coded DCE. Different presentations of the tasks were discussed among the research team and tested with colleagues. Three different versions of the online tool were built before we developed a first fully functional prototype.

Subsequently, the prototype was evaluated and further refined in five iterative rounds of user testing. This involved qualitative online interviews with a total of 22 participants (5+4+4+5+4), recruited via the Prolific platform (https://www.prolific.co). During the interviews, we observed the participants' screens while they were going through the OPUF Tool. After each task, we asked them how they understood the task, how difficult it was, and whether there was anything confusing about it. The interviews took between 15 and 53 minutes. After each round, we revised the tool based on the feedback we received. After the third round, we also conducted a first 'test launch', for which we recruited 50 participants to complete the tool without being directly monitored by the interviewer. Data from the test launch was used to check and refine our analysis plan.

Once we arrived at the final version of the OPUF Tool, we conducted a quantitative pilot to test the feasibility of using it for deriving personal as well as group-level EQ-5D-5L utility functions. The results are described in section 4 (quantitative pilot results).

3.3 The EQ-5D-5L OPUF Tool

The OPUF Tool consists of 10 steps. In the following, we describe each step in more detail and explain how the respective tasks work. However, we consider the visual presentation of the tasks an essential component of the OPUF Tool. Much effort went into developing an intuitive and easy-to-use design. We thus recommend readers to consult the online demo version of the tool while reading through this section. It is available at https://eq5d5l.me.

Steps 1 & 2: Warm-up

The first two tasks aim to familiarise participants with the instrument and the five dimensions it covers. They are asked to self-report their current health on the EQ-5D-5L descriptive system and to rate their overall health status, using the EQ-VAS. To avoid any anchoring effect, we designed a new, empty slider input element, which had no default value.

Step 3: Level rating

In the original PUF, level rating involved five separate tasks, one for each dimension of the EQ-5D-3L. Participants were asked to allocate 100 points between an improvement from extreme to moderate, and from moderate to no problems. Since no and extreme problems are fixed at 100 and 0, in effect, this exercise determined the values of the 'moderate' level on each dimension. For the OPUF Tool, the move from the 3L to the 5L version meant that we had to reconsider the design. Asking participants, for each dimension, to allocate points to four improvements (extreme to severe, severe to moderate, moderate to slight, and slight to no problems) seemed excessive. We thus considered two alternative options:

> A Use the design for the 3L version to elicit a score for the moderate level on each dimension, and then linearly interpolate the scores for the slight and severe level. This assumes that the differences between levels are equal.
>
> B Elicit scores for all levels without any reference to a particular dimension. This assumes that the different levels of severity ('slight', 'moderate' etc.) have consistent interpretations, irrespective of the specific health problem.

We assessed the model coefficients of existing EQ-5D-5L value sets from different countries, to check whether either of the options could be supported by empirical data. However, the evidence was ambiguous and partly contradictory. Ultimately, we chose to implement option B (elicit all level ratings without reference to a specific dimension) because it seemed more convenient for the participants.

The final instructions for the task state that "a person with 100% health has no", and "a person with 0% health has extreme health problems". Participants are then asked: "[h]ow much health does a person with slight health problems have left?". Responses are recorded on a scale that ranges from 100% (= no problem) to 0% (extreme problems). After the participant clicks on the scale, two things happen. Firstly, the label ('slight problems') and a connecting arrow appear right next to the selected value; and secondly, the question changes to the next severity level (i.e. from slight to moderate, and from moderate to severe). The severity levels are highlighted, using a purple background colour (the hue depends on the severity level).

During the entire pilot phase, this task was considered to be difficult by many of the participants. Especially in earlier versions of the tool, participants were often confused by the instructions and we had to revise and simplify the instructions and layout several times.

In a previous version, the task also included default values, i.e. the values of slight, moderate, and severe problems were preset to 75%, 50% and 25%, respectively, and participants were asked to adjust them. Yet, this caused a strong anchoring effect and many participants did not change those values: 26 of 50 participants (52%) kept the preset value for the moderate severity level, for example. Adapting the design, so that it did not show any defaults, was technically challenging, but seemed necessary in light of these early findings.

Step 4: Dimension ranking

Participants are presented with the worst levels of each dimension (i.e. 'I am unable to walk about, I am unable to wash and dress myself, etc), and asked to rank them in order of which problem they would 'least want to have'; ties were not permitted. The task aims to introduce participants to the idea of prioritising one dimension of health over another. Responses to this task are also used to tailor the presentation of the following task to the individual participant.

Step 5: Dimension weighting (Swing weighting)

Five sliders are shown, one for each dimension, describing an improvement from the worst (extreme problems) to the best level (no problems). The sliders are presented in the same order as the participant had just ranked them. The first slider, for the most important dimension, is set to 100. This is given as a fixed yardstick, that participants are asked to use to evaluate the relative importance of the improvements in the other dimensions (which are set to 0 by default).

The instructions are tailored to each participant: if, for example, extreme pain or discomfort was ranked first in the previous task, the instructions state: "If an improvement from 'I have extreme pain or discomfort' to 'I have no pain or discomfort' is worth 100 'health points', how many points would you give to improvements in other areas?".

Step 6: Validation DCE

Three pairwise comparisons between health states are sequentially presented to the participant: they are asked whether they prefer scenario A or B. The health states for the scenarios are personalised. For each participant, the dimension weights and the level ratings are combined into a (1-0 scaled) PUF. This function is then used to value all 3,125 health states, and to establish a preference order. Ties are broken randomly.

Health states for scenario A are selected from the 25th, 50th, and 75th percentile (order randomised) of the participant's personal ranking. The scenario A states are then paired with states that have an absolute utility distance of about 0.1 (hard choice), 0.2 (medium choice), and 0.3 (easy choice), respectively (order randomised). Dominated and dominating states are excluded.

To make it easier for participants to asses the severity of a health state, we used intensity colour coding, i.e. different shades of purple were used as background colours, ranging from light purple for no problems to dark purple for extreme problems, as previously suggested by Jonker *et al.*[29].

The responses to this task were not used in the construction of the PUF – the purpose was to assess how accurately the OPUF approach can predict an individual participant's actual choices in a standard discrete choice experiment task.

Step 7: Position-of-Dead Task

In this task, participants go through up to six paired comparisons between A) a health state and B) 'Being Dead'. In the first comparison, scenario A is the worst health state ('55555'). If the participant prefers that state over dead, the participant immediately proceeds to Step 8. If they prefer dead, a binary search algorithm is initiated, to find the state that is equal to dead.

As before, in Step 6, the participant's individual PUF is used to value and rank all 3,125 health states. After the participant's indicated that state ’55555’ is worse than being dead, the search goes to the median state. From there, it moves up or down, depending on the participant’s choices, in half-intervals. The search stops after five iterations. At this point, the equal-to-dead state is identified with a maximum error of +/- 49 states, corresponding to 1.6% of the total number of states defined by the EQ-5D-5L.

In a previous version of the tool, the dead state was labelled ‘Immediate Death’. Through the qualitative interviews, however, we learned that this made many participants think about the process of dying and they were consequently rather hesitant to ever choose this option. We changed the label to ‘Being Dead’. We also decided not to display any duration for scenario A, because in the QALY framework, utility independence must be assumed.

Step 8: Dead-VAS

Those participants, who indicated they would prefer the worst health state (’55555’) over being dead, are asked to assess the value of that health state on a vertical visual analogue scale. The top anchor point, at 100, is labelled ’No health problems’, and the bottom, at 0, is labelled ’Being Dead’. The description of the worst health state is shown in a box next to the scale. When the participant selects a position value, an arrow is displayed, connecting the box to the respective position on the scale.

A previous version of the tool did not include the Dead-VAS, but instead all participants completed three TTO tasks: two warm-up tasks and then one TTO involving the worst health state. However, this design often lead to inconsistent responses: 19 of 50 participants (38%) reversed their preference between the Position-of-Dead and the TTO task. More specifically, 15 (30%) switched from worst health state $\prec$ dead to dead $\succ$ worst health state, while 4 (8%) switched the other way around. Although smaller, the latter group was more problematic, because their responses made it impossible to anchor their PUFs, at all.

The inconsistent results could be attributable to several factors. First of all, it is a well known (and unavoidable) fact that different valuation techniques yield different utility values, and thus different anchor points [1, p. 49–76]. Other potential explanations might include differences in the interpretation of the tasks, the additional consideration of time (displayed in the TTO, but not in the Position-of-Dead task), or lack of attention.

To ensure that PUFs can be constructed for all participants, we decided to implement the Dead-VAS. The task also appeared to be easier for the participants and also quicker to complete (the TTO took more than 2 minutes, i.e. 20% of the average completion time, in the pre-pilot).

Step 9: demographics

This step includes questions about personal characteristics that are assumed or have shown to explain some of the variability in people’s health preferences, including age, partnership status, sex, having children, nationality, importance of religion, spirituality or faith, and the frequency of engaging in religious activities, level of education, work status, income, and experience with poor health[10,30].

Add-on: Personal results page

As a thank-you, some of the PUF results are fed back to the participants at the end of the survey. Presented are the dimension ranking and the level rating tasks, as well as estimated utility values for four different health states. Participants could compare their results with aggregate results from the overall sample of participants in each study, and with the value sets for EQ-5D-5L obtained from the English general population using conventional decompositional methods, as reported by Devlin *et al.*[9].

Most participants found it difficult to interpret the results; the meaning of the health state values were unclear. Notwithstanding, many participants appreciated the results page, if only as a gesture, and found it interesting to compare their own results with those from the general population.

Other learnings from the qualitative pilot

The online interviews played a key role in the development of the OPUF Tool. The feedback from participants helped us to identify many minor and major issues, and the tool underwent significant changes over the course of the pilot. The changes affected almost every aspect, including the wording of questions, the presentation of the tasks, the overall layout, and the mechanics of different tasks.

A main challenge in the development process was to strike the right balance between rigour/completeness and ease of use. For example, we started with long descriptions for all tasks, which often included examples, and some also contained animations (e.g. to demonstrate how sliders work). We realised, however, that when descriptions were too long or complicated, participants would skip over them and/or disengage with the tasks. We therefore gradually shortened the descriptions and simplified the language. Overall this seemed to be more effective in conveying the relevant information. The final version only contains very short instructions, and we sought to apply an intuitive design, which eliminates the need for elaborate explanations.

Through the pilot we also learned that from interactions with other websites, most people have developed very clear expectations about interacting with online surveys. When elements (such as buttons, sliders, etc) were presented in a slightly unusual way, it often caused confusion and participants sometimes got stuck on a task. To give just one example, in a previous version, the OPUF Tool included a text box next to a visual analogue scale. The text box would show the value that the participant selected on the scale. At the beginning (when the participant had yet not selected a value), however, the box would be empty. This led several participants to assume that they were expected to enter a value into the box manually. They tried to click on it and to type in a number. Since this did not work, they got frustrated and it took them a while until they realised they had to use the scale instead. This problem was easily resolved by just hiding the box in the beginning, and only showing it after the participant had clicked on the scale and selected a value. In another context, we implemented loading animations, to draw the participants' attention to specific parts of the page when they changed. Otherwise, participants often did not notice that a new task had already started and they were waiting for something to happen. These small 'tricks' very much helped to improve the user experience, which seemed suboptimal, in earlier versions of the OPUF Tool.

The usability of the final version received very positive feedback, and participants described it as "easy to navigate", "clear", or "easy to red and understand". One participant stated that "it felt like everything clicked into place".

4 Quantitative pilot results

We conducted a quantitative pilot study to assess the feasibility of OPUF Tool in practice. As for the qualitative pilot, recruitment was conducted through the Prolific platform without any restrictive inclusion criteria or quota – any adult person from the UK with a prolific account could participate. The main points of interest were the plausibility of the responses, the consistency across tasks, and the participants' engagement with the online tool. We also tested our methods of analysis: the collected preference data was used to construct individual and social value functions, and to value all 3,125 EQ-5D-5L health states. We did not attempt any further exploratory or confirmatory analysis of the data, since this was only a pilot study, without a representative sample.

Sample

Fifty participants were recruited. Of these, 23 (46%) were younger than 30 years of age, 18 (36%) were between 30 and 39, and 9 (18%) were 40 years of age or older. Thirty (60%) participants were female, 20 (40%) were male. A majority of 32 (64%) participants had a high level of education (degree or post-graduate).

Step 1+2: Warm-up

Fourteen (28%) participants reported to be in perfect health. The remaining 36 (72%) participants also mostly reported slight or moderate health problems. Self-reported health on the visual analogue scale ranged from 100 to 40, with a mean (SD) and median (IQR) of 78 (14) and 80 (21.25), respectively.

Step 3: Level ratings

Mean (SD) ratings for the level slight, moderate, and severe were 79.10 (11.45), 54.92 (13.41), and 23.46 (11.27) (the ratings of no and extreme problems were fixed at 100 and 0). Figure 1 shows the full distributions of values assigned to the three levels.

Forty (80%) and 41 (82%) participants set their own values for the slight and severe levels, i.e. they changed the default values. For the moderate level, only 26 (52%) changed the value, which may be an indication for the presence of an anchoring effect.

Step 4: Dimension ranking

Table 1 shows the results of the ranking exercise. Twenty-three (46%) participants considered Pain/Discomfort the most most important criterion. The average ranking of this dimension was 2.2. It was followed by Mobility (mean rank = 2.4), Self-Care (3.0), Anxiety/Depression (3.6), and, lastly, Usual Activities (3.8).

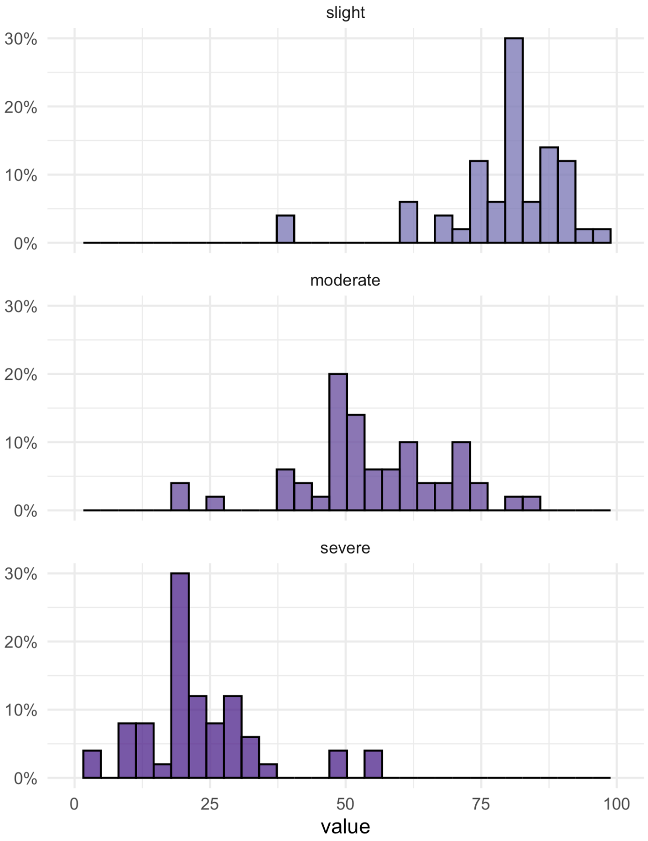

Figure 1. Level ratings for 'slight', 'moderate', and 'severe problems'.

Table 1. Summary of the dimension ranking exercise.

Rank	MO	SC	UA	PD	AD
1st	15 (30%)	8 (16%)	1 (2%)	23 (46%)	3 (6%)
2nd	14 (28%)	11 (22%)	7 (14%)	8 (16%)	10 (20%)
3rd	10 (20%)	14 (28%)	12 (24%)	7 (14%)	7 (14%)
4th	9 (18%)	9 (18%)	10 (20%)	10 (20%)	12 (24%)
5th	2 (4%)	8 (16%)	20 (40%)	2 (4%)	18 (36%)

MO = Mobility; SC = Self-Care; UA = Usual Activities; PD = Pain/Discomfort; AD = Anxiety/Depression

Step 5: Dimension weighting (swing weighting)

Figure 2 shows the distribution of the weights assigned to the five EQ-5D-5L dimensions. The dimension with the highest mean (SD) weight was Mobility at 85.16 (23.51), followed by Pain/Discomfort at 83.08 (26.41), Self-Care at 77.38 (30.22), Usual activities at 69.78 (30.22), and then Anxiety/Depression

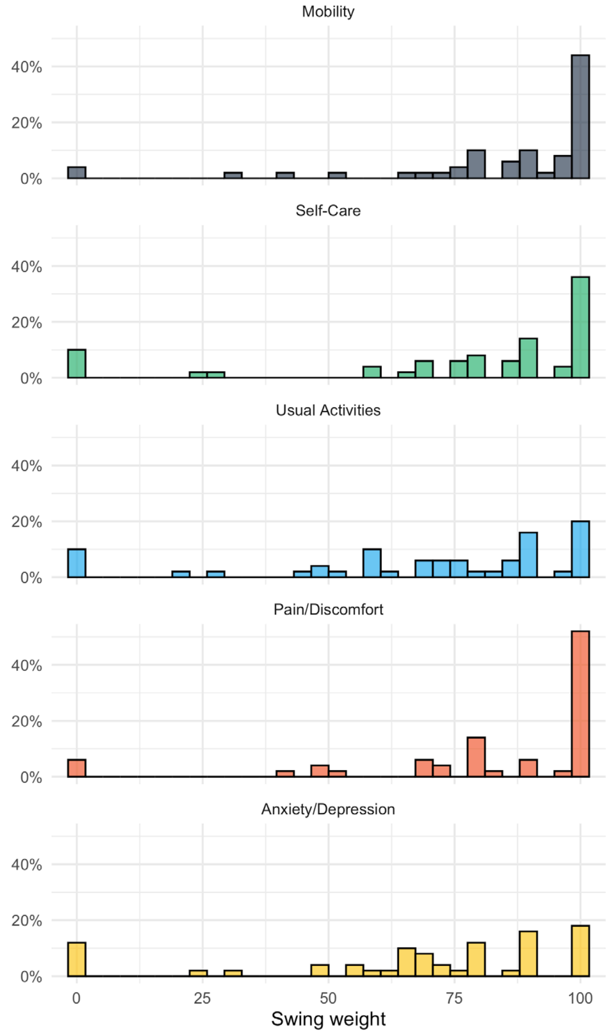

Figure 2. Swing weights for dimension MO = Mobility, SC = Self-care, UA = Usual activities, PD = Pain/discomfort, AD = Anxiety/depression.

at 67.78 (30.78). Four (8%) participants assigned a value of 100 to all dimensions; 7 (14%) assigned a value of zero to one or more dimensions.

The weights of 30 (60%) participants implied different preference order, i.e. at least one preference reversal, compared to the order specified in the previous ranking task (ties were not considered an order violation). As noted above, these inconsistencies do not necessarily signify that participants did not pay attention. In the qualitative pilot, some participants deliberately chose a different ranking, in response to the slightly differently phrased question.

Step 6: Validation DCE

Each participant completed three paired comparisons. Of the 150 choices, 120 (80%) were consistent with the choices predicted by participants' PUFs. More specifically, 28 (56%) participants made no inconsistent choice, 15 (30%) made one, six (12%) participants made two, and one (2%) participants made three 'errors'.

We also found that the larger the utility difference between the two states in a choice set, the smaller the error rate: at a distance of about 0.1 (on a normalised 0-1 scale, dominating/dominated states were excluded), the error rate was 26%, at 0.2, it was 24%, and at 0.3, it was 10%.

Step 7: Position-of-Dead Task

A total of 18 (36%) participants stated that they would prefer the worst health state ('55555') over 'being dead'. Another nine (18%) preferred 'being dead' in the first choice set, but then choose the health state in the next five sets. Of the remaining participants, the position of dead varied greatly. The number of states considered worse than dead ranged from 0 (0%) to 2,883 (92%), with a mean and median of 483 (15%) and 50 (2%).

Step 8: Dead-VAS

The 18 participants, who considered the worst health state better than 'being dead', completed the Dead-VAS task. Their valuations of the worst health state on a scale between 100 ('no health problems') and 0 ('being dead') ranged from 5 to 70, with a mean (SD) and median (IQR) of 23.22 (21.03) and 19.5 (21.75).

Step 9: Demographics

Some of the collected demographic information (age, sex, level of education) are provided above in the description of the study sample. Further data are not reported here, since this is only a pilot study, and we did not attempt to make any inferences about participants personal characteristics.

Survey duration

On average, it took participants about seven minutes (range: 3.6 – 18.2 mins) to complete all tasks. The longest time (76 secs) participants spent completing the survey was on the dimension weighting task and the demographic questions. The shortest duration was observed for the subjective health status (EQ-VAS) (21 seconds). Further details on the time participants spent on different tasks are shown in Table 2. With only very few exceptions (e.g. one participants spent only 4 seconds on the dimension ranking task), the observed times seemed by and large plausible and suggested that participants did engage with the tasks.

How to construct PUFs from participants' responses?

Constructing a participant's PUF required two steps: firstly, level ratings were combined with the dimension weights. Secondly, the resulting model coefficients were anchored on to the QALY scale.

In the first step, level ratings, ranging from 100 (no problems) to 0 (extreme problems) were converted to disutilities, ranging from 0 (no problems) to 1 (extreme problems). For convenience, dimension weights were also normalised so that the sum of all five weights summed up to 1. By taking the outer product of these two vectors, we derived a (1-0 scaled) set model coefficients.

Table 2. Survey completion times (in seconds).

	Mean	SD	Min	25th perc.	Median	75th perc.	Max
Own Health State	29	17	11	18	23	30	96
EQ-VAS	21	18	6	11	15	24	116
Level Rating	58	33	17	36	49	66	177
Dimension Ranking	51	33	4	33	41	58	184
Dimension Weighting	76	47	18	50	62	89	274
Validation DCE	63	27	20	45	57	70	165
Position-of-Dead Task	48	34	7	17	44	64	172
Dead-VAS (conditional)	26	12	15	17	22	32	56
Demographics	76	26	43	62	72	85	195
Total	431	178	215	318	356	508	1091
Total (Minutes)	7.2	3.0	3.6	5.3	5.9	8.5	18.2

In the second step, these coefficients were anchored on the QALY scale, using either the state that was determined to be approximately equal to 'being dead' in the position-of-dead task (for 32 participants who considered one or more health states worse than 'being dead'), or the value that was assigned to the worst health state ('55555') in the Dead-VAS task (for the other 18 participants).

To illustrate the computation with a simple example: suppose an individual rated the five severity levels (denoted l) in the following way: $l_{no} = 100$, $l_{slight} = 90$, $l_{moderate} = 50$, $l_{severe} = 30$, and $l_{extreme} = 0$. Furthermore, they assigned the following weights (denoted w) to the five dimensions: $w_{MO} = 100$, $w_{SC} = 60$, $w_{UA} = 45$, $w_{PD} = 80$, and $w_{AD} = 70$. After converting to level ratings to disutilties and normalising the weights, we get the following two vectors:

$$l = \begin{bmatrix} 0 & 0.1 & 0.5 & 0.7 & 1 \end{bmatrix}; \; w = \begin{bmatrix} 0.29 & 0.17 & 0.11 & 0.23 & 0.2 \end{bmatrix}$$

Taking the outer product provides a (scaled) matrix $\tilde{M}$, containing all 25 level-dimension coefficients (see below). These coefficients can already be used to value (on a 0-1 scale) and rank health states. The value for '12345', for example, is $1 - (0 + 0.02 + 0.06 + 0.16 + 0.20) = 0.56$. It should be noted that this procedure is also used within the OPUF Tool, in order to determine the algorithm for the Position-of-Dead and also to select choice sets for the DCE validation task.

$$l \otimes w = \tilde{M} = \begin{array}{r|ccccc|} & w_{MO} & w_{SC} & w_{UA} & w_{PD} & w_{AD} \\ l_{no} & 0 & 0 & 0 & 0 & 0 \\ l_{slight} & 0.03 & 0.02 & 0.01 & 0.02 & 0.02 \\ l_{moderate} & 0.14 & 0.09 & 0.06 & 0.11 & 0.10 \\ l_{severe} & 0.20 & 0.12 & 0.08 & 0.16 & 0.14 \\ l_{extreme} & 0.29 & 0.17 & 0.11 & 0.23 & 0.20 \end{array}$$

Suppose that for this individual, the health state '51255' was identified as being approximately similar to being dead in the Position-of-Dead task. After we compute the (scaled) disutility for state '51255' (= 0.29 + 0 + 0.02 + 0.23 + 0.2 = 0.74), we can anchor and rescale the coefficient matrix, by simply dividing it by this value:

$$\frac{\tilde{M}}{0.74} = M = \begin{array}{r|ccccc|} & w_{MO} & w_{SC} & w_{UA} & w_{PD} & w_{AD} \\ l_{no} & 0 & 0 & 0 & 0 & 0 \\ l_{slight} & 0.04 & 0.02 & 0.02 & 0.03 & 0.03 \\ l_{moderate} & 0.19 & 0.12 & 0.08 & 0.15 & 0.14 \\ l_{severe} & 0.27 & 0.16 & 0.11 & 0.22 & 0.19 \\ l_{extreme} & 0.39 & 0.23 & 0.15 & 0.31 & 0.27 \end{array}$$

Now, we have derived the individual's PUF. It sets '51255' to 0 (1 − (0.39 + 0 + 0.02 + 0.31 + 0.27) = 0); '11111' is still equal to 1 (1−(0+0+0+0+0) = 1), and the worst health state ('55555') is set to -0.35 (1 − (0.39 + 0.23 + 0.15 + 0.31 + 0.27) = −0.35).

Individual and social PUF

We constructed PUFs for all 50 participants. The descriptive statistics are provided in Table 3. The first column shows the mean coefficients. These mean values may also be taken as the group-level value set (i.e. the group tariff). The 95% confidence intervals were bootstrapped using 10,000 iterations. The width of the confidence intervals suggests that, even with a small sample size of only 50 participants, the OPUF approach allowed us to estimate a group tariff with reasonable precision.

Figure 3 illustrates all 50 personal, as well as the average, group-level utility function for a small subset set of EQ-5D-5L health states. Shown are the values for 50 health states, ranked 1st, 65th, 129th, 192th, 256th, ..., 3125th, according to the group-level utility function.

Table 3. Descriptive statistics for 50 PUFs (i.e. personal model coefficients).

Dim	Lvl	Mean (95% CI)	Min.	25th perc.	Median	75th perc.	Max.
MO	2	0.072 (0.064; 0.099)	0.000	0.031	0.048	0.083	0.573
	3	0.150 (0.138; 0.188)	0.000	0.075	0.126	0.185	0.679
	4	0.250 (0.234; 0.302)	0.000	0.137	0.219	0.309	0.793
	5	0.344 (0.316; 0.437)	0.000	0.175	0.282	0.354	1.554
SC	2	0.057 (0.053; 0.070)	0.000	0.027	0.045	0.076	0.207
	3	0.121 (0.112; 0.151)	0.000	0.068	0.099	0.160	0.622
	4	0.207 (0.192; 0.258)	0.000	0.139	0.176	0.242	1.057
	5	0.282 (0.254; 0.375)	0.000	0.167	0.247	0.309	2.073
UA	2	0.051 (0.047; 0.063)	0.000	0.020	0.040	0.069	0.166
	3	0.103 (0.097; 0.124)	0.000	0.055	0.090	0.144	0.357
	4	0.182 (0.170; 0.221)	0.000	0.102	0.174	0.213	0.629
	5	0.234 (0.219; 0.281)	0.000	0.131	0.219	0.265	0.761
PD	2	0.062 (0.057; 0.078)	0.000	0.030	0.051	0.079	0.281
	3	0.132 (0.123; 0.160)	0.000	0.067	0.114	0.159	0.500
	4	0.225 (0.211; 0.273)	0.000	0.138	0.185	0.269	0.840
	5	0.291 (0.274; 0.351)	0.000	0.173	0.249	0.339	1.000
AD	2	0.052 (0.046; 0.071)	0.000	0.020	0.042	0.066	0.413
	3	0.104 (0.096; 0.130)	0.000	0.045	0.093	0.133	0.489
	4	0.175 (0.163; 0.213)	0.000	0.092	0.154	0.201	0.572
	5	0.231 (0.214; 0.288)	0.000	0.124	0.205	0.259	1.086

MO = Mobility; SC = Self-Care; UA = Usual Activities; PD = Pain/Discomfort; AD = Anxiety/Depression

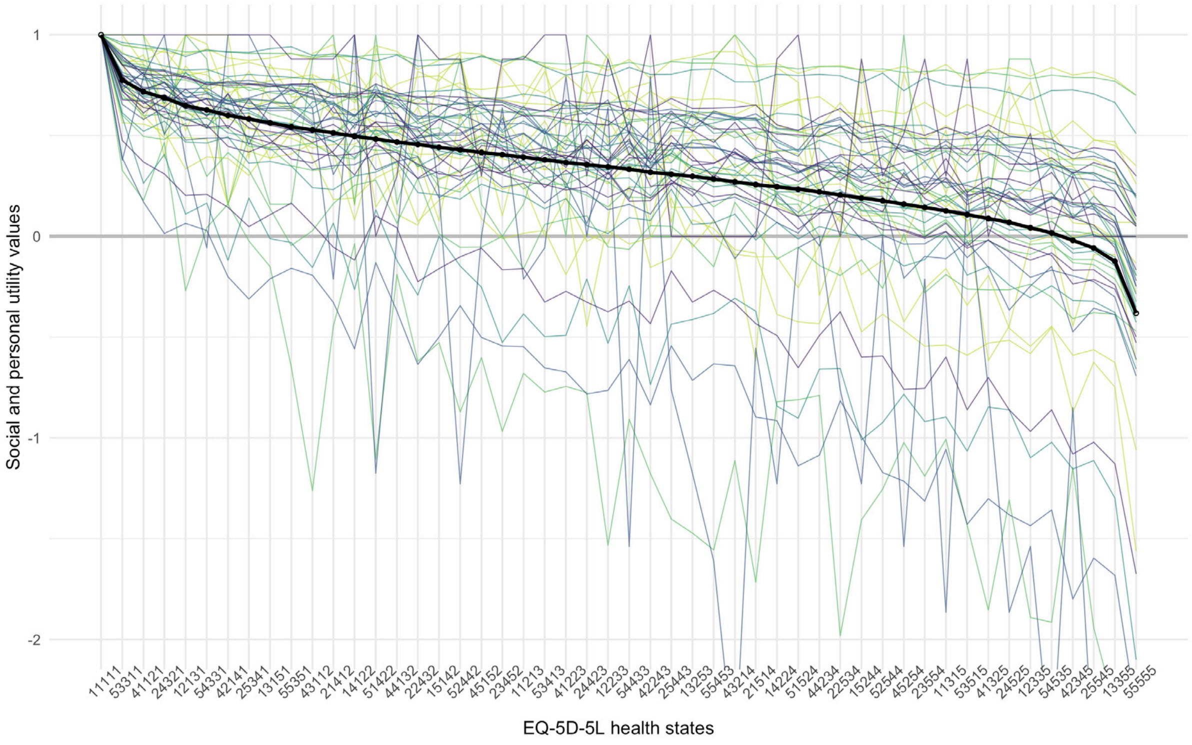

Figure 3. Personal and group-level utility functions for 50 health states, ordered from best to worst, according to the group preference. The thick lines represent the group preference, and the thin lines represent the 50 underlying personal utility functions. The different colours are used to distinguish between separate individuals and have no other meaning.

It can be seen from the graphs that health state preferences of the participants differed considerably. Two separate processes can be distinguished: firstly, lines depicting personal utility values go up and down, and cross each other, while the group preference is monotonically decreasing. This illustrates individual differences in the relative ranking of health states. Secondly, the range of utility values also varies greatly between participants. For some participants, all health states have high values, within a narrow range, while for others, the range of utility values is much wider. Accordingly, the value of the worst health state ('55555') ranges from a maximum of 0.7 to a minimum of -3.2, with a mean and median of -0.4 and -0.2. For comparison, the population estimate reported by Devlin *et al.* is -0.285[18].

It may be interesting to note the difference between the mean and the median, as it shows the effect that outliers, with a wide utility range, have on the overall group tariff. This is not an uncommon finding in valuation studies and for the construction of a social value set, one may want to consider following the common practice of rescaling the negative values to have a lower limit of -1, or using the median, instead of the mean, to aggregate preferences across individuals[31].

5 Discussion

This study provides a comprehensive description of the new OPUF Tool. It covers the theoretical background, reports on the iterative development, and provides a pilot study, which demonstrates that it is feasible to use the online tool for eliciting personal, as well as group-level, preferences for EQ-5D-5L health states.

We think the OPUF Tool provides a flexible, conceptually attractive, and potentially useful new approach for deriving value sets for the EQ-5D-5L (or any other health descriptive system). It could be used as a standalone solution, or to complement established (decompositional) methods, by providing more detailed preference information. The compositional preference elicitation techniques included in the OPUF Tool have several advantages over the more commonly used decompositional methods, which may make the approach particularly attractive to other researchers.

In contrast to the TTO, which is generally administered in face-to-face interviews (though can be online), the OPUF is applied online, which makes it easier and cheaper to collect preference data. The qualitative feedback received during the online interviews even suggests that participants tended to find the online survey to be interesting and engaging. Furthermore, the OPUF approach provides value sets which are anchored on the QALY scale (i.e. at full health and dead), and not only on a latent scale (i.e. un-anchored), which is usually the case in conventional DCE surveys.

Another advantage of the OPUF approach over other conventional valuation methods is the statistical power: fewer participants are required to derive a group tariff or social value set. Note that even with data from just 50 participants, we were able to derive relatively precise estimates for an EQ-5D-5L group tariff. The OPUF Tool may thus allow estimating value sets for smaller groups (e.g. local communities, patient groups), which could practically not be estimated using decompositional methods.

As we have demonstrated, utility functions can even be estimated on the individual-level. This enables researchers to investigate the heterogeneity of health state preferences between individuals in an unprecedented level of detail. It could potentially be useful for other applications beyond health economics (e.g. individualised cost-effectiveness analyses[32]). For example, the OPUF approach could be used as a patient decision aid and to facilitate shared decision making in a clinical context. Explicitly weighing different aspects of health might help patients, who face complex treatment decisions, to better understand the trade-offs that are involved, and what aspects are most important to them.

Furthermore, we would like to draw attention to the fact that the calculations required to construct individual and group-level preferences in the OPUF approach are relatively simple. This makes the underlying model more transparent and potentially easier to communicate to decision makers than more sophisticated statistical models, such as a mixed conditional logit, or a Bayesian hybrid model[18,33].

Finally, another benefit of compositional preference elicitation techniques may be that they break down the valuation of health states into sub-tasks (level rating, dimension weighting, anchoring). The original PUF approach made use of this and encouraged participants to reflect on their preferences at every step of the survey. The OPUF Tool could also be adapted for this purpose and be applied in computer-assisted personal interviews. A study that uses a modified version of the tool to facilitate deliberative discussions among groups of participants is currently under way.

This study also has several important limitations that need to be considered.

Firstly, in the development of the OPUF Tool, 'ease of use' was a main goal. Some valuation tasks were thus simplified, in order to reduce the burden for the participants. For example, we used a single level rating task for all dimensions combined, instead of having separate tasks for each. This assumes the that the relative positions of slight, moderate, and severe problems are the same across all five EQ-5D dimensions. In the absence of any authoritative guidance, it remains unclear whether we struck the right balance between rigour and ease of use.

Secondly, every task has a design which shapes how participants respond to it and which may influence their decision making. This is referred to as choice architecture[34]. Further evaluation of the OPUF Tool could help to assess to what extent participants' responses are sensitive to changes in the presentation of the different tasks, and to improve the quality and robustness of the survey.

Thirdly, an important limitation of compositional preference elicitation techniques is that they cannot easily be used to test for interaction effects. Rather, a functional form must be assumed a priori. In our study, we assumed an additive, main effects model. This seemed reasonable, because it is commonly used to represent health state preferences – most EQ-5D-5L value sets are based on such a model. When studies test for and include interaction effects, authors also often find only minor improvements in the explanatory power[35].

Finally, some important challenges of the OPUF Tool are likely not methodological, but normative. Over the last decades, decompositional preference elicitation methods, have been used extensively in the valuation of health and are by now well established. The compositional methods, used in the OPUF Tool, on the other hand, are new. Decision makers may be less familiar with them, and they may also appear to be conceptually different. This raises the question, *are the derived value sets equally valid*?

Assessing validity of a new method for valuing health is an intricate problem, as there is no gold standard against which it could be compared. At the moment, several valuation methods (SG, TTO, DCE, etc) are used side by side, and numerous studies have shown that these different methods, and even variations of the same method, produce different results[1,36–38]. It is not clear, which, if any, of these methods should be considered to be *the best*. Nevertheless, the findings from this study indicate at least a high level of consistency between the OPUF approach and DCE. We included three standard DCE tasks in the survey and found that the constructed PUF of a particular participant predicted their choices in a DCE task with an accuracy of 80%.

Irrespective of the comparably high level of agreement with DCE, some readers may argue that eliciting preferences requires observing choices involving trade-offs and potentially also risk and uncertainty. Compositional techniques may then seem principally inappropriate. To this, we would reply that MAVT/MAUT provide broad theoretical frameworks, on the basis of which different methods can be justified. Moreover, deviations from formal (Welfare) economic theory are common in health economics and other areas. Simplifications are often made to make certain applications practically feasible. The QALY framework, for example, can be viewed as a major simplification, yet it proved to be immensely useful to inform resource allocation in health care. Similarly, the OPUF Tool may be based on a simpler conception of individual preferences, but it enables new types of analyses (e.g. preferences heterogeneity) and makes it possible to derive value sets on the individual level and in settings in which it would otherwise be unfeasible (e.g. small patient groups).

Next steps

The immediate next step will be to replicate the pilot in a larger study, not only to show that the OPUF can be used to estimate a country-specific social tariff, but also to demonstrate how information on individuals' personal preferences can be harnessed to investigate the heterogeneity of preferences between individuals and/or societal subgroups.

Furthermore, it should be noted that the OPUF approach is not specific to the EQ-5D instrument. The approach is, in principle, applicable to any health descriptive system. This might be true not only on the conceptual level, but also on the technical: the OPUF Tool was programmed in R/Shiny[28]. For the implementation, we developed several generic methods and input elements. This means, the tool could quickly be adapted for different settings (e.g. other country) or instruments (e.g. SF-6D)[39]. Several steps in the development could then be automated. With some further abstraction, the underlying code could potentially provide a flexible, modular software platform for creating valuation tools for any health descriptive system.

Conclusion

Using an iterative design approach, we developed the OPUF Tool; a new type of online survey to derive value sets for the EQ-5D-5L. Based on compositional preference elicitation techniques, it allows the estimation not only of social, but also of personal utility functions. In this study, we successfully tested the OPUF Tool and demonstrated its feasibility in a in a sample of 50 participants from the UK. Even though the development is still in an early stage and further refinement is required, we see several potential applications for the OPUF approach.

Data availability

Underlying data

Zenodo: bitowaqr/opuf$_d$emo: *OPUF zenodo version 1*, *https://doi.org/10.5281/zenodo.5773915*.

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Ethical approval

The study was approved by the Research Ethics Committee of the School of Health and Related Research at the University of Sheffield (ID: 030724). We obtained written informed consent from all participants for the use and publication of their data.

Acknowledgements

We are very grateful to Siobhan Daley, Jack Dowie, Barry Dewitt, Irene Ebyarimpa, Paul Kind, Johanna Kokot, Simon McNamara, Clara Mukuria, Monica Oliveira, Krystallia Pantiri, Donna Rowan, Erik Schokkaert, Koonal Shah, Robert Smith, Praveen Thokala, Ally Tolhurst, Evangelos Zormpas, and the participants of the 2021 Summer HESG virtual meeting for helpful comments, discussions of the ideas expressed in this paper, and/or for providing feedback on earlier versions of the OPUF Tool.We would also like to thank all participants who took part in the pilot study.

References

1. Brazier J, Ratcliffe J, Saloman J, *et al.*: **Measuring and valuing health benefits for economic evaluation.** OXFORD university press, 2017.
Publisher Full Text
2. Whitehead SJ, Ali S: **Health outcomes in economic evaluation: the QALY and utilities.** *Br Med Bull.* 2010; **96**: 5–21.
PubMed Abstract | Publisher Full Text
3. Keeney RL, Raiffa H, Rajala DW: **Decisions with Multiple Objectives: Preferences and Value Trade-Offs.** *IEEE Transactions on Systems, Man, and Cybernetics.* 1979; **9**(7): 403.
Publisher Full Text
4. Marsh K, IJzerman M, Thokala P, *et al.*: **Multiple criteria decision analysis for health care decision making--emerging good practices: report 2 of the ISPOR MCDA Emerging Good Practices Task Force.** *Value Health.* 2016; **19**(2): 125–137.
PubMed Abstract | Publisher Full Text
5. Belton V, Stewart T: **Multiple criteria decision analysis: an integrated approach.** Springer Science & Business Media, 2002.
Publisher Full Text
6. Gandhi M, Xu Y, Luo N, *et al.*: **Sample size determination for EQ-5D-5L value set studies.** *Qual Life Res.* 2017; **26**(12): 3365–3376.
PubMed Abstract | Publisher Full Text
7. de Bekker-Grob EW, Donkers B, Jonker MF, *et al.*: **Sample size requirements for discrete-choice experiments in healthcare: a practical guide.** *Patient.* 2015; **8**(5): 373–384.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Torrance GW, Feeny DH, Furlong WJ, *et al.*: **Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2.** *Med Care.* 1996; **34**(7): 702–722.
PubMed Abstract | Publisher Full Text
9. Devlin NJ, Shah KK, Mulhern BJ, *et al.*: **A new method for valuing health: directly eliciting personal utility functions.** *Eur J Health Econ.* 2019; **20**(2): 257–270.
PubMed Abstract | Publisher Full Text | Free Full Text
10. MVH Group: **The measurement and valuation of health: Final report on the modelling of valuation tariffs.** Centre for Health Economics, University of York, 1995.
Reference Source
11. Richardson JRJ, Mckie JR, Bariola EJ: **Multiattribute Utility Instruments and Their Use.** In *Encylopedia of Health Economics.* Elsevier, 2014; **2**: 341–357.
Publisher Full Text
12. Torrance GW, Boyle MH, Horwood SP: **Application of multi-attribute utility theory to measure social preferences for health states.** *Oper Res.* 1982; **30**(6): 1043–1069.
PubMed Abstract | Publisher Full Text
13. Torrance GW, Furlong W, Feeny D, *et al.*: **Multi-attribute preference functions. Health Utilities Index.** *Pharmacoeconomics.* 1995; **7**(6): 503–520.
PubMed Abstract | Publisher Full Text
14. Rowen D, Brazier J, Ara R, *et al.*: **The role of condition-specific preference-based measures in health technology assessment.** *Pharmacoeconomics.* 2017; **35**(Suppl 1): 33–41.
PubMed Abstract | Publisher Full Text
15. Herdman M, Gudex C, Lloyd A, *et al.*: **Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L).** *Qual Life Res.* 2011; **20**(10): 1727–1736.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Shah KK, Ramos-Goñi JM, Kreimeier S, *et al.*: **An exploration of methods for obtaining 0 = dead anchors for latent scale EQ-5D-Y values.** *Eur J Health Econ.* 2020; **21**(7): 1091–1103.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Dolan P: **Modeling valuations for EuroQol health states.** *Med Care.* 1997; **35**(11): 1095–1108.
PubMed Abstract | Publisher Full Text
18. Devlin NJ, Shah KK, Feng Y, *et al.*: **Valuing health-related quality of life: An EQ-5D-5L value set for England.** *Health Econ.* 2018; **27**(1): 7–22.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Costa CABE, Vansnick JC: **The MACBETH Approach: Basic Ideas, Software, and an Application.** In: *Advances in decision analysis.* Springer, 1999; **4**: 131–157.
Publisher Full Text
20. Danner M, Hummel JM, Volz F, *et al.*: **Integrating patients' views into health technology assessment: Analytic hierarchy process (AHP) as a method to elicit patient preferences.** *Int J Technol Assess Health Care.* 2011; **27**(4): 369–375.
PubMed Abstract | Publisher Full Text
21. Oliveira MD, Agostinho A, Ferreira L, *et al.*: **Valuing health states: is the MACBETH approach useful for valuing EQ-5D-3L health states?** *Health Qual Life Outcomes.* 2018; **16**(1): 235.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Oliveira MD, Mataloto I, Kanavos P: **Multi-criteria decision analysis for health technology assessment: addressing methodological challenges to improve the state of the art.** *Eur J Health Econ.* 2019; **20**(6): 891–918.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Thokala P, Devlin N, Marsh K, *et al.*: **Multiple criteria decision analysis for health care decision making--an introduction: report 1 of the ISPOR MCDA Emerging Good Practices Task Force.** *Value Health.* 2016; **19**(1): 1–13.
PubMed Abstract | Publisher Full Text
24. Angelis A, Kanavos P: **Multiple criteria decision analysis (MCDA) for evaluating new medicines in health technology assessment and beyond: the advance value framework.** *Soc Sci Med.* 2017; **188**: 137–156.
PubMed Abstract | Publisher Full Text
25. Feeny D, Furlong W, Torrance GW, *et al.*: **Multiattribute and single-attribute utility functions for the health utilities index mark 3 system.** *Med Care.* 2002; **40**(2): 113–128.
PubMed Abstract | Publisher Full Text
26. Determann D, Lambooij MS, Steyerberg EW, *et al.*: **Impact of survey administration mode on the results of a health-related discrete choice experiment: online and paper comparison.** *Value Health.* 2017; **20**(7): 953–960.
PubMed Abstract | Publisher Full Text
27. Soekhai V, de Bekker-Grob EW, Ellis AR, *et al.*: **Discrete choice experiments in health economics: past, present and future.** *Pharmacoeconomics.* 2019; **37**(2): 201–226.
PubMed Abstract | Publisher Full Text | Free Full Text
28. RStudio Inc: **Easy web applications in R**. 2013.
Reference Source
29. Jonker MF, Donkers B, de Bekker-Grob E, *et al.*: **Attribute level overlap (and color coding) can reduce task complexity, improve choice consistency, and decrease the dropout rate in discrete choice experiments.** *Health Econ.* 2019; **28**(3): 350–363.
PubMed Abstract | Publisher Full Text | Free Full Text
30. Golicki D, Jakubczyk M, Graczyk K, *et al.*: **Valuation of EQ-5D-5L health states in Poland: the first EQ-VT-based study in Central and Eastern Europe.** *Pharmacoeconomics.* 2019; **37**(9): 1165–1176.
PubMed Abstract | Publisher Full Text | Free Full Text
31. de Charro F, Busschbach J, Essink-Bot ML, *et al.*: **Some considerations concerning negative values for EQ-5D health states.** In *EQ-5D concepts and methods: A developmental history.* Springer, 2005; 171–179.
Publisher Full Text
32. Ioannidis JP, Garber AM: **Individualized cost-effectiveness analysis.** *PLoS Med.* 2011; **8**(7): e1001058.
PubMed Abstract | Publisher Full Text | Free Full Text
33. Ramos-Goñi, JM, Pinto-Prades JL, Oppe M, *et al.*: **Valuation and modeling of EQ-5D-5L health states using a hybrid approach.** *Med Care.* 2017; **55**(7): e51–e58.
PubMed Abstract | Publisher Full Text | Free Full Text
34. Johnson EJ, Shu SB, Dellaert BGC, *et al.*: **Beyond nudges: Tools of a choice architecture.** *Mark Lett.* 2012; **23**: 487–504.
Publisher Full Text
35. Nicolet A, Groothuis-Oudshoorn CGM, Krabbe PFM: **Does inclusion of interactions result in higher precision of estimated health state values?** *Value Health.* 2018; **21**(12): 1437–1444.
PubMed Abstract | Publisher Full Text
36. Green C, Brazier J, Deverill M: **Valuing health-related quality of life. A review of health state valuation techniques.** *Pharmacoeconomics.* 2000; **17**(2): 151–165.
PubMed Abstract | Publisher Full Text
37. Attema AE, Edelaar-Peeters Y, Versteegh MM, *et al.*: **Time trade-off: one methodology, different methods.** *Eur J Health Econ.* 2013; **14 Suppl 1**(Suppl 1): S53–64.
PubMed Abstract | Publisher Full Text | Free Full Text
38. Lipman SA, Brouwer WBF, Attema AE: **What is it going to be, TTO or SG? A direct test of the validity of health state valuation.** *Health Econ.* 2020; **29**(11): 1475–1481.
PubMed Abstract | Publisher Full Text | Free Full Text
39. Brazier J, Roberts J, Deverill M: **The estimation of a preference-based measure of health from the SF-36.** *J Health Econ.* 2002; **21**(2): 271–292.
PubMed Abstract | Publisher Full Text

Chapter 7

Not just another EQ-5D-5L value set for the UK: using the OPUF approach to study health preferences on the societal, group and individual level

Directly following on from chapter 6, this chapter presents the first application of the Online elicitation of Personal Utility Functions (OPUF) approach in a larger sample of the UK general population. The aim was to construct and compare EQ-5D-5L preferences on the societal, group and individual level, and OPUF made it indeed possible to not only estimate an alternative UK value set, but also to investigate the heterogeneity of preferences in granular detail.

This paper was written with three co-authors: Nancy Devlin, Ben van Hout and John Brazier. PS and JB conceived the project and designed the study. PS developed and implemented the survey software, conducted the analysis, and wrote the first draft. JB and ND supervised the project and provided input on the survey design and the analysis. All authors contributed to the interpretation of results, provided comments on the manuscript and approved the final version.

ABSTRACT

A new method has recently been developed for valuing health states, called 'Online elicitation of Personal Utility Functions' (OPUF). In contrast to established methods, such as time trade-off or discrete choice experiments, OPUF does not require hundreds of respondents, but allows estimating utility functions for small groups and even at the individual level. In this study, we used OPUF to elicit EQ-5D-5L health state preferences from the UK general population, and then constructed and compared utility functions on the societal-, group-, and individual level. A demo version of the survey is available at: https://eq5d5l.me. Data from 874 respondents were included in the analysis. For each respondent, we constructed a personal EQ-5D-5L value set. These personal value sets predicted respondents' choices in three hold-out discrete choice tasks with an accuracy of 78%.

Overall, preferences varied greatly between individuals. However, PERMANOVA analysis showed that demographic characteristics explained only a small proportion of the variability between subgroups. While OPUF is still under development, it has important strengths as a method: it produces plausible mean values for patient reported outcome instruments such as EQ-5D-5L, while also allowing examination of underlying preferences in an unprecedented level of detail.

INTRODUCTION

Preference-based measures of health, such as the EQ-5D-5L, are a widely used component of health economic evaluations. They map health states to a common currency, that is usually referred to as health state 'utility'. Utility values are needed to compute quality-adjusted life years (QALYs) and to assess and compare the health effects of different treatment options (Drummond et al., 2015; Whitehead & Ali, 2010).

Preference-based measures of health have two components. Firstly, a descriptive system which defines a number of mutually exclusive health states. Secondly, a value set, which assigns each health state a utility value. These utility values are preference-based. They require the preferences of a target population, in most cases the general population, but occasionally also patients, as input (Brazier et al., 2017b).

Health state preferences can be elicited using various different methods. Time trade-off (TTO), standard gamble (SG) and discrete choice experiments (DCE) are those most commonly used (Brazier et al., 2017a). These methods, however, have a severe limitation: they only allow the elicitation of (large) group preferences. Since little information is obtained from each individual, data from hundreds, if not thousands of individuals are required to estimate a preference function. Work by Oppe & van Hout (2017) suggests, for example, that the minimum sample size required to derive a preference model (with 20 coefficients) for the EQ-5D-5L is about 1,000 participants. While this may not be an issue when eliciting average preferences from the general population, it makes it difficult, if not impossible, to elicit preferences from smaller groups, such as patients with a particular disease. Since models can only be estimated on a group level, it is also difficult to study the heterogeneity within groups, and it is generally impractical to draw inferences about the preferences of any given individual.

We recently developed a new preference elicitation method, called Online elicitation of Personal Utility Functions (OPUF) (Schneider et al., 2022). The approach is based on previous work by Devlin et al. (2019), and allows estimating preferences on the individual person-level.

Thus far, the new method has only been applied in small pilot studies. Here, we report on the results of a larger survey of the UK population, in which we used OPUF to elicit health state preferences for the EQ-5D-5L. We exploit the approach's ability to construct preferences on the social, group, subgroup, and individual level, and to study the heterogeneity of preferences in an unprecedented level of detail. Specifically, we investigated to what extent health preferences differ between members of the UK general public, and how much of these differences can be explained by demographic characteristics.

METHODS

Sample

We recruited 1,000 participants through the Prolific online platform (Palan & Schitter, 2018) in August 2021. The sample was selected to be broadly representative of the UK general population in terms of age, sex, and ethnicity. All participants completed the EQ-5D-5L OPUF survey.

The EQ-5D-5L instrument

The EQ-5D-5L instrument is a generic preference-based measure of health-related quality of life (Herdman et al., 2011). It consists of two components: a descriptive system, which defines mutually exclusive health states and, secondly, a set of (social) values, that reflect their respective desirability.

The descriptive system defines health states along five dimensions: mobility (MO), self-care (SC), usual activities (UA), pain or discomfort (PD), and anxiety or depression (AD). Each dimension has five levels: *no*, *slight*, *moderate*, *severe*, and extreme problems or unable to do. The instrument can describe a total of 3,125 health states. These states are usually referred to by a 5-digit code, representing the severity levels: '11111' denotes full health, for example; '21111' denotes slight mobility problems but no problems on any other dimension; and '55555' denotes the (objectively) worst health state (Herdman et al., 2011, Devlin et al., 2018).

The social value set maps each health state to a utility value. Utility values range from 1, assigned to perfect health ('11111') to 0, assigned to dead. Health states that are considered worse than being dead have a negative utility value.

EQ-5D-5L health state preferences are most commonly represented by a linear additive model. It includes 20 coefficients, – four on each dimension – representing the disutility associated with the move from no problems to slight, moderate, severe, and extreme problems (Devlin et al., 2018).

The online elicitation of personal utility functions (OPUF) approach

The OPUF approach is an adaptation of the Personal Utility Function (PUF) method (Devlin et al., 2019) for use as a stand alone online survey. In contrast to traditional preference elicitation techniques (TTO, DCE, SG, etc), which are alternative-based (decompositional), the OPUF approach is attribute-based (compositional). The theoretical foundation for both, compositional and decompositional methods, lie in multi-attribute value theory. The difference between the two is the *direction* in which preferences are (de)constructed (Belton & Stewart. 2002; Keeney & Raiffa, 1993; Thokala et al., 2016).

Decompositional methods start with valuing health states. In a second step, the responses are decomposed into their components, using statistical methods. This means, the 20 EQ-5D-5L preference model parameter coefficients are inferred from respondents' holistic evaluation of health states.

In a compositional approach, the partial values for the different components of health states are elicited directly. The components are 1) dimension weights, which determine the relative importance of each dimension; 2) level ratings, which determine the relative position of the five severity levels (no, slight, moderate, severe, extreme) within each dimension; and 3) anchoring, which maps the dimension weights and level ratings on to the QALY scale. These components are then combined to construct values for entire health states.

The EQ-5D-5L OPUF survey

The EQ-5D-5L OPUF survey consists of nine steps, of which four are essential for the construction of PUFs. In the following, the steps will be briefly described. A more detailed description of the OPUF survey and its development is provided in Schneider et al., (2022). Much effort went into the design of an intuitive and easy-to-use interface. We thus recommend readers to consult the online demo version of the OPUF survey while reading through this section. It is available at: https://eq5d5l.me.

1) Warm-up (own EQ-5D-5L health state, EQ VAS)

The survey began with a question asking the participants to report their own EQ-5D-5L health state and to rate their overall health status, using the EQ VAS.

2) Level rating

Level ratings were elicited by asking participants to position ‘slight’, ‘moderate’, and ‘severe health problems’ on a visual analogue scale between 0% and 100%. The instructions stated that “a person with 100% health has no health problems”, and“a person with 0% health has extreme health problems”. Respondents are then asked "[h]ow much health does a person with slight, moderate, and severe health problems have left?”.

The level descriptions of the EQ-5D-5L are similar across dimensions. The second best level is referred to as ‘slight’ on all five dimensions, for example. We thus decided to elicit the level ratings for health problems *in general*, i.e. without reference to any particular dimension, and then applied the level ratings to all five dimensions. However, this should be seen as a simplification. The description of the worst level differs between dimensions (*extreme problems* and *being unable to do*), and, irrespective of the wording, the ratings of levels might also differ by dimension. Ideally, level ratings should thus be obtained for each dimensions separately.

3) Dimension ranking

Participants were asked to rank the worst levels of the five EQ-5D-5L dimensions (i.e. ‘I am unable to walk about’, ‘I am unable to wash and dress myself’, etc) from worst to less worse. Ties were not permitted. The selected rank order was used to tailor the presentation of the following task (4) to the individual participant.

4) Dimension swing weighting

The task showed five sliders, one for each EQ-5D-5L dimension, describing an improvement from the worst (extreme problems) to the best level (no problems) on the respective dimension. The sliders were presented in the same order as the participant had ranked them before. The first slider (the most important dimension) was set to 100. Participants were asked to use this as a yardstick to evaluate the importance of the four other dimensions. The instructions for this task were personalised. If, for example, pain/discomfort was ranked first in the previous exercise, the instructions stated: “If an improvement from ‘I have extreme pain or discomfort’ to ‘I have no pain or discomfort’ is worth 100 'health points', how many points would you give to improvements in other areas?”.

5) Validation DCE

The survey also included three DCEs. The choice sets were personalised, to cover a broad range in terms of severity (mild, moderate, severe health states) and utility differences between scenarios (easy, moderate, difficult). The choice sets always involved trade-offs, i.e. dominant or dominated states were excluded. The responses were not used to construct PUFs. The task was only included to assess the consistency between PUFs and participants’ DCE choices.

6) Anchoring I: position-of-dead

Two different methods were used to anchor PUFs on the QALY scale: all participants were asked to consider a pairwise comparison between the worst health state ‘55555’ (scenario A) and being dead (scenario B). If they preferred ‘55555’ over ‘being dead’, they immediately moved on to task 7. If they preferred ‘being dead’ over ‘55555’, a binary search algorithm was initiated, during which the health state shown in scenario A changed, adaptively, depending on the participant’s choices, to find the health state that they considered to be equivalent to ‘being dead’ (Devlin et al., 2019; Sullivan et al., 2020).

To enable the search algorithm, all 3,125 EQ-5D-5L health states were ranked from the best to the worse, based on the participant's responses to the level rating and dimension weighting. After the first comparison ('55555' vs 'being dead'), the algorithm selects the median state (which may be different for each participant). It then jumps up or down, narrowing down to the health state that is equal to being dead. After six iterations, the search ended. At this point, the equal-to-dead state is being identified with a maximum error of +/- 49 ranks (corresponding to 1.6% of the total number of EQ-5D-5L health states).

7) Anchoring II: dead-VAS

If participants prefer the worst health state, '55555', over 'being dead', the utility of '55555' could take any value between 1 and 0. We therefore asked those participants to locate the position of '55555' on a visual analogue scale between 'No health problems' (=100) and 'being dead' (=0). The selected value was then used as the anchor point for the PUF.

8) Demographic questionnaire

The OPUF survey included questions about personal characteristics, which were previously shown to be associated with EQ-5D-5L health preferences. These included: age, sex, having children, importance of religion or spirituality, the frequency of engaging in religious or spiritual activities, level of education, income, and experience with severe health problems – see *table 1* for more details (Golicki et al., 2019; MVH, 1995; Feng et al., 2018; Peeters & Stiggelbout, 2010).

9) Results page

As a thank-you to the participants, the last page of the survey showed a comparison between some of their own responses and aggregate results from English general population (obtained from Devlin et al., 2018).

Constructing Personal Utility Functions (PUFs)

PUFs were constructed for all participants. In this section, we provide an overview of the preference construction procedure and illustrate the steps with an example.

Overview

1. The level ratings for no, slight, moderate, severe, and extreme health problems were rescaled between 0 (no problems) and 1 (extreme problems).
2. The five dimension weights were normalised to sum 1.
3. The outer product of the dimension weights and the level ratings was taken to generate a set of 20 (un-anchored) model coefficients (+5 zero coefficients).
4. Depending on whether the participants considered state '55555' better or worse than dead, we either used the response from the 'dead-VAS' or from the 'position-of-dead' task to anchor the model coefficients and map them on to the QALY scale.
5. Finally, the model coefficients were used to generate utility values for all 3,125 EQ-5D-5L health states – this vector of utility values represents the PUF

Example

To illustrate the procedure, suppose a participant gave the following level ratings l with $l_{no} = 100,\ l_{slight} = 90,\ l_{moderate} = 50,\ l_{severe} = 30$, and $l_{extreme} = 0$; and the following dimension weights w with $w_{MO} = 100,\ w_{SC} = 60,\ w_{UA} = 45,\ w_{PD} = 80$, and $w_{AD} = 70$.

After rescaling the level ratings and the dimension weights, we derive the two vectors:

$$l' = \begin{bmatrix} 0 \\ 0.1 \\ 0.5 \\ 0.7 \\ 1 \end{bmatrix}; \quad w' = \begin{bmatrix} 0.29 \\ 0.17 \\ 0.11 \\ 0.23 \\ 0.2 \end{bmatrix}$$

Taking the outer product provides a matrix $\widetilde{M}$, containing 20 (1-0 scaled) coefficients (+ zero coefficients for ‘no problems’ on each dimension).

$$l' \otimes w' = \widetilde{M} = \begin{array}{cc} & \begin{array}{ccccc} w_{MO} & w_{SC} & w_{UA} & w_{PD} & w_{AD} \end{array} \\ \begin{array}{c} l_{no} \\ l_{slight} \\ l_{moder.} \\ l_{severe} \\ l_{extreme} \end{array} & \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.03 & 0.02 & 0.01 & 0.02 & 0.02 \\ 0.14 & 0.09 & 0.06 & 0.11 & 0.10 \\ 0.20 & 0.12 & 0.08 & 0.16 & 0.14 \\ 0.29 & 0.17 & 0.11 & 0.23 & 0.20 \end{bmatrix} \end{array}$$

Suppose the respondent considered state '51255' (approximately) equivalent to being dead in the ‘Position-of-Dead’ task. To rescale and anchor $\widetilde{M}$ on the QALY scale, we first compute the scaled disutility for the state equal to being dead with $\widetilde{u}(51255) =$ 0.29+0+0.02+0.23+0.2= 0.74. Subsequently, we set the utility of that state to zero and rescale the entire matrix accordingly, by simply dividing it by that value:

$$\frac{\widetilde{M}}{0.74} = M = \begin{array}{cc} & \begin{array}{ccccc} w_{MO} & w_{SC} & w_{UA} & w_{PD} & w_{AD} \end{array} \\ \begin{array}{c} l_{no} \\ l_{slight} \\ l_{moder.} \\ l_{severe} \\ l_{extreme} \end{array} & \begin{bmatrix} 0 & 0. & 0 & 0 & 0 \\ 0.04 & 0.02 & 0.02 & 0.03 & 0.03 \\ 0.19 & 0.12 & 0.08 & 0.15 & 0.14 \\ 0.27 & 0.16 & 0.11 & 0.22 & 0.19 \\ 0.39 & 0.23 & 0.15 & 0.31 & 0.27 \end{bmatrix} \end{array}$$

Note that the constructed preference model assigns state '51255' a value of 0 (= 1 - (0.39+0+ 0.02+0.31+0.27)); '11111' is still equal to 1 (= 1 - (0+0+0+0+0)), and the worst health state ('55555') now has a value of −0.35 (= 1-(0.39+ 0.23+0.15+0.31+0.27)). The model can be used to assign utility values to all EQ-5D-5L health states. The resulting vector of 3,125 utility values is taken to be a representation of the participant's PUF.

Preference Heterogeneity

Investigating the heterogeneity of preferences between individuals, requires a measure of dis/similarity to quantify how far apart two PUFs are. As stated above, a PUF was represented by a vector of 3,125 utility values (one for each EQ-5D-5L health state). It would not be useful to compare the utility values of individual health states, nor would it provide much insight to compute means or medians in this case. Instead, we assessed the dissimilarity between PUFs using the euclidean distance (ED) measure.

Analogous to a line between two points on a two dimensional plane, the ED between two PUFs denotes the shortest path length in a 3,125 dimensional space. It is computed as the square root of the sum of the squared differences between the PUFs of individuals *i* and *j*:

$$d_{EUD}(i,j) = \sqrt{\sum \left(u_i(s_1) - u_j(s_1)\right)^2 + \ldots + \left(u_i(s_{3125}) - u_j(s_{3125})\right)^2}$$

with s = {11111, 21111, ..., 55555}

The ED has a lower bound of 0, which indicates that two PUFs are identical. Theoretically, it does not have an upper bound, but due to the design of the EQ-5D-5L OPUF survey, the maximum ED between two PUFs was 1,789.

Statistical analysis

After we constructed PUFs for all participants, we computed all pairwise ED. We then performed permutational multivariate analysis of variance (PERMANOVA) to investigate the heterogeneity of preferences between subgroups.

PERMANOVA is a geometric partitioning of variation across a multivariate data cloud, defined in the space of any given dissimilarity measure, in response to one or more groups (Anderson, 2014; Anderson & Walsh, 2013). Originally developed to test for differences in dispersion in ecological data (e.g. Souza et al., 2013), in this study, we used it to investigate the variability in EQ-5D-5L health state preferences. *Analogous to ANOVA,* PERMANOVA decomposes the total distances between observations (SS_T) into within-groups (SS_W) and between groups sum-of-squares (SS_B), with

$$SS_T = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d(i,j)^2 \text{ ; and } SS_W = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d(i,j)^2 \epsilon_{ij}^{\ell}/n_{\ell}$$

where N is the total sample size (=874), $d(i,j)^2$ is the squared distance between the PUFs of participants *i* and *j*,ϵ_{ij} is an indicator which is 1, if participants *i* and *j* belong to the same group, and 0 if they do not, and n_ℓ is the size for group ℓ. Then,SS_B can then be calculated as $SS_B = SS_T - SS_W$, which allows calculating the pseudo F statistic for p groups: $F = \frac{\left(\frac{SS_B}{p-1}\right)}{\left(\frac{SS_W}{N-p}\right)}$

Semiparametric inference is achieved by permutations. The data is resampled (without replacement) and each time the F statistic is recorded. The original F statistic is then compared to the F statistics of the permutations to derive a p-value. This allows robust statistical inference in situations where more response variables than participants are observed or when the data is non-normal or zero-inflated. The null hypothesis that is investigated is that the centroids and the dispersion (however defined by the distant measure) are equivalent for all groups. The null hypothesis can be rejected either because the centroids or the spread of the distances is different. PERMANOVA was performed on the ED matrix. We first tested each of the group characteristics shown in table 1 individually, and then combined them all in one model. P-values were based on 10,000 permutations and a value below 0.05 was considered statistically significant.

RESULTS

Sample

We recruited 1,000 participants through the Prolific online platform. Data from 126 participants, who skipped one or more valuation steps, had to be excluded, because no meaningful PUF could be constructed for them. Characteristics of the 874 participants included in the study are shown in table 1.

Although we sought to recruit a representative sample of the UK population, the included sample tended to be younger (e.g. only 3% were aged 70+ versus 15% in the UK population), and more highly educated (e.g. 56% had a degree versus 40% in the population).

EQ-5D-5L OPUF survey results

On average, it took participants about nine minutes to complete the survey. The median was eight; the shortest duration was three; and the longest was 32 minutes.

TABLE 1 Sample characteristics

	n (%%)
Sex	
Female	456 (52%)
Male	413 (47%)
Other/prefer not to say	5 (1%)
Age	
18-29	189 (22%)
30-39	188 (22%)
40-49	162 (19%)
50-59	147 (17%)
60-69	164 (19%)
70+	23 (3%)
Prefer not to say	1 (0%)
Children	
No	410 (47%)
Yes	458 (52%)
Prefer not to say	6 (1%)
Education	
without qualifications	10 (1%)
GCSE/Standard grade	93 (11%)
A-Level/Higher grade	161 (18%)
Certificate/Diploma/NVQ	118 (14%)
Degree	305 (35%)
Post-graduate	181 (21%)
Prefer not to say	6 (1%)
Income	
£0 - £20,000	207 (24%)
£20,001 - £30,000	161 (18%)
£30,001 - £50,000	216 (25%)
£50,001 - £70,000	132 (15%)
£70,001+	99 (11%)
Prefer not to say	59 (7%)
Religious/spiritual practice	
Never/practically never	545 (62%)
A few times a year	132 (15%)
A few times a month	47 (5%)
Once a week	32 (4%)
A few times a week	48 (5%)
Every day	60 (7%)
Prefer not to say	10 (1%)
Importance of religion/spirituality	
Not important	476 (54%)
Slightly important	201 (23%)
Moderately important	100 (11%)
Very important	88 (10%)
Prefer not to say	9 (1%)
Experience with health problems*	
Health care professional	76 (9%)
Carer	86 (10%)
Family member	429 (49%)
Past own experience	199 (23%)
Present own experience	49 (6%)
No experience	285 (33%)
Prefer not to say	11 (1%)

*non-exclusive categories

Warm-up (own EQ-5D-5L health state, EQ VAS)

Most participants had no or only mild health problems: 216 (25%) were in full health and 404 (46%) reported slight problems on one or more dimensions.

Overall, problems were most frequently reported for the AD (n=470; 53%) and the PD dimension (n=458, 52%). The mean (SD) and median (IQR) EQ VAS score was 77.56 (15.59) and 80 (70-90), with a range of 12 to 100.

Level ratings

The mean (SD) ratings assigned to the 'slight', 'moderate', and 'severe health problems' were 80.23 (11.23); 55.61 (11.55); and 23.47 (13.18), respectively. Participants often assigned round values: 182 (21%) participants assigned a rating of 80 to the 'slight' level, and 112 (13%) assigned it a value of 90, for example.

Dimension weights

The EQ-5D-5L dimension that was, on average, considered to be most important was pain/discomfort with a mean (SD) weight of 90.05 (16.61), followed by mobility and self-care, which nearly identical weights of 82.88 (20.71) and 82.87 (20.47), and then anxiety/depression with a mean weight of 75.80 and the highest standard deviation of 24.15. The least important dimension was usual activities, with a mean (SD) weight of 73.71 (22.15).

Anchoring (position-of-dead and dead-VAS)

For 342 (39%) participants, who indicated that they would prefer state ‘55555’ over ‘being dead’, we took the anchor point from the dead-VAS task. For the remaining 532 (61%) participants, who considered ‘55555’ worse than dead, we anchored the PUF using their responses to the position-of-dead task. Figure 1 below shows the resulting bi-modal distribution of utility values for state ‘55555’. The mean (SD) utility of state ‘55555’ was -0.37 (0.83), and the lowest and highest values were -9.42 and 1.

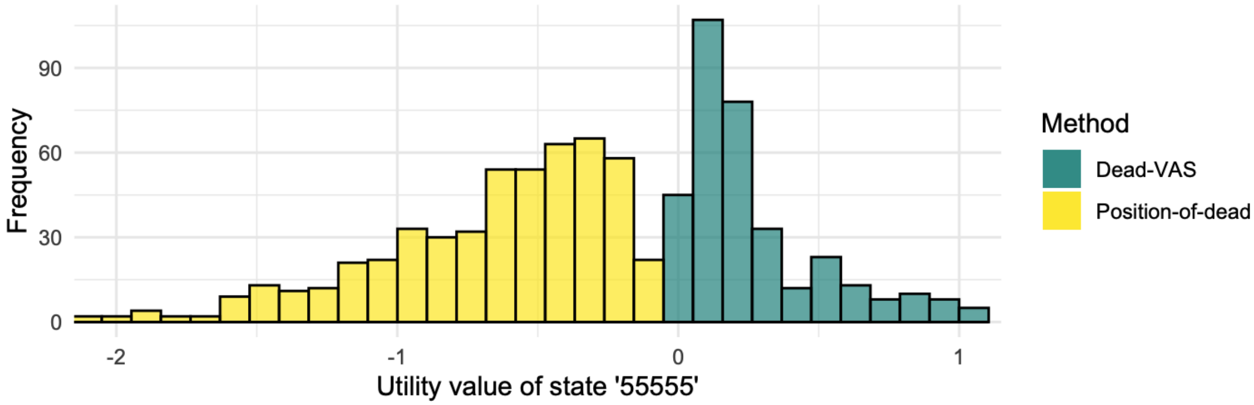

FIGURE 1 Distribution of utility values for state ‘55555’, based on the responses from either the dead-VAS or the position-of-dead task. Values below -2 are not shown (n=24).

Personal utility functions and an alternative UK EQ-5D-5L social value set

Descriptive statistics for the constructed personal EQ-5D-5L preference models are provided in table 2. The reported mean or median model coefficients may be interpreted as a social utility function, and could be used to generate an alternative EQ-5D-5L social value set for the UK.

TABLE 2 Descriptive statistics of personal EQ-5D-5L models (n=874)

	Mean (95% CI)	Median (Q1-Q3)
Mobility		
Level 2	0.055 (0.053; 0.059)	0.044 (0.024; 0.071)
Level 3	0.123 (0.121; 0.130)	0.109 (0.071; 0.156)
Level 4	0.213 (0.210; 0.223)	0.193 (0.128; 0.267)
Level 5	0.283 (0.278; 0.297)	0.252 (0.168; 0.346)
Self-Care		
Level 2	0.055 (0.054; 0.058)	0.045 (0.026; 0.071)
Level 3	0.124 (0.122; 0.130)	0.110 (0.072; 0.158)
Level 4	0.213 (0.210; 0.222)	0.192 (0.133; 0.267)
Level 5	0.282 (0.278; 0.294)	0.256 (0.174; 0.350)
Usual activities		
Level 2	0.048 (0.047; 0.051)	0.038 (0.022; 0.062)
Level 3	0.108 (0.106; 0.113)	0.096 (0.062; 0.138)
Level 4	0.186 (0.184; 0.194)	0.168 (0.110; 0.236)
Level 5	0.248 (0.245; 0.260)	0.220 (0.150; 0.317)
Pain/Discomfort		
Level 2	0.060 (0.059; 0.063)	0.050 (0.029; 0.080)
Level 3	0.136 (0.134; 0.141)	0.122 (0.082; 0.171)
Level 4	0.234 (0.231; 0.243)	0.214 (0.147; 0.293)
Level 5	0.309 (0.305; 0.322)	0.275 (0.190; 0.387)
Anxiety/Depression		
Level 2	0.049 (0.048; 0.052)	0.040 (0.020; 0.065)
Level 3	0.111 (0.110; 0.117)	0.099 (0.061; 0.145)
Level 4	0.192 (0.189; 0.200)	0.173 (0.114; 0.246)
Level 5	0.254 (0.250; 0.266)	0.227 (0.153; 0.322)

*95% CI = 95% confidence intervals, based on 10,000 bootstrap iterations; Q1 = first quartile; Q3 = third quartile

Validation DCE

Overall, PUFs predicted participants' DCE responses between non-dominant pairs with an accuracy of 78.5%. The responses of 453 (52%) participants were fully consistent, while 299 (34%) made one, 101 (12%) made two, and 21 (2%) made three 'mistakes'. We found that the consistency varied by difficulty of the DCE choice set. When the utility difference between the two presented health states was large (>0.3, measured on the personal 1-0 utility scale) 82% (325 of 395) choices were consistent. Yet, even when the utility difference was small (<0.1) and the choice was difficult, a participant's PUF still predicted their choices with an accuracy of 68% (143 of 209 of choices).

Preference heterogeneity

The average utility values for the EQ-5D-5L health states ranged from 1 to -0.37. The variability of utility values increased with severity: the mean and standard deviation (SD) of states '22222', '33333', '44444', and '55555' were 0.73 (0.22), 0.40 (0.38), -0.04 (0.60), and -0.37 (0.83), respectively. (N.B.: by definition, '11111' has a value of 1).

Figure 2 illustrates the substantial variation in participants' health state preferences. It shows the average utility values across all participants, i.e. the social value set, for a subset of 100 health states, ranked from the best to the worst (according to the social preference). The thin lines represent the 874 individual PUFs. The colour of the line indicates the ED from the average social value set.

We computed the ED between the PUFs of all participants, which yielded a 874 x 874 distance matrix with 381,501 unique pairwise comparisons. The mean (SD) and median (IQR) ED was 23.36 (23.02) and 17.95 (9.72; 29.37). The highest and lowest observed ED were 259.93 and 0.

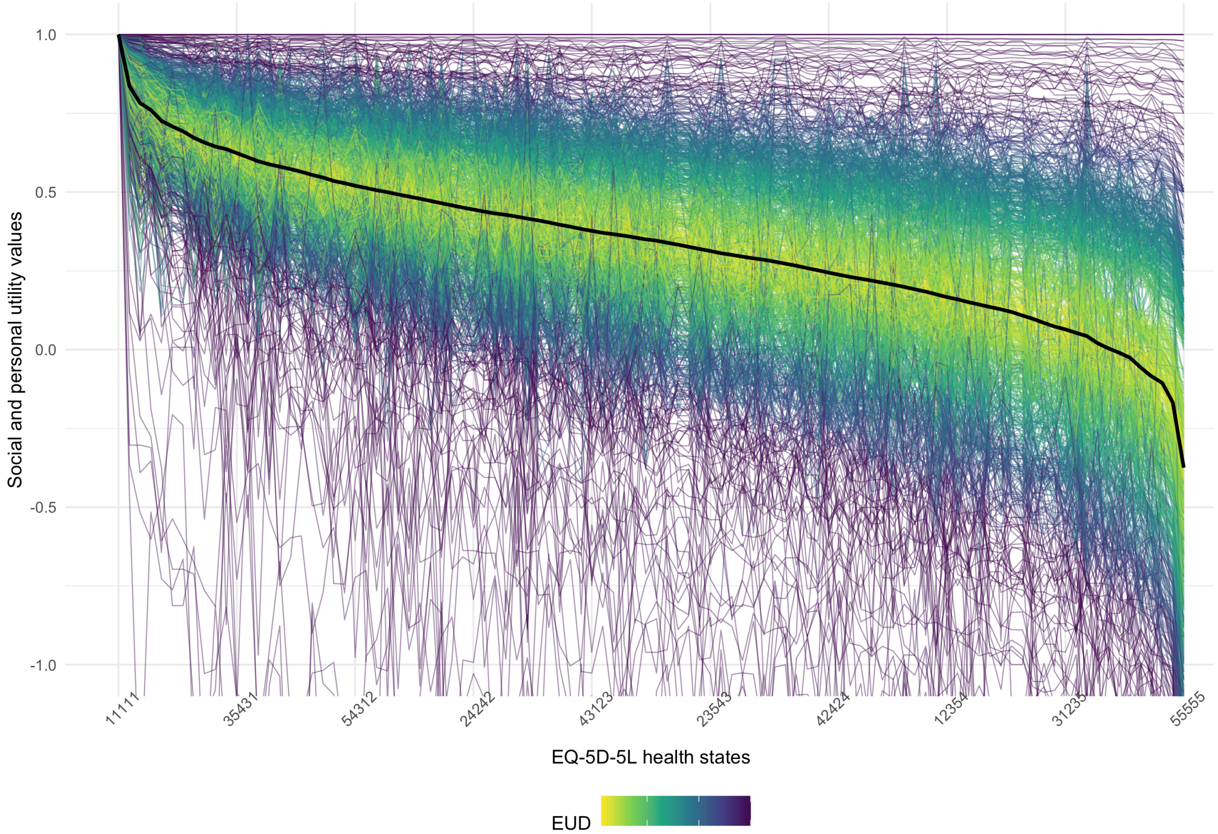

FIGURE 2 Simplified illustration of the aggregate group preference (thick black line) and the PUFs of all 874 participants. Shown are the utility values for a sample of 100 health states, ranked from the best on the left to the worst on the right (according to the aggregate group preference). The colours of the individual PUF lines indicate their euclidean distance from the average preference. Values below -1 are not shown.

PERMANOVA

Table 3 provides the results of the PERMANOVA. Shown are the within-group sum-of-squares (SS_W) for each group individually and for all groups combined, and the corresponding R^2, pseudo F, and p values. The between groups sum-of-squares (SS_B) can be computed by subtracting the SS_W from the SS_T. Significant differences between groups were observed for four group characteristics: age, having children, importance of religion/spirituality, and own EQ VAS quintiles. In addition, the effect of currently experiencing severe health problems ('present own experience') was borderline significant (p=0.0504). However, the proportions of the variance that were explained by these group characteristics individually were rather small: R^2 values ranged between 2.6% (for age) and 1.2% (for importance of religion/spirituality). The effects of group characteristics that reflected experience with health problems (e.g. being a healthcare professional, carer) were not statistically significant. The model that included all group characteristics explained 8.5% of the differences between participants' PUFs.

TABLE 3 Results of PERMANOVA – testing for differences in EQ-5D-5L health state preferences between groups characteristics

Group variable	SS_W	Df	R^2	F	p
Sex	473	2	0.1 %	0.44	0.630
Age	12180	6	2.6 %	3.85	0.008*
Having children	7877	2	1.7 %	7.43	0.008*
Education	4142	6	0.9 %	1.29	0.238
Income	4160	5	0.9 %	1.55	0.166
Importance of religion/spirituality	5708	4	1.2 %	2.67	0.034*
Religious/spiritual practice	5698	6	1.2 %	1.78	0.098
Experience w/ health problems					
Health care professional	410	1	0.1 %	0.76	0.373
Carer	188	1	0.0 %	0.35	0.569
Family member	146	1	0.0 %	0.27	0.633
Past own experience	179	1	0.0 %	0.33	0.582
Present own experience	1977	1	0.4 %	3.69	0.050
No experience	180	1	0.0 %	0.33	0.586
EQ VAS (quintiles)	5699	4	1.2 %	2.67	0.027*
All groups together	36794	41	7.8 %	1.73	0.018*
Total (SS_T)	469540	873			

SS_T = total sum-of-squares; SS_W = within-group sum-of-squares; df = degrees of freedom; F = pseudo F statistics; p values based on 10,000 permutations; * = $p<0.05$

To give some intuition for kind of differences that existed between groups, the (sub)group-specific value sets for different age groups are shown in figure 3 as an example. The colours of the plotted group-level (thick lines) and personal utility functions (thin lines) indicate group membership. For simplicity, the 'prefer not to say' group is not shown.

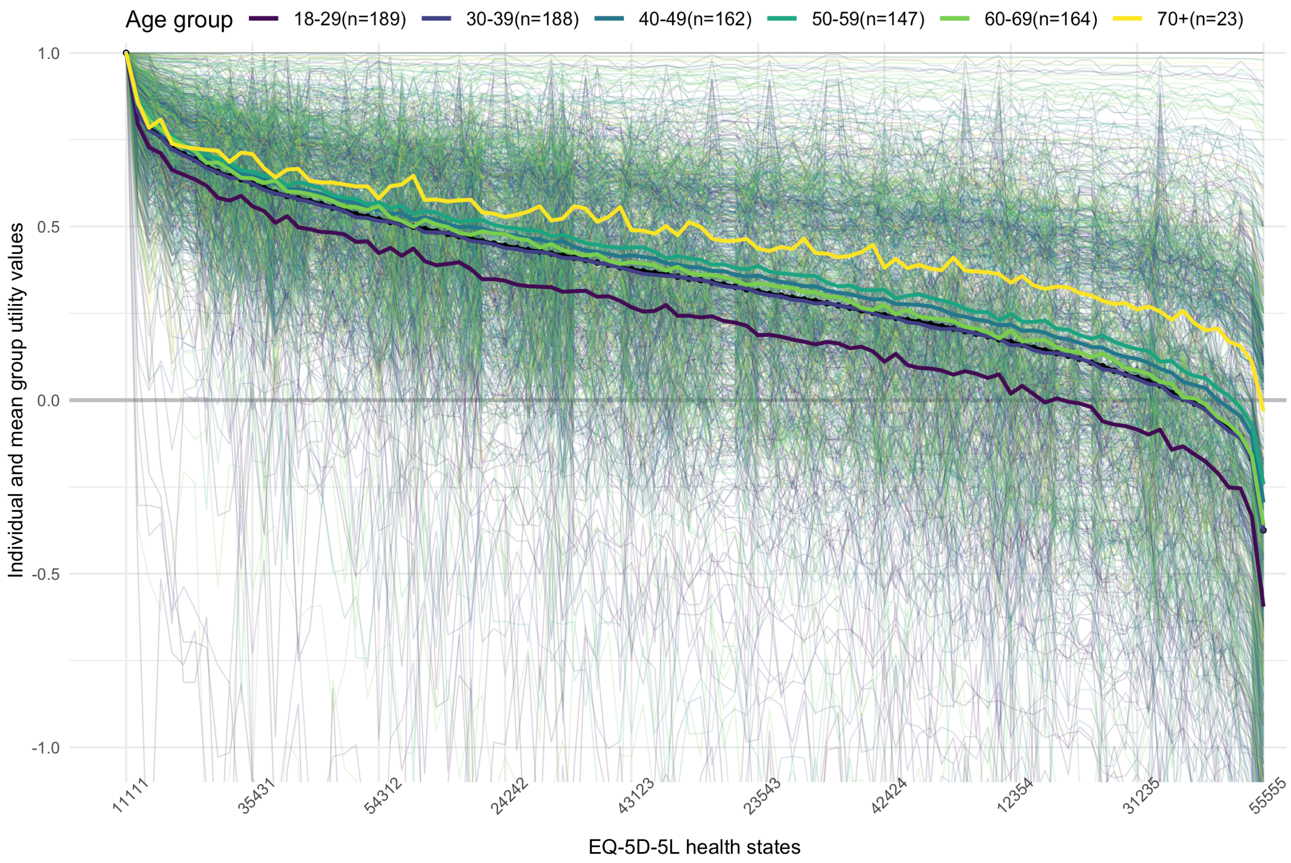

FIGURE 3 Age-group specific EQ-5D-5L health state preferences. Shown are the group level value sets (thick lines) and the underlying PUFs (thin lines), as well as the social value set (thick black line). Values below -1 and the 'prefer not to say' group are not shown.

The age group specific value sets differ from each other in two ways. Firstly, there appears to be some differences in *scale*. The curve for the youngest group (age 18-29) is the lowest. The curve then seem to move upwards with increased age, and the curve for the oldest age group (70+) is the highest. This suggests that the older the participants are, the higher they set their anchor point against dead. Secondly, the group-specific curves are not strictly decreasing, i.e. they move up and down. This indicates differences in the relative importance of health state attributes, i.e. groups assign different weights to the five EQ-5D-5L dimensions and/or differ in their level ratings. As a result, the rank order of the health states differs, and the graph fluctuates when compared to the overall social rank order. Due to the simplified visualisation of EQ-5D-5L utility functions (we only show 100 of the 3125 utility scores) this effect may appear smaller than it actually is.

DISCUSSION

This study is the first application of the newly developed OPUF approach for eliciting health state preferences in a large sample of the UK population. We constructed EQ-5D-5L value sets on the societal-, group-, and individual person level, to explore the, hitherto largely ignored, heterogeneity of health state preferences.

We found that health state preferences systematically differed between groups. Significant effects were observed in the PERMANOVA for age, having children, importance of religion/spirituality, and the EQ VAS quintile. However, the variability of preferences within groups was substantial, and individual group characteristics explained only small proportions of the ED between PUFs. For other demographic factors (sex, education, income), we observed no systematic differences between groups. Contrary to our expectations, participants' experience with severe health problems (captured by 6 non mutually exclusive categories) were also not associated with the differences in PUFs. It should be noted though, that the participants in our sample were quite ‘healthy’ – a large majority reported no or only slight problems in any of the EQ-5D dimensions.

When all characteristics were taken into account together, group membership accounted for just 8% of the variance. This result should not be considered surprising. The formation of health preferences is a complex task, which is likely to be influenced by various emotional, cognitive, and social factors (Russo et al. 2019). There is no compelling reason why demographic factors, such as age, should be good predictors of people’s health preferences. The results illustrate that aggregate group-level value sets usually say little about the preferences of any given individual – in our study, preferences differed greatly between individuals within all the groups that we considered.

In addition to allowing us to better understand the heterogeneity of EQ-5D-5L preferences within the UK general population , the OPUF method also produced highly plausible aggregate social values. Another advantage of the OPUF method is that, like DCE, it can be administered as a stand-alone online survey, thereby avoiding the cost and complexity of TTO. Yet, unlike DCE, OPUF can yield utility values that are anchored at dead (=0) and full health (=1). The fact that the survey can also be completed relatively quickly (the median completion time in our study was eight minutes), in combination with the compositional nature of the method, might make it also applicable to longer, more complex descriptive systems, like the EORTC QLC10 (King et al., 2016) or the EQ-HWB (Brazier et al., 2022).

Our study has some limitations that should be considered when interpreting the findings.

Firstly, the participants that were included in the analysis were younger and more highly educated than the general UK population. We also did not attempt to apply quality control criteria (e.g. remove participants with very fast completion times, test for response biases). The reported mean EQ-5D-5L model coefficients may not yield a representative social value set.

Secondly, preference heterogeneity can be investigated in many different ways. Designing this study thus required making several, somewhat contingent methodological choices. Instead of computing the ED between health state utility vectors, we could have assessed the differences in participants' model coefficients, or we could have computed a different distance measure – the Kendall correlation distance, for example, could be used to compare preference orderings (i.e. ordinal instead of cardinal preferences). Results may not be robust to these kinds of methodological choices.

Thirdly, we explored the variability of EQ-5D-5L health state preferences in a general sense. This means, we neither specified any hypotheses about the type or the direction of differences, nor did we test differences between subgroups. Even though the OPUF approach would have allowed us to study the health state preferences of small subgroups, in the absence of predefined hypotheses about subgroup differences, it did also not seem useful to consider the (up to 240) interaction effects between groups. For investigating more specific research questions, such as, '*do older people with strong religious beliefs people assign higher utility values to health states than the general public?*', PERMANOVA may not be the most appropriate statistical approach.

Finally, a key consideration for the interpretation of our findings is the validity of the OPUF approach. It is a new method, based on a different paradigm (compositional approach) than other, established preference elicitation methods, such as TTO, DCE, or SG (decompositional). Even though we observed a high consistency of 78%, between the constructed PUFs and participants' DCE choices, more research is needed to better understand how the OPUF approach compares to other methods, and to determine how the online survey design affects participants' preference formation. Further refinement of the survey may also be help to prevent people from skipping essential valuation tasks, and thereby reduce the number of participants who have to be excluded from the analysis.

CONCLUSION

The OPUF approach provides a flexible, conceptually attractive, alternative approach for eliciting health state preferences. The ability to construct utility functions on the individual person level opens up new and, we think, exciting avenues for research. As demonstrated in this study, the OPUF approach makes it possible to investigate the heterogeneity of health states preferences in an unprecedented level of detail. It may also enable researchers to derive value sets for small groups of participants (e.g. patients with rare diseases), for which this would otherwise be practically infeasible. Even though the OPUF approach has, thus far, only been implemented for the EQ-5D-5L, in principle, it could be applied to any descriptive system or patient-reported outcome measure.

Acknowledgements

We are very grateful to Siobhan Daley, Jack Dowie, Barry Dewitt, Irene Ebyarimpa, Job van Exel, Anthony Hatswell, Paul Kind, Johanna Kokot, Simon McNamara, Clara Mukuria, Monica Oliveira, Krystallia Pantiri, Donna Rowan, Erik Schokkaert, Koonal Shah, Robert Smith, Praveen Thokala, Ally Tolhurst, David Tordrup, Evangelos Zormpas, and the participants of the 2022 lolaHESG and the 2022 Summer HESG meeting for helpful comments, discussions of the ideas expressed in this paper, and/or for providing feedback on earlier versions of the EQ-5D-5L OPUF survey. We would also like to thank all participants who took part in this study. The usual disclaimer applies.

Data availability

The R shiny source code for the OPUF survey tool is openly available at: https://github.com/bitowaqr/opuf_demo, and all data and an annotated version of the R source code used for this study are available at: https://github.com/bitowaqr/opuf_uk upon request.

Ethical approval

The study was approved by the Research Ethics Committee of the School of Health and Related Research at the University of Sheffield (ID: 030724).

REFERENCES

Anderson MJ. Permutational multivariate analysis of variance (PERMANOVA). Wiley statsref: statistics reference online. 2014 Apr 14:1-5.

Anderson MJ, Walsh DC. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing?. Ecological monographs. 2013 Nov;83(4):557-74.

Belton V, Stewart T. Multiple criteria decision analysis: an integrated approach. Springer Science & Business Media; 2002.

Brazier J, Ratcliffe J, Saloman J, Tsuchiya A. Measuring and valuing health benefits for economic evaluation. OXFORD university press; 2017a.

Brazier J, Ara R, Rowen D, Chevrou-Severac H. A review of generic preference-based measures for use in cost-effectiveness models. Pharmacoeconomics. 2017b Dec;35(1):21-31.

Brazier J, Peasgood T, Mukuria C, Marten O, Kreimeier S, Luo N, Mulhern B, Pickard AS, Augustovski F, Greiner W, Engel L. The EQ Health and Wellbeing: overview of the development of a measure of health and wellbeing and key results. Value in Health. 2022 Mar; 25(4):482-91.

Devlin NJ, Shah KK, Mulhern BJ, Pantiri K, van Hout B. A new method for valuing health: directly eliciting personal utility functions. The European Journal of Health Economics. 2019 Mar;20(2):257-70.

Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: An EQ-5 D-5 L value set for England. Health economics. 2018 Jan;27(1):7-22.

Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. Oxford university press; 2015 Sep 25.

Feng Y, Devlin NJ, Shah KK, Mulhern B, van Hout B. New methods for modelling EQ-5D-5L value sets: An application to English data. Health Economics. 2018 Jan;27(1):23-38.

Golicki D, Jakubczyk M, Graczyk K, Niewada M. Valuation of EQ-5D-5L health states in Poland: the first EQ-VT-based study in Central and Eastern Europe. Pharmacoeconomics. 2019 Sep;37(9):1165-76.

Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Quality of life research. 2011 Dec;20(10):1727-36.

Keeney RL, Raiffa H, Meyer RF. Decisions with multiple objectives: preferences and value trade-offs. Cambridge university press; 1993.

King MT, Costa DS, Aaronson NK, Brazier JE, Cella DF, Fayers PM, Grimison P, Janda M, Kemmler G, Norman R, Pickard AS. QLU-C10D: a health state classification system for a multi-attribute utility measure based on the EORTC QLQ-C30. Quality of Life Research. 2016 Mar;25(3):625-36.

MVH Group. The measurement and valuation of health: Final report on the modelling of valuation tariffs. Centre for Health Economics, University of York. 1995.

Ombler F, Albert M, Hansen P. How significant are “high” correlations between EQ-5D value sets?. Medical Decision Making. 2018 Aug;38(6):635-45.

Oppe M, Van Hout B. The “power” of eliciting EQ-5D-5L values: the experimental design of the EQ-VT. EuroQolWorking Paper Series. 2017 Oct;17003.

Palan S, Schitter C. Prolific. ac—A subject pool for online experiments. Journal of Behavioral and Experimental Finance. 2018 Mar 1;17:22-7.

Peeters Y, Stiggelbout AM. Health state valuations of patients and the general public analytically compared: a meta-analytical comparison of patient and population health state utilities. Value in Health. 2010 Mar 1;13(2):306-9.

Russo S, Jongerius C, Faccio F, et al. Understanding patients' preferences: a systematic review of psychological instruments used in patients' preference and decision studies. Value in Health. 2019 Apr 1;22(4):491-501.

Schneider PP, van Hout B, Heisen M, Brazier J, Devlin N. The Online Elicitation of Personal Utility Functions (OPUF) tool: a new method for valuing health states. Wellcome Open Research. 2022 Jan 14;7:14.

Souza AT, Dias E, Nogueira A, Campos J, Marques JC, Martins I. Population ecology and habitat preferences of juvenile flounder Platichthys flesus (Actinopterygii: Pleuronectidae) in a temperate estuary. Journal of Sea Research. 2013 May 1;79:60-9.

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’. Social Science & Medicine. 2020 Feb 1;246:112707.

Thokala P, Devlin N, Marsh K, et al. Multiple criteria decision analysis for health care decision making—an introduction: report 1 of the ISPOR MCDA Emerging Good Practices Task Force. Value in health. 2016 Jan 1;19(1):1-3.

Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. British medical bulletin. 2010 Dec 1;96(1):5-21.

Chapter 8

Using OPUF to derive a patient-based value set

This chapter reports on a study testing the OPUF approach to derive an EQ-5D-5L value set from a relatively small sample of patients with rheumatic diseases in Germany. The survey was generally well received and we were able to derive a logically consistent value set.

The sample was chosen because collaborators from the German Institute for Quality and Efficiency in Health Care (IQWIG) were interested in contrasting (informed) patient-based with (uninformed) population-based value sets. Rheumatic disease patients are likely to have experienced problems in all five domains of the EQ-5D, making a future comparison particularly informative.

Following the study reported in the previous chapter, the OPUF survey was further refined. An important change was made to the 'anchoring task'. While in the previous version, two different methods were used, depending on whether or not the worst state was preferred over being dead, the new survey uses only one method (see methods section below), to make it more internally consistent. The layout and design of the survey was also revised, to make the tool more flexible.

This paper was written with four co-authors, Katharina Blankart, John Brazier, Ben van Hout, and Nancy Devlin. PS and KB conceived the study. PS developed and implemented the survey software, recruited participants, conducted the analysis and wrote the first draft of the paper. KB, JB, ND provided input to the survey design and analysis. KB and JB supervised the project. All authors reviewed, edited, and approved the final version.

Using the Online Elicitation of Personal Utility Functions (OPUF) approach to derive a patient-based EQ-5D-5L value set: a study in 122 patients with rheumatic diseases from Germany

Schneider P[1,2], Blankart K[2], Brazier J[1], van Hout B[1,3], Devlin N[1,4]

[1]University of Sheffield, Sheffield, UK; [2]University of Duisburg/Essen, Essen, Germany;
[3]Open Health, York, UK; [4]University of Melbourne, Melbourne, Australia

Abstract

Objectives

Traditional preference elicitation methods, such as DCE or TTO, usually require large sample sizes. This can limit their applicability in patient populations, where recruiting a sufficient number of participants can be challenging.

The objective of this study was to test a new method, called the Online elicitation of Personal Utility Functions (OPUF) approach, to derive an EQ-5D-5L value set from a relatively small sample of patients with rheumatic diseases.

Methods

OPUF is a new type of online survey that implements compositional preference elicitation techniques. Central to the method are three valuation steps: (1) dimension weighting, (2) level rating, and (3) anchoring. An English demo version of the OPUF survey can be accessed at https://valorem.health/eq5d5l.

From the responses, a personal EQ-5D-5L utility function can be constructed for each participant, and a group-level value set can be derived by aggregating model coefficients across participants.

Results

A total of 122 rheumatic disease patients from Germany completed the OPUF survey. The survey was generally well received, and most participants completed the survey in less than 20 minutes. We derived a plausible, logically consistent EQ-5D-5L value set. The precision of mean coefficients was high, despite the small sample size.

Conclusions

Our findings demonstrate that OPUF can be used to derive an EQ-5D-5L value set from a relatively small sample of patients. Even though the method is still under development, we think that it has the potential to be a valuable preference elicitation tool and to complement traditional methods in several areas.

HIGHLIGHTS

- Using the recently developed OPUF method, we successfully derived an EQ-5D-5L value set from a sample of 122 rheumatic disease patients from Germany.
- The valuation survey was easy to implement, and the results suggest that it was feasible and acceptable to the participants.
- We think OPUF is a promising new approach for eliciting health preferences from patients, which could complement standard methods, especially when sample sizes are small.

INTRODUCTION

Many health technology assessment agencies recommend that QALY-weights are to be derived from the preferences of the general public. Nevertheless, decision makers may – and probably should - be keen to also consider the patients' perspective when making decisions about the allocation of limited healthcare resources. This can be done formally, by including patient preferences in the decision-making process.1–7

Eliciting preferences from patients, however, can be difficult.[4,8] Traditional elicitation methods, such as discrete choice experiments (DCEs), time trade-off, or standard gamble, require large sample sizes:[9,10] several hundred participants are commonly needed to estimate a reliable preference model. Recruiting such a large number of patients for a study can be difficult, and in many cases (e.g. rare diseases) it will not be feasible at all. This limits the availability of quantitative evidence on patient preferences and thus the use of patient preferences to inform health policy decision making.

Recently, a new method, called Online elicitation of Personal Utility Functions (OPUF), was developed.[11,12] It implements multiple compositional preference elicitation techniques into an adaptive and easy to use online tool. OPUF allows constructing preferences for small groups and even on the individual person level.

Here, we report the results of a valuation study that was conducted to test the feasibility of using OPUF to elicit EQ-5D-5L health state preferences from patients with rheumatic diseases in Germany. We demonstrate how the new method allows constructing a value set based on a sample of just 122 participants.

METHODS

Respondents

The valuation study was conducted in Germany between May and July 2022. Participants were recruited through the German Rheumatism Association (Deutsche Rheuma-Liga e.V., DRL). The invitation to participate in our study was distributed to their members through a newsletter and social media. Participants were offered a financial incentive of €5. The survey was open to anyone who identified as a patient with a rheumatic disease. We did not specify any exclusion criteria.

EQ-5D-5L

The EQ-5D-5L is the most widely used generic measure of health-related quality of life.[13,14] It consists of five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), each of which has five levels (no problems, slight problems, moderate problems, severe problems, and extreme problems). EQ-5D-5L health states are commonly denoted by a 5-digit code, representing the respective dimension-levels. The system can describe a total of 3,125 health states, ranging from '11111' (full health, no health problems) to '55555' (the worst state with extreme problems on all five dimensions). While population-based EQ-5D-5L value sets currently exist for 25 countries (more studies are presently ongoing),[15] patient-based value sets are scarce.[16,17]

OPUF – general method

OPUF is a web-based version of the Personal Utility Function (PUF) approach, originally developed by Devlin et al.[12]. The main difference to traditional valuation methods is that OPUF based on a different paradigm: it is a compositional, instead of a decompositional, preference elicitation technique.[18,19] While DCE or TTO require participants to evaluate entire health states, from which partial values for dimension-level coefficients are subsequently inferred (= decompositional approach), in the OPUF method, partial values are elicited directly from the participants. The process broadly consists of three steps:

Firstly, criteria weighting, which determines the relative importance of the different dimensions of the HRQoL measure, on a scale from 0 to 100.

Secondly, level rating, which determines, within each dimension, the relative importance of any intermediate levels (e.g. slight, moderate, severe problems walking about), on a scale that is anchored at the worst (unable to walk about = 0) and the best level (no problems walking about = 100).[20]

Finally, anchoring, which is an additional step that allows mapping the values obtained in the previous steps on to the QALY scale, which is anchored at full health (=1) and dead (=0).

From the responses, an additive preference model, containing one coefficient for each dimension-level, can then be derived, for each participant, by multiplying the level ratings with the respective dimension weights, and rescaling the resulting values to the QALY scale, using the anchoring factor.

This model can then be used to derive values for any health state, by multiplying the dimension-level coefficients with the respective dimension-levels of the health state, and summing up the values, and subtracting them from 1.

A more detailed description of the OPUF method can be found in[11], and a formal description of method used to construct personal preference models from individual level responses, and a simple example, can be found in the appendix (see S1).

OPUF - Implementation

For this study, we adapted a previous (English) version of the EQ-5D-5L OPUF survey tool and translated it into German.[11] The survey was built using modern JavaScript frameworks (Vue.js for the front-end; Node.js for the back-end). The different valuation tasks were tsted in online interviews with a small group of lay persons and experts, and piloted it in a small study with 25 participants, recruited though prolific.[21]

A demo version of the final survey (in German) that was used for this study can be found under the following url: https://valorem.health/rl2022. An English translation of (a newer version of) the survey is available at https://valorem.health/eq5d5l.

In the following, we provide a brief overview of the different valuation steps and how they were implemented in the study.

Warm-up

At the beginning of the survey, each participant was asked to report their own EQ-5D-5L health state and rate their subjective level of health using the EQ VAS. This was done not only to assess the health of the participants. but also to familiarise participants with the instrument.

Criteria weighting

First, participants were shown a list of the worst problems of each dimensions. Depending on their choice, the respective dimension was set to 100 in the subsequent task, in which participants had to complete a swing rating exercise. The task was to assign values between 0 and 100 to swings from the worst to the best level on each dimension. The 100 points, assigned to the most important dimension, were fixed, and used as a yard stick to help participants to determine the relative importance of the other four dimensions. The order of the dimensions was randomised.

Level rating

For each dimension, the participants were asked to place the 3 intermediate levels (slight, moderate, severe problems) on a scale from 0 (no problems) to 100 (extreme problems). For this, participants had to 'drag and drop' labels with the respective level description onto the scale. This method avoids any anchoring effects. The order of the five dimensions was randomised.

Anchoring

The anchoring task consists of two steps. The task begins with a pairwise comparison between the objectively worst EQ-5D-5L health state '55555' (Option A) and 'being dead' (Option B). Depending on their stated choice, participants got to see different tasks.

Option A, If a participant preferred '55555' over 'being dead', they were asked to locate the position of '55555' on a visual analogue scale between 'No health problems' (=100) and 'being dead' (=0). The selected value was then used as the anchor point for the personal utility function.

Option B, if participants preferred 'being dead' over '55555', a binary search algorithm was initiated, in which the state that was shown as option A adaptively changed, to find the health state that they considered to be equivalent to 'being dead'.[12,22] For this, all 3,125 EQ-5D-5L health states were ranked from the best to the worse, based on the participant's responses to the previous tasks. After the first comparison ('55555' vs 'being dead'), the algorithm selects the median state (which may be different for each participant). Depending on the participant's subsequent choices, the algorithm then jumps up or down in half interval steps. After six iterations, the rank of the equal-to-dead state is identified with a maximum error of +/- 49 ranks, and the search ends. The normalised utility value of the equal-to-dead state is then used to rescale and anchor the personal utility function.

Demographic questions and feedback

At the end of the survey, we asked for basic demographic information and rheumatic diseases diagnoses. Participants were also invited to share feedback on the survey and to make suggestions for improvement.

Data analysis and modelling

Participants' responses were analysed on multiple levels.

First, we assessed the raw response data of the three valuation steps separately.

Secondly, we constructed personal EQ-5D-5L utility functions for each participant. The utility functions were specified as additive models with 20 coefficients – four for each of the five dimensions, representing the utility decrement associated with levels 2-5. The models were constructed using the procedure described above. For a detailed description of the procedure used to construct the personal preference models from individual level responses, and a simple example, Please see S1 in the appendix.

Finally, we aggregated the personal utility functions model coefficients to derive a preference function for the group as a whole. This was done by averaging the coefficients across all participants. The aggregate preference function was then used to generate an EQ-5D-5L value set, i.e. QALY-weights for all 3,125 EQ-5D-5L health states.

All analyses were conducted in R 4.2.1.[23]

Engagement

One indicator of engagement that we assessed was the time participants spent on completing the survey. Furthermore, the levels of the EQ-5D-5L instrument have a predefined order: slight problems, for example, (weakly) dominates moderate problems. We can therefore, utilise the level rating task to check participants' understanding and their engagement with the task, by assessing the frequency of implausible level ratings, that means ratings that violate the correct order of the levels.

DCE hold-out tasks

Participants completed three forced choice DCE holdout tasks. The tasks were generated adaptively, based on participants' PUF. For each participant, choices were selected to have a utility difference between the two states of around 0.05 (hard), 0.1 (moderate), and 0.25 (easy) on the personal utility scale. The order of the DCEs was randomised. Trivial choices, involving dominated or dominating alternatives, were excluded. We predicted participants' choices in the DCE hold-out tasks, based on personal utility functions, and then compared those against the observed choices.

Feedback

At the end of the survey, participants were invited to share feedback on the survey and make suggestions for improvement. A formal qualitative analysis of the responses is beyond the scope of this paper. However, we performed a crude thematic analysis to assess how the survey was received and whether any potential issues were raised, which would indicate problems with the response data.

Ethical approval

The study was approved by the ethics committee of the University of Sheffield, UK and the University of Duisburg/Essen, Germany. All participants provided informed consent.

Data and code availability

The data and source code of the survey as well as the analysis scripts have been uploaded onto Github. We are happy to share access to the respective repositories upon request. Please contact the corresponding author for more information.

RESULTS

Sample demographic and health characteristics

A total of 122 participants completed the survey between May and July 2022. Most participants were female (n = 111, 91%). Further demographic information is provided in table 1.

Table 1: sample characteristics

Group	N (%)
Age	
18-29	18 (14.8%)
30-39	23 (18.9%)
40-49	20 (16.4%)
50-59	36 (29.5%)
60-69	18 (14.8%)
70+	7 (5.7%)
Sex	
Female	111 (91%)
Male	9 (7.4%)
Other	2 (1.6%)
Highest secondary education degree	
University entrance qualification	67 (54.9%)
Entrance qualification for universities of applied sciences	25 (20.5%)
Intermediate secondary education	22 (18%)
Basic secondary education	4 (3.3%)
None/other	4 (3.2%)
Total number of participants	122 (100%)

Participants reported various rheumatic disease diagnoses. The most common conditions were: rheumatoid arthritis (n = 72, 59%), osteoarthritis (n = 31, 25%), psoriatic arthritis 21 (17%), fibromyalgia (n = 21, 17%) and chronic pain syndrome (n = 19, 16%). Sixty-four (52%) participants reported more than one condition. Further details can be found in the appendix (see table S2).

Corresponding to the prevalent health conditions, many participants reported poor EQ-5D-5L health states. Only one (1%) participant was in full health, and 22 (18%) reported having only mild health problems, while 41 (34%) reported having severe or extreme problems on at least one EQ-5D dimension. The most frequently affected dimension was pain/discomfort (n = 118, 97%) with a mean (SD) severity score of 2.9 (0.78); followed by usual activities (n = 104, 85%), with a score of 2.5 (0.79); anxiety/depression (n = 90, 74%) with a score of 2.2 (0.95); mobility (n = 87, 71%) with a score of 2.2 (0.97); and lastly self-care (n = 52, 43%), with a score of 1.6 (SD = 0.79). The mean (SD) and median (IQR) EQ VAS was 61.1 (18.14) and 61.5 (48.0 - 75.0), respectively.

OPUF survey results

Criteria weighting

The most important EQ-5D dimension was pain/discomfort (mean = 90.1, SD = 16.4), followed by usual activities (mean = 67.6, SD = 17.1), self-care (mean = 85.8, SD = 20.2), and mobility (mean = 84.9, SD = 15.9). The least important dimension was anxiety/depression (mean = 72.9, SD = 27.7).

Level rating

Level ratings were similar across all five dimensions. The mean ratings for the intermediate levels, slight, moderate, and severe problems, were around 75, 50, and 22, respectively. The ratings for the best and the worst level were fixed at 100 and 0. Full results of the level ratings are provided in table S3 in the Appendix.

Anchoring

The majority of participants (n = 89, 73%) indicated that they would prefer 'being dead' over the worst EQ-5D-5L health state ('55555'). The mean (SD) and median (IQR) utility values for the worst health state were -0.32 (0.52) and -0.26 (0.1; -0.65), respectively. A total of 11 (9%) participants had utility scores below -1, with one participant having a utility score as low as -1.9 for state '55555'.

Personal and group-level utility functions

For each participant, we successfully constructed a personal EQ-5D-5L utility function, using their individual responses from the criteria weighting, the level rating, and the anchoring task.

Model coefficients were aggregated across participants, by means of averaging, to obtain a utility function that reflects the preferences of the group of patients with rheumatic diseases as a whole. Table 2 below shows the resulting mean coefficient estimates and bootstrapped 95% confidence intervals.

The reported coefficients can be used to construct utility values for any of the 3,125 EQ-5D-5L health states. For example, the utility value for the health state '12345' is $1-(0+0.066+0.141+0.222+0.222)=0.35$. The utility values of states '22222', '33333', '44444', and '55555' are 0.69, 0.35, -0.05, and -0.32, respectively, to give just a few more examples.

A simplified visualisation of the group value set - compared to all 122 individual patient value sets - is provided in figure 1. The graphs illustrate that even though the study was conducted in a population of patients with similar health problems, the personal value sets showed a considerable degree of heterogeneity.

Table 2: EQ-5D-5L value set based on the preference data of 122 patients with rheumatic diseases from Germany – shown are mean coefficients and bootstrapped 95% confidence intervals, based on 10,000 iterations.

	Mobility	Self-care	Usual activities	Pain/ discomfort	Anxiety/ depression
lvl2	0.063 (0.055; 0.072)	0.066 (0.056; 0.078)	0.066 (0.058; 0.075)	0.061 (0.053; 0.070)	0.054 (0.047; 0.062)
lvl3	0.129 (0.117; 0.142)	0.133 (0.119; 0.146)	0.141 (0.128; 0.155)	0.136 (0.123; 0.150)	0.115 (0.102; 0.128)
lvl4	0.208 (0.190; 0.227)	0.212 (0.192; 0.232)	0.227 (0.207; 0.248)	0.222 (0.202; 0.242)	0.176 (0.159; 0.194)
lvl5	0.269 (0.248; 0.290)	0.270 (0.247; 0.293)	0.280 (0.258; 0.303)	0.281 (0.261; 0.301)	0.222 (0.202; 0.243)

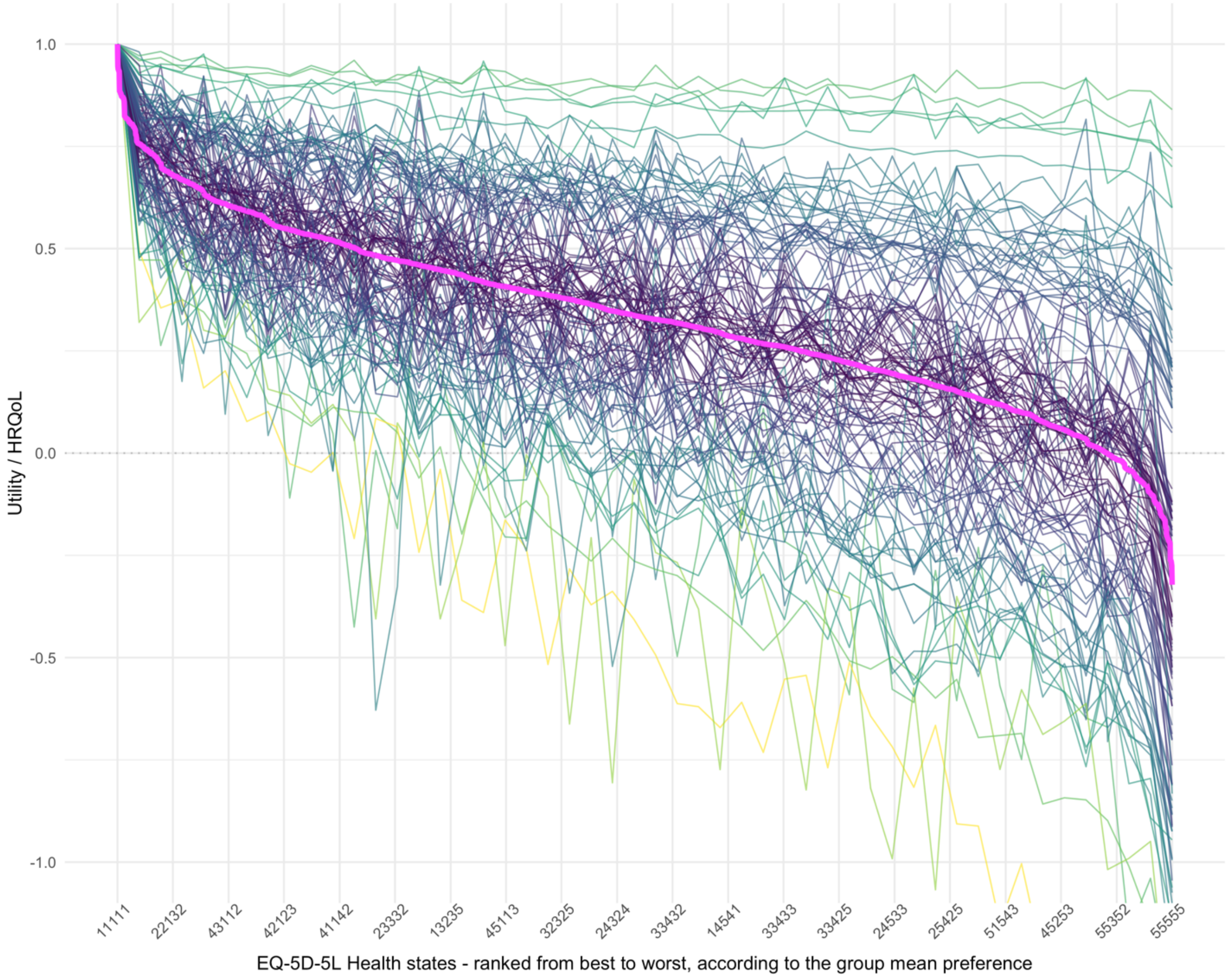

Figure 1: Simplified illustration of the aggregate group preference (thick line) and the personal utility functions of all 122 participants. Shown are the utility values for a sample of 50 health states, ranked from the best on the left to the worst on the right (according to the aggregate group preference). The colours of the lines representing personal preference functions indicate their euclidean distance from the average preference: purple (smaller distance) to yellow (greater distance). Utility values below -1 are not shown.

Engagement

On average, participants spent about 17 minutes (SD = 16) on the OPUF survey. The median completion time was 13.6 minutes (IQR 9.9; 18.7). The fastest participant completed the survey in 4.5 minutes, which is still within reasonable time limits.

Using the 15 individual level rating tasks as a means to assess participants' understanding and attention, we found that participants were generally able to rate the intermediate levels in a consistent way across the five EQ-5D dimensions. Only 13 participants (11%) made two or more errors, while 10 participants (8%) made one error, and 99 participants (81%) made no errors. The results suggest a high level of engagement and understanding of the level rating task.

DCE hold-out tasks

Overall, the constructed personal utility functions predicted participants' observed choices in the three hold-out DCE tasks with an accuracy of 67.8%. The predictive accuracy varied by difficulty: the accuracy was 60.7% for the hard DCE task (with a utility difference of 0.05 between the two states in the choice set), 61.5% for the medium task (0.1 difference), and 81.1% for the easy task (0.25 difference). Thirty-nine participants (32%) made no 'error', 52 (43%) made one, 27 (22%) made two, and 4 (3%) made three 'errors'.

Feedback

Of the 122 participants, 49 (40%) sent feedback and/or suggestions to improve the survey. The median word count was 19 (IQR: 10; 30). The crude thematic analysis identified five main themes, in which most of the responses could be categorised (responses could be categorised in multiple themes) - see list below. No major issues were identified, but some suggestions for improvement were made.

1. Introspection/reflection: Twenty participants (41%) found the survey interesting or thought-provoking. Some also reported that the survey helped them to better understand their own priorities.

2. Difficulties: Fourteen participants (29%) made suggestions for improvement and/or reported difficulties with certain aspects of the survey, including the navigation on specific tasks, instructions for the level rating task, and the handling of the sliders.

3. Overall assessment: Eight participants (16%) submitted an overall evaluation of the survey, which ranged from 'very good' ["sehr gut"] to 'so-so' ["geht so"].

4. Unrealistic states: Five participants (10%) commented on the DCE task and noted that some states were unrealistic or implausible (note: states were generated randomly, and thus not necessarily realistic).

5. Other: Eight participants (16%) provided comments that were not directly related to the survey.

DISCUSSION

Main findings

Based on a sample of 122 participants, we were able to construct an EQ-5D-5L value set for patients with rheumatic disease from Germany. Despite the small sample size, the precision of mean coefficient estimates was high and the resulting model was internally consistent. The reported health state utilities could be readily used in health economic evaluations to gain insight into the to value of different treatments for rheumatic diseases from the patients' perspective.

This is the first study to apply the OPUF method in a population of patients with rheumatic diseases - or any patient population, for that matter. So far, OPUF has only been used in smaller pilot studies and in samples of the general population.[11,12] The results show that the method is feasible and acceptable. The feedback we received further suggests that the OPUF survey was generally well received; many participants even found it to be interesting and thought-provoking. Notwithstanding, several participants also reported difficulties with the navigation, handling, or instructions on specific tasks. These issues should be addressed in future studies.

Comparison with previous studies

Even though the OPUF method is still under development and further refinement may be needed, our findings can be compared with the results from previous study. Only very recently, Ludwig et al.[16] conducted a DCE study to elicit EQ-5D-5L health state preferences from 453 patients with rheumatic arthritis in Germany. The ranking of the dimensions, as well as the absolute coefficient values reported in their paper are considerably different from our results. The most important dimension for the patients in Ludwig et al's study was self-care (level 5 coefficient = 0.364), followed by mobility (0.355), pain/discomfort (0.339), anxiety/depression (0.330), and lastly usual activities (0.272). In our study, in contrast, we found that the most important dimension was pain/discomfort (level 5 coefficient = 0.281), followed by usual activities (0.280), self-care (0.270), mobility (0.269), and anxiety/depression (0.222). It should also be noted that their final model contains three logical inconsistencies in the level ordering. Within the pain/discomfort dimension, for example, the level 2 coefficient has a higher value than the level 3 coefficient (0.121 vs 0.089).

There are several possible explanations for these differences. Firstly, Ludwig et al. used a different method, namely DCE, to elicit preferences. Secondly, the characteristics of the patient sample were considerably different from our sample in terms of age, sex, and the reported health conditions. Thirdly, because of the DCE design, patient preferences were estimated on a latent scale, which was then anchored using the pits state value (=-0.661) from the official German EQ-5D-5L value set[24], which may also have influenced the results.

A direct comparison between our results and the official German value set[24] may be difficult to interpret, because of the differences in the target population. It seems noteworthy, however, that despite the large differences in sample sizes, the precision of the coefficient estimates was similar in both studies. The official German EQ-5D-5L valuation study used the EQ-VT protocol and included 1,158 participants. The reported standard errors around mean estimates range between 0.008 and 0.011. This corresponds to 95% confidence interval widths of (0.006*3.92=) 0.024 to (0.011*3.92=)0.043, which is comparable to the 95% confidence intervals we achieved in our study with a sample of 122 participants.

Strengths and limitations

Our study has several strengths but also limitations that should be considered when interpreting the results.

First of all, we would like to note that any comparison of our findings with results from other studies should take into account the differences in the valuation method. It may well be the case that value sets generated with OPUF systematically differ from value sets generated with other methods. Some indication for this can be seen in the fact that the personal utility functions we constructed were not good predictors of participants' choices in DCE hold-out tasks; overall, the predictive accuracy was 67.8%. This is not necessarily surprising. It is well established that different valuation methods tend to produce different results.[25,26] Moreover, OPUF is based on a different theoretical framework; it is a compositional preference elicitation technique, in which participants evaluate each dimension, and each dimension-level individually.[18] This approach has several advantages, as demonstrated in this study, but it also requires making stronger assumptions about the underlying preference structure. In particular, the OPUF method assumes an additive model, which may not always be appropriate.

Decompositional methods, like DCE or time trade-off, on the other hand, involve holistic evaluations of entire health states. This requires more participants, and often more intricate statistical modelling,[27] but in principle, it is able to accommodate more complex, non-additive preference structures - although, in practice, few studies have actually done so, and it was found that interaction terms generally do not markedly improve model fit.[28]

We think that neither approach can be said to be inherently superior to the other. To reiterate a common refrain, there is no gold standard. The choice of method should be based on the research question and the context.[19,29]

The OPUF method may be particularly well suited when it is difficult to recruit a large sample of participants (note, using DCE, it would most likely not have been feasible to construct an EQ-5D-5L value set for 122 patients). Moreover, the fact that dimensions are evaluated individually may make the OPUF method useful for valuing instruments that are more complex than the EQ-5D-5L (such as the EQ-HWB[30]) - when there many dimensions, it can difficult for participants to evaluate all dimensions simultaneously.

From a practical perspective, the OPUF method is very flexible. It can be adapted to different settings and instruments, is easy to implement as a stand-alone online survey, and can be completed in a short time; most participants completed the survey in less than 20 minutes.

Notwithstanding the potential advantages of the OPUF method noted above, it is important to acknowledge that the OPUF method is still under development. The feedback we received suggests that further refinement may be needed to ensure that all participants can complete the survey without difficulties. Since this was a self-complete online study, we cannot rule out the possibility that some participants did not fully understand the tasks or did not pay attention, even though completion times and error rates indicated a good level of engagement.

Finally, since this study was conducted online and participants were recruited through a patient organisation, participants in our study are unlikely to be representative of patients with rheumatic diseases in Germany.

CONCLUSION

This study demonstrates, for the first time, that it is possible to use the OPUF approach to derive an internally consistent EQ-5D-5L value set from a relatively small sample of patients with rheumatic diseases. Our results show that the OPUF method is feasible and the feedback we received further suggests that it was generally well received by the participants. Even though OPUF is still under development, we think that it has the potential to complement traditional preference elicitation methods, especially in situations where it is difficult to recruit a large sample of participants.

ACKNOWLEDGEMENTS

We would like to thank the German Rheumatism Association (Deutsche Rheuma-Liga e.V., DRL) and in particular Jürgen Clausen, for their support in recruiting participants for this study. We would also like to thank Jörg Lauterberg, Sarah Mostardt, and Anja Schwalm from the Institute for Quality and Efficiency in Health Care (IQWiG) for their useful comments on the study design. Finally, we are grateful to all participants for their time and effort.

This work was supported by the Wellcome Trust (108903/Z/19/Z), the University of Sheffield, and the CINCH Health Economics Research Center, at the University of Duisburg/Essen.

REFERENCES

1. Bekker-Grob EW de, Berlin C, Levitan B, et al. Giving patients' preferences a voice in medical treatment life cycle: The PREFER public–private project. *The Patient-Patient-Centered Outcomes Research*. 2017;10:263-266.

2. Hoos A, Anderson J, Boutin M, et al. Partnering with patients in the development and lifecycle of medicines: A call for action. *Therapeutic innovation & regulatory science*. 2015;49(6):929-939.

3. Bouvy JC, Cowie L, Lovett R, Morrison D, Livingstone H, Crabb N. Use of patient preference studies in HTA decision making: A NICE perspective. *The Patient-Patient-Centered Outcomes Research*. 2020;13(2):145-149.

4. Dirksen CD. The use of research evidence on patient preferences in health care decision-making: Issues, controversies and moving forward. *Expert review of pharmacoeconomics & outcomes research*. 2014;14(6):785-794.

5. Rowen D, Brazier J, Ara R, Azzabi Zouraq I. The role of condition-specific preference-based measures in health technology assessment. *Pharmacoeconomics*. 2017;35(1):33-41.

6. Longworth L, Yang Y, Young T, et al. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: A systematic review, statistical modelling and survey. *Health Technology Assessment*. 2014.

7. Versteegh M, Brouwer W. Patient and general public preferences for health states: A call to reconsider current guidelines. *Social Science & Medicine*. 2016;165:66-74.

8. Van Overbeeke E, Janssens R, Whichello C, et al. Design, conduct, and use of patient preference studies in the medical product life cycle: A multi-method study. *Frontiers in pharmacology*. 2019;10:1395.

9. Bekker-Grob EW de, Donkers B, Jonker MF, Stolk EA. Sample size requirements for discrete-choice experiments in healthcare: A practical guide. *The Patient-Patient-Centered Outcomes Research*. 2015;8(5):373-384.

10. Oppe M, Devlin NJ, Hout B van, Krabbe PF, Charro F de. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value in Health*. 2014;17(4):445-453.

11. Schneider PP, Hout B van, Heisen M, Brazier J, Devlin N. The online elicitation of personal utility functions (OPUF) tool: A new method for valuing health states. *Wellcome Open Research*. 2022;7:14.

12. Devlin NJ, Shah KK, Mulhern BJ, Pantiri K, Hout B van. A new method for valuing health: Directly eliciting personal utility functions. *The European Journal of Health Economics*. 2019;20(2):257-270.

13. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of life research*. 2011;20(10):1727-1736.

14. Richardson JRJ, Mckie JR, Bariola EJ. Multiattribute utility instruments and their use. In: *Encylopedia of Health Economics, Volume 2*. Elsevier; 2014:341-357.

15. Devlin N, Pickard S, Busschbach J. The development of the EQ-5D-5L and its value sets. In: *Value Sets for EQ-5D-5L*. Springer, Cham; 2022:1-12.

16. Ludwig K, Ramos-Goñi JM, Oppe M, Kreimeier S, Greiner W. To what extent do patient preferences differ from general population preferences? *Value in Health*. 2021;24(9):1343-1349.

17. Leidl R, Reitmeir P. An experience-based value set for the EQ-5D-5L in germany. *Value in Health*. 2017;20(8):1150-1156.

18. Marsh K, IJzerman M, Thokala P, et al. Multiple criteria decision analysis for health care decision making—emerging good practices: Report 2 of the ISPOR MCDA emerging good practices task force. *Value in health*. 2016;19(2):125-137.

19. Keeney RL, Raiffa H, Rajala DW. Decisions with multiple objectives: Preferences and value trade-offs. *IEEE transactions on Systems, man, and cybernetics*. 1979;9(7):403-403.

20. Michel YA, Augestad LA, Barra M, Rand K. A norwegian 15D value algorithm: Proposing a new procedure to estimate 15D value algorithms. *Quality of Life Research*. 2019;28(5):1129-1143.

21. Palan S, Schitter C. Prolific. Ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*. 2018;17:22-27.

22. 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." *Social Science & Medicine*. 2020;246:112707.

23. R Core Team. *R: A Language and Enviro*nment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. https://www.R-project.org/.

24. Ludwig K, Graf von der Schulenburg J, Greiner W, et al. German value set for the EQ-5D-5L. Pharmacoeconomics. 2018;36(6):663-674.

25. Robinson A, Spencer AE, Pinto-Prades JL, Covey JA. Exploring differences between TTO and DCE in the valuation of health states. *Medical Decision Making*. 2017;37(3):273-284.

26. Brazier J, Ratcliffe J, Saloman J, Tsuchiya A. *Measuring and Valuing Health Benefits for Economic Evaluation*. OXFORD university press; 2017.

27. 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. *Medical care*. 2017;55(7):e51.

28. Poudel N, Fahim S, Qian J, Garza K, Chaiyakunapruk N, Ngorsuraches S. Methodological similarities and variations among EQ-5D-5L value set studies: A systematic review. *Journal of Medical Economics*. 2022;25(1):571-582.

29. Belton V, Stewart T. *Multiple Criteria Decision Analysis: An Integrated Approach*. Springer Science & Business Media; 2002.

30. Brazier J, Peasgood T, Mukuria C, et al. The EQ health and wellbeing: Overview of the development of a measure of health and wellbeing and key results. *Value in Health*. 2022.

APPENDIX

S1: Method for constructing a personal preference function from individual level responses - general model and a simple example

General Model

This section first describes the general model for constructing a personal preference function from individual level responses. Below, we then provide a simple example to illustrate the procedure.

The formula below shows how dimension level coefficients can be derived directly from the responses to the dimension weighting, level rating, and the anchoring task. In short, the level ratings and multiplied by the dimension respective weights, and the resulting values are rescaled to a 0-1 utility scale, where 0 is the best possible health state and 1 is the worst possible health state. The values are then multiplied with the anchoring factor, which either specifies the position of dead, on a scale between full health and the worst state (if the participants prefers the worst state over being dead), or the position of the worst state on a scale between full health and dead (if the participants prefers being dead over the worst state).

More formally, the model is defined as follows:

$$c_{ij} = \frac{\left(1 - \frac{l_{i,j}}{100}\right) * w_i}{\sum_{k=1}^{n} \left(w_k\right) * f(a)}$$

Where c_{ij} refers to the coefficient for jth level on dimension i. Accordingly, $l_{i,j}$ refers to the rating for level j on dimension i, w_i refers to the weight for dimension i, n refers to the number of dimensions, $f(a)$ refers to the anchoring factor. Note that $\sum_{k=1}^{n} \left(w_k\right)$ is simply the sum of all dimension weights, used to rescale the values to a 0-1 utility scale.

As stated above, the specification of the anchoring factor is dependant on whether the participant prefers the worst state over being dead, or being dead over the worst state. It can be defined as follows:

$$f(a) = \begin{cases} pits \succcurlyeq dead, \frac{1}{1 - v_{pits}} \\ pits \prec dead, 1 - v_{dead} \end{cases}$$

Where v_{dead} is the position of 'being dead' on 0-1 scale, anchored at full health (=1) and the worst health states (=0); and v_{pits} is the rating of position of the worst health state on 0-1 scale, anchored at full health (=1) and dead (=0).

To derive the utility for any given health states, first the sum of the respective dimension-level coefficients $c_{i,j}$ needs to be computed. Since the model is expressed in terms of disutilities, the resulting value needs to be subtracted from 1. The simple example below illustrates the method.

Example

Suppose a health-related quality of life measure with 3 dimensions (d_1, d_2, d_3), each of which with 3 ordered levels ($l_{11}, l_{12}, \ldots, l_{33}$).

Suppose that a participant provided the following responses:

- Dimension weightings: d_1 = 80; d_2 = 100; d_3 = 50;
- Level ratings: l_{11} = 100; l_{12} = 50; l_{13} = 0; l_{21} = 100; l_{22} = 40; l_{23} = 0; l_{31} = 100; l_{32} = 70; l_{33} = 0;
- Anchoring: the participants preferred the worst state ('333') over being dead. They located the worst state at 0.2 on the scale between full health (1) and dead (0). The anchoring factor is thus: $f(a) = \dfrac{1}{1 - 0.2} = 1.25$

Applying the formula above, we can derive the coefficients for each dimension and level:

$$c_{1,2} = \frac{(1 - \frac{50}{100}) * 80}{230 * 1.25} \approx 0.14, c_{1,3} = \frac{(1 - \frac{0}{100}) * 80}{230 * 1.25} \approx 0.28$$
$$c_{2,2} = \frac{(1 - \frac{40}{100}) * 100}{230 * 1.25} \approx 0.17, c_{2,3} = \frac{(1 - \frac{0}{100}) * 100}{230 * 1.25} \approx 0.35$$
$$c_{3,2} = \frac{(1 - \frac{70}{100}) * 50}{230 * 1.25} \approx 0.09, c_{3,3} = \frac{(1 - \frac{0}{100}) * 50}{230 * 1.25} \approx 0.18$$

N.B.: Note that level 1 coefficients are being equal to 0 and that $\sum_{k=1}^{n} (w_k) = 80 + 100 + 50 = 230$

These coefficients can then be used to compute the value of any given health state. The value of states 111, 123, 3333 are consequently:

$$v_{111} = c_{1,1} + c_{2,1} + c_{3,1} = 0 + 0 + 0 = \underline{0}$$
$$v_{123} = c_{1,1} + c_{2,2} + c_{3,3} = 0 + 0.17 + 0.18 = \underline{0.35}$$
$$v_{333} = c_{1,3} + c_{2,3} + c_{3,3} = 0.28 + 0.35 + 0.18 = \underline{0.81}$$

Table S2: Sample characteristics: reported health conditions

Disease	frequency
Rheumatoid arthritis	72 (59.02%)
Psoriatic arthritis	21 (17.21%)
Fibromyalgia	21 (17.21%)
Chronic pain syndrome	19 (15.57%)
Ankylosing spondylitis	15 (12.3%)
osteoporosis	11 (9.02%)
Sjogren’s syndrome	9 (7.38%)
Lupus erythematosus	8 (6.56%)
Gout	3 (2.46%)
Polymyalgiarheumatica/RZA	3 (2.46%)
Scleroderma	1 (0.82%)
Other degenerative	6 (4.92%)
Other endocrine	3 (2.46%)
Other inflammatory	8 (6.56%)
Other vasculitis	6 (4.92%)
Other	4 (3.28%)
prefer not to say	1 (0.82%)

Note: 64 participants reported more than one health condition.

Table S3: Mean (SD) level ratings

	MO	SC	UA	PD	AD
lvl_1	100 (0)	100 (0)	100 (0)	100 (0)	100 (0)
lvl_2	75.5 (17.1)	74.4 (19)	75.8 (15.2)	77.2 (17.5)	74.5 (19.3)
lvl_3	52 (14.6)	50.5 (17)	49.4 (14)	52.7 (14.8)	49.3 (17.3)
lvl_4	23.6 (17.4)	22 (18.3)	20.5 (17.5)	22 (19.8)	21.4 (15.3)
lvl_5	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

DISCUSSION

‘I could tell you my adventures – beginning from this morning,’ said Alice a little timidly: ‘but ‘it’s no use going back to yesterday, because I was a different person then.’

– Lewis Caroll, Alice in Wonderland

Chapter 9

Discussion and Conclusion

Overview

In this thesis, I have sought to contribute to both the understanding, as well as the practice, of health valuation in the context of HTA, and more specifically to the way in which individual health preferences are elicited, aggregated, and used to create value sets. The thesis has two parts. In Part I (Chapters 2-5), I used an applied ethics approach to explore normative issues related to the aggregation of individual preferences and the ethical implications of currently used methods were evaluated from different theoretical perspectives. Part II (Chapters 6-8) reported on the development of the OPUF approach, a new method for valuing health, which allows for the construction of preferences in small samples and even on the individual level. The underlying theme, connecting both parts, is the challenge of determining the value of health in a plural society, consisting of many individuals, with unique preferences, values, and perspectives. By exploring related normative and methodological issues, I have sought to call into question current standard practices, and derive starting points for a more informed, considered, and pluralistic approach.

This final ninth Chapter is structured as follows: First, I summarise the main contributions to the field of health valuation and the impact of the research reported in this thesis. I then reflect on the limitations of this work before, finally, offering potential avenues for future research and a conclusion.

Key contributions and impact

This thesis makes a number of original contributions to the field of health valuation. In part I, I developed four original ideas, outlining new perspectives on the relationship between individual preferences and social value sets.

Part I: normative issues

In Chapter 2 (*Social tariffs and democratic choice*), I explored how social value sets can be understood as a major instrument of democratic participation, and how this would have implications for both the method used to aggregate individual preferences, and the set of individuals whose preferences should count. This provides a novel perspective on what value sets are, or could be, i.e. thinking through the idea of incorporating 'the will of the people' into health economic evaluations. In addition, I point out interesting paradoxes that can occur when using the median, instead of the average, to aggregate preferences, including the finding that the median does not necessarily represent the preferences of a majority. This is certainly not a new insight, but illustrating the implications in the context of health state preferences aggregation provides an original contribution to the field.

In Chapter 3 (F*air interpersonal utility comparison*), I developed a relative utilitarian approach to the aggregation of individual preferences as an alternative to the current practice of simply taking the arithmetic average. To the best of my knowledge, this is the first time such an approach has been proposed in the context of health economics. More generally, my research also sheds light on certain peculiar characteristics of the utility scale in a QALY framework, namely that the scale is anchored between two potentially incomparable states: full health, which has a time and a health-related quality of life dimension, and being dead, which most people understand as a timeless state without any qualities whatsoever. Locating other health states on this scale involves complex evaluations of both health-related quality of life and life time. For intrapersonal utility comparisons, i.e. within a single individual, this may not be a problem, but when comparing or aggregating preferences of different individuals, it seems questionable whether utilities are (directly) interpersonally comparable.

The scaling between full health and dead is investigated further in Chapter 4 (S*etting dead to zero*?), which adds to the knowledge base by critically reviewing the arguments that have been used to justify the practice of setting dead to a utility value of zero. It is shown that setting dead to a different value may well be permissible. Several authors have already proposed anchoring the utility scale on a different state (e.g. Sampson et al., 2020), but, as far as I am aware, this is the first time the question has been engaged with on an quasi-axiomatic basis, and it refutes previously made claims that setting dead to a different value would violate rationality rules or measurement theory (Roudijk et al., 2018).

Finally, Chapter 5 (*The QALY is ableist*) contributes a new argument to the long-lasting debate on whether or not the QALY discriminates against people in poor health and those living with disability (Ubel et al., 2000). Previous ethical objections primarily focused on relative prioritisation, i.e. is a person with poor HRQoL more or less deserving of health care (e.g. Harris, 1995)? Proponents of the QALY could then argue that, since resources are limited, some form of discrimination is unavoidable, and that HRQoL seems to be a reasonable allocation criterion (Whitehurst and Engel, 2018). As I have shown, states worse than dead add another layer to this debate, because not only do they involve judgments about relative prioritisation, but about the absolute value of people's lives. Assigning negative QALYs to additional life years leads to QALY gains from people (who would prefer to live longer) dying earlier. This result seems utterly absurd, and there does not seem to be any justification for the implied discriminatory value judgement. It is striking that this point has not been made before, and I hope this work will contribute to a more informed debate on the ethical implications of current practices in health valuation and decision analytical modelling. I was recently invited to give a talk to the *Evidence Review Group* at the University of Exeter, who worked on a technology appraisal in which states worse than dead posed a challenge. Anecdotally, it was my impression that the group was interested in the arguments I presented, and that it might have affected the way in which the appraisal was conducted.

Potentially more important than the individual issues addressed in the four chapters may be the overall realisation that the aggregation of individual preferences into a social value set is a complex process, requiring careful consideration of both, methodological and normative issues. Throughout Part I of this thesis, I sought to make explicit the strong normative assumptions underlying current practices, and to stress the contingency of methodological choices. This applies not just to preference scaling and aggregation but more generally to other aspects of health valuation, including questions about what should be valued, whose values should count, and how values should be elicited. I would even argue that when the outcome of a research project – such as a value set – is supposed to inform health policy decision making, possibly affecting the health of millions of people, even seemingly mundane research methods, such as taking measurements, recruiting respondents for a survey, or applying statistical models, may suddenly become highly charged, value-laden activities. A more nuanced understanding of the methodological and normative choices underlying health valuation seems necessary, in order to ensure decisions about health care are based on evidence that is both methodologically rigorous *and* ethically sound.

Part II: practical tools

In the second part of the thesis, I sought to provide a more practical solution to the problem of how to account for the heterogeneity of preferences between individuals. Chapters 6-8 report on the development and piloting of the OPUF tool, a new method eliciting health state preferences, based on compositional preference elicitation techniques.

The initial survey development, using an iterative design approach, is documented in Chapter 6. The OPUF survey was then first tested on 50 respondents. Subsequently, in Chapter 7, OPUF was applied to derive an alternative EQ-5D-5L value set in a larger sample of 1,000 respondents, that were broadly representative of the UK general population. The collected data allowed me to demonstrate that it is feasible to use OPUF to construct and compare subgroup preferences in a level of granularity that is not possible with time trade-off, or other traditional methods. Finally, in Chapter 8, I showed that OPUF can also be used to construct an EQ-5D-5L value set in a relatively small group of patients (n=122) with rheumatic diseases. While the results of the individual studies are discussed in detail in the respective chapters, the overall finding is that OPUF provides a promising alternative approach, which has been demonstrated to generate plausible value sets (for the EQ-5D-5L) with good precision, even in small samples.

Notwithstanding the developmental work described in this thesis, some might argue that OPUF is not a particularly original contribution, because it heavily builds on previous work: compositional preference elicitation methods have been used since at least the 1970s (Keeney et al., 1979), and even the specific combination of compositional methods used for OPUF is not new. It was first developed by Devlin et al. (2019) in a study from 2015. However, their approach was designed for face-to-face interviews, with an emphasis on deliberative and reflective elements. My primary contribution to this work stream has been to take this rather elaborate approach and translate it into a short, easy-to-use online survey. This came at a cost: deliberative elements, such as the ability to reflect on and revise responses, or to tap into personal motivations, had to be omitted. However, in doing so, the tool was significantly easier (and cheaper) to apply at scale, while retaining considerable benefits of compositional preference elicitation.

A distinguishing feature of OPUF, in comparison to other elicitation methods, may, in fact, be its design as a 'web-native' method. In the development process, particular attention was paid to the 'user experience', i.e. much effort went into making the survey as user-friendly and aesthetically appealing as possible, so that it is intuitive to use and easy to navigate, making it accessible to a broad audience, including patients and members of the general public. Throughout the three studies described in Part II of this thesis, the OPUF surveys were generally well received, as evidenced by the positive qualitative feedback many of the respondents provided.

Ease-of-use was an important design criterion, not only for the respondent-facing survey part, but also for the researcher-facing 'backend'. The first version of OPUF, used for the research reported in Chapters 6 and 7, was developed in R, making use of the 'shiny' package to create the user interface (Chang et al., 2017). The EQ-5D-5L instrument was hard-coded into the survey, making it difficult to reuse and adapt the tool for other research project. The current version uses a modern web development framework consisting of Vue.js (vue 3) for the frontend (client-side), Node.js as backend (server-side), and MongoDB as database. It now consists of several modules (e.g. ranking, demographic questions, swing weighting, etc.), which can be configured and easily combined with each other to create customised surveys. The tasks provide merely a – so called – 'skeleton'. This means, all page content (e.g. instructions, questions, labels, etc) is retrieved from the database and filled-in programmatically. This enables researchers to create bespoke OPUF surveys for any standard measure of health-related quality of life (such as EQ-5D, SF-6D, etc) within minutes, allowing fast testing, piloting, and implementation. The platform also contains an ‘admin dashboard’, which displays survey status (e.g. number of completes, duration, etc.), and automates some common analytical steps.

The entire OPUF source code is released under a permissive open source licence, so that it can be checked, copied, re-used and adapted by anyone. Although the current lack of documentation creates a significant barrier to entry, I am confident that, once this is in place, OPUF can become a widely used health state preferences elicitation platform, to which other developers can also contribute.

So far, the reception of the OPUF method by the research community has been positive. I have been given the opportunity to present my work at over 20 conferences, seminars, and workshops, had many interesting discussions, and several researchers contacted me to express their interest in applying the method for their own research. Collaborative projects in a range of different applications have been initiated, which I will support by providing technical assistance and methodological advice. Three examples are listed below.

1. **OPUF for the MobQoL-7D:** the MobQoL-7D is a new instrument to capture mobility-related quality of life (i.e. it is a condition-specific measure), recently developed by Bray et al. (2022). In September 2022, an OPUF study was successfully conducted to derive value sets for the instrument based on the preferences of both members of the UK general public (n= 504) and individuals with mobility problems (n=368). The results have been submitted to ‘Disability and Rehabilitation’ and are currently under review (‘*De-*

veloping preference-based value sets for the MobQoL-7D: Practical application of the Online Elicitation of Personal Utility Functions (OPUF) tool').

2. **OPFU for the EQ-HWB:** the EQ-HWB-S is a recently developed generic measure that goes beyond health to include aspects of wellbeing (Brazier et al., 2022; Peasgood et al. 2022). The measure has nine dimensions, which makes it difficult to derive a value set using conventional time trade-off or discrete choice experiment, i.e. the number of questions and/or participants would need to be quite large to derive reliable estimates. Pilot studies, funded by the EuroQol Group, are currently underway in the UK and in Germany to elicit EQ-HWB-S preferences from members of the general public as well as from patients (rheumatic diseases and diabetes mellitus). The study includes a test-retest part, to assess the reliability of OPUF. At the time of writing, the data collection for the re-test has just been completed.

3. **OPUF for the EQ-5D-5L in South Africa:** at the moment, there is no EQ-5D-5L value set for South Africa. Estimating a value set using the 'official' EQ-VT protocol entails an involved and expensive process, which currently does not seem to have enough policy support (Al Shabasy et al., 2021). Researchers from the University of the Witwatersrand in Johannesburg are currently using OPUF to derive an experimental value set, with a particular emphasis on exploring preference heterogeneity with respect to socio-economic status and 'race'. Preliminary discussions concluded that the use as a stand-alone online survey was not feasible, and thus, the survey was adapted for computer-assisted personal interviewing. At the time of writing, the data collection is ongoing.

Further OPUF valuation studies currently underway or being planned include applications to value the EQ-5D-Y in the UK (Wille et al., 2010), the CHU9D in Australia (Stevens, 2009), the FACT-8D and the QLU-C10D in China (King el al., 2016), the EQ-5D-5L in Hungary, and the SanQoL in Mozambique (Ross et al., 2022). Moreover, the use of OPUF to create dashboards for patients with dementia and their family and care givers, as well as a decision-aid in clinical practice, are being explored.

Overall, OPUF provides a versatile and pragmatic tool to elicit values in a range of different contexts and settings. It is easy to use (for the respondents) and easy to implement (for the researcher), while still providing detailed, granular data on preferences of individuals and small groups. It is my hope that OPUF will be taken up by researchers and practitioners, and that it will contribute to making context-specific preference information more readily available to decision makers.

Limitations

The research program reported in this thesis has a number of limitations that deserve mentioning. Study specific limitations have been discussed in the respective chapters, but there are some limitations that relate to the approach taken in this thesis more generally.

First, this thesis has covered a wide range of issues, which, while connected by a general theme, are still quite distinct from each other. The four chapters in part I are essentially independent works; they could stand on their own, and are not even mutually reinforcing. My aim was merely to exemplify the range of ethical issues raised by preference aggregation methods.

This also applies to parts I and II of the thesis, which focus on normative issues and on the provision of a practical solution, respectively. The two parts are connected by the underlying theme of accounting for individual preferences in health valuation in a plural society, yet there is a risk of a 'disconnect' between the two parts. More in-depth work on each topic, and/or a clear demonstration of the implications of the findings from Part I, using the tools developed in Part II, would have been desirable. However, this would have required a whole programme of work, well beyond the limited scope and resources of my PhD thesis, which, I hope, provides useful starting points for further research.

Secondly, the empirical studies reported here are pilot studies, which means the results should be interpreted with caution. More psychometric testing is required to determine the reliability and validity of the OPUF method. More work is also needed to explore the implications of the OPUF method for the construction of value sets, and to demonstrate the usefulness of the OPUF in a range of different contexts, and to determine whether, and if so, where, the new method can best complement (or even replace) existing methods.

Thirdly, health valuation, and the resulting value sets, are but one component of health economic evaluations, which, in turn, are but one component within HTA. For none of the work described in this thesis have I demonstrated whether, and if so how, alternative approaches to valuing health, in general, and to preference aggregation, in particular, would make any difference in the broader context of HTA: is the decision, whether or not to reimburse a given health technology (and potentially its price) affected by the way in which preferences are aggregated? Would different decisions be made, if, instead of the current reference case, patient preferences, elicited using OPUF, were used? Further research is needed to answer these questions, and demonstrate that the intellectual exercises displayed in this thesis have any relevance for the real world. However, the informational requirement for this would be substantial: one would need to compare the outcomes of a ‘standard’ economic model with the outcomes of a counter-factual model that uses the alternative health valuation method. Detailed information on the health states of the underlying population would be required to estimate QALY weights. In order to generate a robust evidence base that does (or does not) show that the alternative method propagates through the model and affects cost-effectiveness estimates in a relevant way, a single case study is unlikely to be sufficient, and so one would need either to investigate many different economic models, or devise a compelling simulation study. To complicate things further, one should account for the fact that NICE’s HTA process is not deterministic. Many factors other than the incremental cost-effectiveness are considered. Some change in the incremental cost-per-QALY from just under to just above the threshold is unlikely to affect the decision in the real world, which makes demonstrating a practical impact of alternative health valuation techniques even more difficult.

Finally, I should acknowledge that, although I have repeatedly criticised the lack of coherent normative framework for the valuation of health in this thesis, I have not made any attempt to develop one of my own. In fact, in Part I, I adopted different, potentially incompatible, perspectives (democratic, relative utilitarian, and liberal) from which to examine the status quo and to develop alternative approaches. The OPUF approach developed in Part II is also not based on any particular normative framework. Rooted in multi-criteria decision analysis, it provides a distinctly pragmatic method for the elicitation of health state preferences, which is agnostic to the underlying value system. This may seem somewhat contradictory, but it is in line with my overall aim to explore the complexity of health valuation.

On a personal note, I may add, that it is also in line with my own view on these issues. When I started my PhD, I (perhaps naively) assumed that I would finish with a well-formed idea for how health should be valued and how individual preferences should be conceptualised and aggregated. Yet, the opposite is the case. Any sense of certainty I may have had in the beginning has given way to an ever-deepening appreciation of the (rewarding) complexity of the task of determining the value of health in a plural society.

Further research

As indicated above, the work presented in this thesis provides the foundation for further research.

First, the practical implications of the arguments presented in Part I of this thesis should be explored. This could include, for example, using granular data on individual preferences, elicited through the OPUF method, to compare the results of a 'standard' average value set with alternative approaches.

Secondly, a comprehensive research program could be devised to validate the OPUF approach and explore its potential applications. Some work is already ongoing to 1) evaluate OPUF in various different settings; 2) assess the test-retest reliability; and 3) qualitatively explore respondents' thought processes and experiences. Further validation studies should also be conducted. However, I would like to add two caveats: before trying to investigate criterion validity, and comparing OPUF to other methods, one should in advance determine how results will be interpreted. It is known, for example, that different preference elicitation methods – and even variations of the same method – can produce different results (Brazier et al., 2017, p. 72). Since there is no gold standard, neither agreement nor disagreement between the results of different methods can be taken as an indication of validity, and it is not entirely clear what can be learned from these studies. Furthermore, OPUF does not necessarily need to be one standardised method, but could provide a modular platform for which different components could be adjusted, depending on the context.

Thirdly, one particularly interesting further research avenue for OPUF would be to explore its use as a decision aid for patients. Many respondents who completed an OPUF study reported that they found the survey thought-provoking and interesting. Given the increasing use of patient-reported outcome measures in clinical practice, OPUF could provide a useful tool to summarise the often complex information that these measures convey, and more generally to help patients better understand their own preferences priorities and make informed decisions about their care.

A key challenge to the adoption of OPUF remains the accessibility of the method to other researchers. Although the source code is provided under a permissive open source licence, comprehensive documentation of the existing source code is still needed. Yet this is unlikely to be sufficient. Most health economists do not have web development experience. To make the method more accessible, I have recently started exploring the possibility of creating a web-based, graphical user interface for OPUF, i.e. a ‘survey builder tool’, which would allow researchers to create customised surveys without any coding. However, this will require a significant amount of development work and some sustainable source of funding to ensure that the platform is maintained and updated.

Finally and most importantly, this thesis shows the need for more research on the normative foundation of health valuation. Ideally, this should involve a broad and sustained participation of all relevant stakeholders (members of the general public, patients, care givers, health care professionals, decision makers, etc.) in an open, participatory process. As already argued on multiple occasions, I do not think one can expect to derive a single, consistent normative framework from this process. Rather, the aim should be to identify core values and principles that can then guide further methodological research in and the practice of health valuation.

Conclusion

This thesis seeks to make an original contribution to both the understanding as well as the practice of health valuation. It provides a critical examination of the normative issues arising from the aggregation of individual preferences into a value set, and it develops a new, pragmatic preference elicitation method – Online Elicitation of Personal Utility Functions (OPUF) – which allows for the construction of preferences in small samples and even on the individual level. In the absence of consensus on the 'correct' way to value health, and no reasonable expectation that such a consensus will soon, or ever, be reached, a more pluralistic and transparent approach is needed. This thesis provides some potential starting points such an approach will require.

REFERENCES

Al Shabasy S, Abbassi M, Farid S. EQ-VT protocol: one-size-fits-all? Challenges and innovative adaptations used in Egypt: a cross-sectional study. *BMJ Open.* 2021 Dec 1;11(12): e051727.

Bray N, Tudor Edwards R. Preference-based measurement of mobility-related quality of life: developing the MobQoL-7D health state classification system. *Disability and Rehabilitation.* 2022 Jun 5;44(12): 2915-29.

Brazier J, Peasgood T, Mukuria C, Marten O, Kreimeier S, Luo N, Mulhern B, Pickard AS, Augustovski F, Greiner W, Engel L. The EQ Health and Wellbeing: overview of the development of a measure of health and wellbeing and key results. *Value in Health.* 2022 Mar 8.

Brazier J, Ratcliffe J, Saloman J, Tsuchiya A. *Measuring and Valuing Health Benefits for Economic Evaluation.* Oxford: Oxford University Press; 2017.

Chang W, Cheng J, Allaire J, Xie Y, McPherson J. Shiny: web application framework for R. R package version. 2017;1(5):2017.

Devlin NJ, Shah KK, Mulhern BJ, Pantiri K, van Hout B. A new method for valuing health: directly eliciting personal utility functions. *The European Journal of Health Economics.* 2019 Mar 15;20: 257-70.

Harris J. Double jeopardy and the veil of ignorance--a reply. *Journal of Medical Ethics.* 1995 Jun 1;21(3): 151-7.

Keeney RL, Raiffa H, Rajala DW. Decisions with multiple objectives: Preferences and value trade-offs. *IEEE transactions on Systems, Man, and Cybernetics.* 1979;9(7): 403-403.

King MT, Norman R, Viney R, Costa D, Brazier J, Cella D, Gamper E, Kemmler G, McTaggart-Cowan H, Peacock S, Pickard AS. Two new cancer-specific multi-attribute utility instruments: EORTC QLU-C10D and FACT-8D. *Value in Health.* 2016 Nov 1;19(7): A807.

Peasgood T, Mukuria C, Brazier J, Marten O, Kreimeier S, Luo N, Mulhern B, Greiner W, Pickard AS, Augustovski F, Engel L. Developing a new generic health and wellbeing measure: Psychometric survey results for the EQ-HWB. *Value in Health.* 2022 Apr 1;25(4): 525-33.

Ross I, Greco G, Opondo C, Adriano Z, Nala R, Brown J, Dreibelbis R, Cumming O. Measuring and valuing broader impacts in public health: Development of a sanitation-related quality of life instrument in Maputo, Mozambique. *Health Economics.* 2022 Mar;31(3): 466-80.

Roudijk B, Donders AR, Stalmeier PF. Setting dead at zero: applying scale properties to the QALY model. *Medical Decision Making.* 2018 Aug;38(6): 627-34.

Sampson C, Parkin D, Devlin N. Drop dead: Is Anchoring at 'Dead' a Theoretical Requirement in Health State Valuation? *OHE Research Paper.* 2020 Nov.

Stevens K. Developing a descriptive system for a new preference-based measure of health related quality of life for children. *Quality of Life Research.* 2009; 18: 1105–13

Ubel PA, Nord E, Gold M, Menzel P, Prades JL, Richardson J. Improving value measurement in cost-effectiveness analysis. *Medical care.* 2000 Sep 1: 892-901.

Whitehurst DG, Engel L. Disability discrimination and misdirected criticism of the quality-adjusted life year framework. *Journal of Medical Ethics.* 2018 Nov 1;44(11): 793-5.

Wille N, Badia X, Bonsel G, Burström K, Cavrini G, Devlin N, Egmar AC, Greiner W, Gusi N, Herdman M, Jelsma J. Development of the EQ-5D-Y: a child-friendly version of the EQ-5D. *Quality of life research.* 2010 Aug;19(6): 875-86.