
Article:
Spencer, A., Rivero-Arias, O., Wong, R. orcid.org/0000-0002-4536-4794 et al. (6 more authors) (2022) The QALY at 50 : one story many voices. Social Science & Medicine, 296. 114653. ISSN 0277-9536

https://doi.org/10.1016/j.socscimed.2021.114653

The QALY at 50: one story many voices

Spencer A[1] PhD, Rivero-Arias O[2] DPhil, Wong R[3] PhD, Tsuchiya A[4] PhD, Bleichrodt H[5] PhD, Edward RT[6] DPhil, Norman R[7] PhD, Lloyd A[8] PhD, Clarke P[2] PhD

Affiliations
[1] University of Exeter, UK
[2] Nuffield Department of Population Health, University of Oxford, UK
[3] Ruth Wong, PhD, University of Sheffield, UK
[4] Aki Tsuchiya, PhD, University of Sheffield, UK
[5] Erasmus School of Economics, Erasmus University, The Netherlands
[6] Centre for Health Economics and Medicines Evaluation, Bangor University, UK
[7] School of Public Health, Curtin University, Australia
[8] Acaster Lloyd Consulting Ltd., UK

Address correspondence to: Professor Philip Clarke, Health Economics Research Centre, Nuffield Department of Population Health, University of Oxford, UK. Email: philip.clarke@ndph.ox.ac.uk

Running title: The QALY at 50

Funding information: No funding was received for the preparation of this manuscript.

Competing interest statement: Authors declare no conflict of interest.

Author contributions
Editorial responsibilities: Rivero-Arias O, Spencer A, Tsuchiya A, Clarke P

Invited contributions to future research agenda section: Bleichrodt H, Edwards RT, Norman R, Lloyd A, Rivero-Arias O, Tsuchiya A, Spencer A, Clarke P

Information specialist: Wong R.

CRediT author statement (https://www.elsevier.com/authors/policies-and-guidelines/credit-author-statement)
Spencer A: Conceptualization, Methodology, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration; **Rivero-Arias O:** Conceptualization, Methodology, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Project administration; **Wong R:** Formal analysis, Writing - Original Draft, Writing - Review & Editing, Visualization; **Tsuchiya A:** Conceptualization, Methodology, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Project administration; **Bleichrodt H:** Writing - Original Draft, Writing - Review & Editing; **Edward RT:** Writing - Original Draft, Writing - Review & Editing; **Norman R:** Writing - Original Draft, Writing - Review & Editing; **Lloyd A:** Writing - Original Draft, Writing - Review & Editing; **Clarke P:** Conceptualization, Methodology, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Project administration

Abstract

Research on quality adjusted life year (QALY) has been underway for just over 50 years, which seems like a suitable milestone to review its history. The purpose of this study is to provide a historical overview of why the QALY was developed, the key theoretical work undertaken by Torrance, Bush and Fanshel and how two seminal papers shaped its subsequent development.

Moving the QALY forward – there are several historical and reflective exercises. The historical interplay between politics, policy and the challenges facing the NHS in formulating the QALY concept in the UK has been explored in some depth already, whilst the conceptualisation and development of the methodological framework is relatively underexplored. We address this gap by viewing the QALY through the lens of the methodological debates, reflecting upon two key papers underpinning the QALY methodology and how these methods have been developed over time. In part the changes in technology e.g. Google Scholar, and the availability of tools to search for early uses of the QALY allow us to better understand the historical context in which the theoretical development of the QALY has taken place.

Here we celebrate two seminal papers that shaped early QALY development. The first section provides a history of these papers, summaries their contributions and explores the uptake of these papers over time. The second section reviews the methodological debates that have surrounded the QALY over the last 50 years and looks at how the QALY has moved to address these challenges. The third section presents the voices of diverse commentators representing the field of health economics who have contributed to the subsequent development of the QALY in both theoretical and empirical capacities and captures their thoughts about future research and policy use of QALYS.

Key words
History of economic thought; Health outcomes; Cost-effectiveness analysis; Quality adjusted life years (QALY)

Word count (including manuscript text and references): 10,722

Acknowledgments
The article stems from the presentations given at the 50th anniversary of the Quality Adjusted Life Year at the UK Health Economists' Study Group Meeting webinar in July 2020. We are grateful to the Oxford Health Economics Research Centre for organising the webinar; to Helen Dakin and Mara Violato for chairing the webinar; and to Bruce Hollingsworth and Paula Lorgelly for their introductions and overviews. We are also grateful to the Health Economics Group at Exeter for their feedback on a presentation on the methodological debates that influenced the QALY in May 2020, to the administrative support by Leala Watson and Jenny Shilton Osborne. Anne Spencer's research is supported by the National Institute for Health Research Applied Research Collaboration South West Peninsula.

Finally, a special thanks to George Torrance, for his views on the origins of the QALY and for attending the HESG webinar on the 50th anniversary of the Quality Adjusted Life Year. His support and encouragement to develop the presentations into a paper have been gratefully received.

The views expressed in this publication are those of the author(s) and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care.

The QALY at 50: one story many voices

1. Introduction

The analytical methods to capture quality of life represent a key element of the health economists' tool kit for economic evaluation and have been widely adopted in many countries to assess, often relative to a funding threshold (Grosse, 2008), new technologies and public health interventions. The methods used to elicit preferences for health-related quality of life were first proposed in two seminal papers, both published in 1970 (Fanshel & Bush, 1970; Torrance, 1970), which proposed methods such as the time trade-off (TTO), standard gamble (SG) and person trade-off (PTO). All these methods still underlie the measurement of what has become known as the Quality-Adjusted Life Year (QALY).

With the QALY as a concept now entering its sixth decade, it is useful to reflect on the work of Bush, Fanshel and Torrance: first, to understand their contribution in a historical context (i.e. what motivated the development of the QALY); secondly, to understand the impact of their research on shaping subsequent developments (both theoretical and applied); and thirdly, to use history as a lens to speculate on what further developments are required to make the QALY fit for purpose for another fifty years. Viewing the QALY through the lens of these methodological debates complements an existing exploration of the interplay between politics, policy and the challenges operationalizing the QALY concept in countries such as the UK (MacKillop & Sheard, 2018).

This article is divided into four sections. The first provides a historical background including an overview of the contributions of both Torrance (Torrance, 1970; Torrance et al., 1972) and Fanshel and Bush (Fanshel & Bush, 1970). This section was informed by relevant literature including rarely cited studies that were identified from keyword searches in databases such as Google Scholar. It has also draws on material from an interview with Torrance conducted in May 2020 (an edited transcript is included in supplementary material 1). The second section provides a bibliometric analysis of the impact of these papers on subsequent research. The third involves contributions of several health economists and outcome researchers on potential avenues for future QALY research. The last section provides some final reflections and conclusions.

2. Historical background

2.1 Origins of the QALY

While the use of the ratio of costs to outcomes to evaluate health care dates back to the late 19th Century (Anon, 1899), it was not widely applied until the 1960s. Cost-effectiveness analysis was developed a decade earlier in research for the US military (Enthoven, 2019), who used this concept to improve the efficiency of weapon systems (Foster & Hoeber, 1955) and is also the reason behind the phrase “bang for your buck”.

Its initial use for health-related applications in the early 1960s was in the context of the development of technologies to save the maximum number of lives in the event of nuclear war. Following a large appropriation by the US Congress, various types of shelter design were evaluated and “cost-effectiveness under a wide range of attack conditions was determined by dividing the costs of shelters by the number of lives presumptively saved by the shelters” (Walsh, 1963). While such analyses were sophisticated, employing both computer simulation and reporting of costs and outcomes on a plane (Guess, 1965), they did not attempt to move beyond a metric of lives saved to consider the quality of life of survivors.

In parallel with the development of the methods for cost-effectiveness, there was recognition by public health researchers of the need to evaluate interventions and systems of care. For example, Mushkin’s Towards a Definition of Health Economics tentatively defined health economics as a field of inquiry focusing on “optimum use of resources” and “to appraise the efficiency of the organization of health services” (Mushkin, 1958). Feldstein also expressed similar sentiments in a paper in the medical journal the Lancet in regard to improving efficiency within England’s National Health Service (Feldstein, 1963), published a few months prior to Arrow’s seminal article on the welfare economics of medical care (Arrow, 1963). Early health economic evaluations often focused on potential benefits of eliminating a disease, and tended to quantify benefits in terms of reductions in health care costs and gains in earning as the basis of a cost-benefit analysis (Klarman, 1982).

The use of economic evaluation was given great impetus in 1965 by President Johnson who wanted other government departments to adopt methods developed in the Defence Department (Novick, 1968). The US Department of Health quickly responded by producing an overview of both cost-benefit and cost-effectiveness analyses, recommending their use by “third-party payers and health agencies of all kinds” and “suggesting that these tools can help to maximize the value of the services which are provided” (Crystal & Brewster, 1966). They also sought to engage with academics by expanding health economics research in the US.

Packer, who was based at the Research Triangle Institute, was quick to identify the key problems when applying cost-effectiveness to health (Packer, 1968). For example, that the outcomes of health care are inherently multidimensional. While mathematicians were developing methods to generalize cost-effectiveness methods to optimize across multiple outcomes (Zabronsky, 1967), Packer’s focus was on ways to map different health outcomes in a unidimensional index. Chiang had proposed such an approach to capture both mortality and morbidity (Chiang, 1965), but Packer concluded that in “no case has the development progressed to the point of producing an operationally useful index” (Packer, 1968). However, while nominating potential ways forward, including the application of cardinal utility theory and the scope for using patient choice to obtain revealed preferences, he did not pursue these ideas further to produce an empirical health measure.

In parallel with this theoretical work is Herbert Klarman’s work with the US Budget office looking at renal dialysis and renal replacement. Klarman was one of the few academics specializing in health economics and had already contributed the first heath economics textbook (Klarman, 1965). His application of cost-effectiveness of the treatment of chronic renal disease published in late 1968 represents a seminal contribution to the method to evaluate health care. It not only includes detailed estimates of the potential life years gained from renal transplants, but also quality-adjusted these outcomes, as ”fraction of each life-year gained” (Klarman, 1965). While Klarman assumed that patients with renal replacement had around 25% of additional quality-adjusted life years compared to those on dialysis (assumed to be equivalent to full health), there was no empirical basis for this estimate in his study.

The stage was now set for two independent contributions that would define the QALY, as we know it today.

1. George Torrance's research on the QALY is documented in a 336 page report entitled *A generalized cost-effectiveness model for the evaluation of health programs* (Torrance, 1970). The report was commissioned by the Ontario Department of Health and formed his PhD thesis supervised by David Sackett, then Chairman of the Department of Clinical Epidemiology and Biostatistics at McMaster University. Here we focus only on Torrance's contributions to the QALY and these include expositions presented in a subsequent paper (Torrance et al., 1972) of the von Neumann-Morgenstern SG to measure the value of health states and a new approach, which Torrance called the Time Trade-Off, as an alternative method to measure the value of health states. In addition to a theoretical exposition, Torrance conducted a number of empirical studies to elicit values for several health states, notably for kidney transplant and dialysis, which built on the work of Klarman.

2. Sol Fanshel (who came from an engineering background) and James W. Bush (who had a medical background specializing in public health) were based at New York University, and were also commissioned by the New York State Health Planning Commission to come up with methods for evaluating heath care programs (Fanshel & Bush, 1970). Drawing on Thurston's early work from psychology (Thurstone, 1928) Fanshel and Bush argued that "paired comparisons… [can be used] to obtain the values (or weights) to be assigned to social phenomena". Fanshel and Bush proposed two types of trade-offs: one involving comparisons of two populations of different sizes experiencing different levels of health, which later on becomes PTO, and the second at an individual level where functional well-being on a zero-to-one scale is traded off against length of life. They illustrated how these methods could be used to characterize a series of health states ranging from mild "dissatisfaction" to a "coma" state that was equivalent to being dead. Their health status index was in an illustrative evaluation of a program involving screening children for tuberculosis.

While the initial methodological development of the QALY was largely undertaken in North America, the need for an index was also recognised elsewhere. There was a proposal by Grogono and Woodgage in the Lancet (Grogono & Woodgate, 1971). At around the same time Rosser and Watts proposed a health index to measure hospital output (Rosser & Watts, 1972). While these early studies did not use the newly developed methods such as the TTO, the later adoption and use of the QALY in the UK, Europe and elsewhere was critical to its success (MacKillop & Sheard, 2018).

In the next section, we show the cumulative citations to the Torrance, Bush and Fanshel papers over time (Figure 1). The adoption of the QALY started to gather pace in the 80s with articles such as Williams 1985 and Williams 2005, which promoted the concept to a wider medical audience (Williams, 1985, 2005). Similarly, the development of multi-attribute utility instruments such as the EQ-5D aided its adoption by providing a practical way to measure health-related quality of life routinely (EuroQol Group, 1990).

Another important driver, which Torrance highlighted when interviewed in 2020, was the development of formal processes involving economic evaluation adopted by many governments for the reimbursement of health care. This started with Australia's requirement for the economic evaluation of new pharmaceuticals as far back as 1990 (Evans et al., 1990) and subsequently implemented in other countries including Canada and the UK. Such evaluations are subject to guidelines, which have often encouraged the use of QALYs (Drummond, 1991).

As an obituary for James Bush written in the late 1980's noted, "he was often frustrated at the slow rate at which his ideas were absorbed and embraced by others. It is ironic that, in the short time since his death… [c]olleagues in Great Britain and Europe are also expanding many of the lines of inquiry that he had originally proposed" (Anderson et al., 1987).

3. Impact of Torrance (1970), Torrance et al (1972) and Fanshel and Bush (1970) on subsequent literature

3.1. Co-citation evaluation

3.1.1 Methods

We used co-citation analysis to assess the frequency with which these identified papers and their references were "cited together by the later literature" (Small, 1973). This bibliographic method is becoming increasingly popular for providing an overall picture of the published literature (Rodrigues et al., 2014; Taheri et al., 2021). To explore the uptake of these papers we searched the Web of Science Core Collection (Clarivate Analytics) in October 2020 to identify the set of all papers that have cited either Torrance et al (1972) or Fanshel and Bush (1970), or both, and examined how this set of papers are co-cited.

A network visualisation map of the results of the co-citation analysis was created using the software VOSviewer version 1.6.14 (https://www.vosviewer.com/). In a first analysis, we produced a network visualisation map reporting the results of the co-citation analysis at the level of the journal. VOSviewer performed a cluster analysis to identify clusters of closely related publications within journals (Waltman et al., 2010). In a second analysis, we conducted an exploratory analysis to explore the methodological development of the TTO within these publications. This exploratory analysis involved three steps given the uncertainty around the feasibility of identifying methodological publications and topics using co-citation analysis. In step 1, we ran a cluster analysis on all the co-citing publications with 20 citations or more to generate an overall picture. In step 2 we removed from the analysis those papers published in clinical journals, to focus on publications in the journals that were more likely to cover methodological developments. To achieve this, journal titles were examined manually by two reviewers independently (AS and AT), and a consensus reached on which clinical journals to exclude. A new cluster analysis was carried out on VOSviewer to identify clusters of closely related publications using the included papers. In step 3, titles and abstracts of the publications falling within each cluster were examined manually and sorted into themes by two reviewers independently (AS and AT), and a consensus reached on topics represented. These topics were reviewed by a panel of five researchers (AS, AT, PC, ORA, RW) to resolve any disputes and confirm topics.

3.1.2 Results

The full bibliographic details of the set of 716 publications citing Fanshel & Bush (1970, 486 publications) and/or Torrance et al., (1972, 272 publications). Figure 1 illustrates the cumulative citations over this time. The 716 publications included on average 29 references, and after removing duplicates there were 20,509 unique references across 7,099 journals. Each reference has a citation count ranging from 1-480, and the collective number of citations is 31,344. A flow diagram reporting the identification of publications and the process of the co-citation analysis is shown in supplementary material 2. The complete list of the 716 publications is reported in supplementary material 3.

<<Figure 1: Uptake of publications over time>>

For the visualisation map at a journal level in Figure 2, we focus on journals with 50 or more citations to facilitate the interpretation of the map. There are 80 journals that cleared this threshold, covering 14,631 citations. In Figure 2, each journal is represented by a node on the map. Larger nodes display the journal title and cover a larger proportion of the total citations and/or co-citations. For some nodes, the journal title may not be displayed due to the small number of citations. The more frequently publications co-cite one another, the closer together the nodes are positioned on the map, forming clusters of nodes highlighted by VOSviewer in colour. Journals within the same cluster are assigned the same colour.

The map identifies six distinct journal clusters, based on the number of times they are co-cited across these 14,631 publications. These journal clusters fall broadly into three research areas: the upper right area of the visualization display Health Economics journals that publish economic analyses and Health Policy analyses (blue), the lower right area relate to health services and operational research (red), and to the upper left area are medical journals that are general and specialist (green, yellow).

<<Figure 2: Journal co-citation network visualization map>>

In the visualisation map at publication level, we focus on publications with 20 or more publications, where 363 out of the set of 716 publications cleared this threshold. The first cluster analysis on these publications identified 29 clusters based on citations and/or co-citations, 25 of the 29 clusters covering

335 of the publications. Each cluster contains 7 to 24 publications, but the related clusters were very heterogeneous making interpretation difficult. The second cluster analysis based on non-clinical journals included 204 publications and excluded the remaining 131 publications from clinical journals. VOSviewer excluded a further four publications because their references had no links. Figure 3 reports this second cluster analysis, where each publication is represented by a node and labelled by the first author for the larger nodes. In this map, publications within the same cluster are assigned the same colour.

Titles and abstracts of the publications falling within each cluster were examined manually to characterise the topic covered in each. Within the analysis, we identified a number of highly cited publications that aimed to summarise and review the methodological development of the TTO in previous decades. These summaries can be viewed as a precursor to the current review, and include reviews by Bergner with 152 citations (Bergner, 1985), McHorney with 188 citations (McHorney, 1999), Green with 142 citations (Green et al., 2000), Gold with 305 citations (Gold et al., 2002) and Weinstein with 320 citations (Weinstein et al., 2009).

Early on in this process, it became clear that papers from distinct literatures addressing very different topics were placed in the same cluster. This suggested that papers within a given cluster were not necessarily basing referencing decisions on a shared understanding of the relevant literature or following a shared set of decision rules. This is possible, since two papers may be in the same cluster if they cited a similar set of references, irrespective of what the research questions were. The analysis, therefore, moved on to identify related topics across the clusters. We broadly summarise the range of topics identified from this additional review process:

1) Assessing how to measure quality of care in hospitals

There were a set of papers around the 1970s that were exploring the meaningful measure of health with the aim of assessing quality of care in hospitals (Donabedian, 1972; Fanshel, 1972; Rosser & Watts, 1972). There papers were mostly in the cluster around Donabedian that received the second highest number of citations, with 3111 citations (Donabedian, 1972).

2) Describing health states

Early calls for an operational definition of health (Patrick et al., 1973) and the work on the influence of health state descriptions on responses (Llewellyn-Thomas et al., 1984) paved the way to the development of standardisation of classification systems to capture health, including the Sickness Impact Scale (Gilson et al., 1975), the 15D questionnaire (Sintonen, 1981), Health Utilities Index Mark 2 (Torrance et al., 1996) and EQ-5D (EuroQol Group, 1990). Later papers review these classification systems (Coons et al., 2000). The most highly cited publication on the EQ-5D (i.e. Williams 1990 in Figure 3)[1], received 8093 citations, and shows the importance of this work within this literature.

3) Application of Expected Utility Theory and Non-Expected Utility Theory

Publications exploring behavioural theories of respondents' decision making and exploration of heuristics and biases start with Pliskin and colleagues (Pliskin et al., 1980) and Miyamoto and Eraker (Miyamoto & Eraker, 1985), which applied Expected Utility Theory. These were followed by considerations of risk attitude affecting the TTO and SG differently (Stiggelbout et al., 1994) and failures of the axioms underpinning these (Stalmeier et al., 1996). The application of Non-Expected Utility Theory came much later in these searches and highlighted inconsistencies (Bleichrodt & Quiggin, 1997), and the need to control for the associated biases (van Osch et al., 2004).

4) Measurement properties of different methods

There were publications that summarised the measurement properties of different methods, and include exploring: the association of functional status and utility valuation (Tsevat et al., 1991), the biases in assessment (Kaplan & Ernst, 1983), the stability of utility valuation measures over time (Tsevat et al., 1993), and the ways in which the valuation measures are impacted by respondent heterogeneity (Bult et al., 1998).

5) Order effects

[1] The recommended reference details is "EuroQol Group. EuroQol - a new facility for the measurement of health-related quality of life. Health Policy, 1990. 16: 199-208" (https://euroqol.org/eq-5d-instruments/eq-5d-3l-about/faqs/). However, Web of Science uses the corresponding author of the paper as first author instead of the corporate EuroQol Group, which was used in our co-citation analysis software.

The potential for order effects to impact on valuations was discussed in the literature around Healthy Years Equivalent (Gafni et al., 1993). Questions were asked around the predictive performance of the valuations of health states, and if patients evaluation of future health change when they enter them by Llewellyn-Thomas and colleagues (Llewellyn-Thomas et al., 1993). These debates related to an earlier discussion around how best to capture the value of temporary states (Torrance, 1986), which is a theme continued throughout the period (Wright et al., 2009). The Lead time TTO method aimed to overcome order effects by standardising procedures for the valuation of better than dead and worse than dead states (Devlin et al., 2011).

6) Elicitations using ranking data and discrete choice experiments

The application of ordinal methods such as conjoint analysis (Ryan & Hughes, 1997) and ranking data to value health states was seen to offer similar results to cardinal methods (Craig et al., 2009). Discrete choice experiments (DCE) led on from these developments (Ali & Ronaldson, 2012), and was later modified to include information on health state classifications (such as the EQ-5D) and length of life within a DCE (Bansback et al., 2012).

7) Beyond welfarist and the QALY

There was a distinct cluster around discussions of interpreting the QALY from distinct normative frameworks such as welfarism, extra-welfarism and Sen's capability approach around the late 2000s, reviewed by Coast and colleagues (Coast et al., 2008b).

8) Decision rules

Concerns over the decision rules were raised by Birch and colleagues (Birch & Gafni, 1992), including the use of QALY league tables. These concerns led to the consideration of other mathematical programming for efficient allocation of resources (Stinnett & Paltiel, 1996).

<<Figure 3: Publication co-citation network visualization map>>

3.1.4 Findings

Our co-citation analysis show that the QALY has influenced a range of research areas with an impressive range of methodological topics. This reflects the success of the QALY to adapt and incorporate new methodological developments over time. In the next section, we illustrate how the concept of the QALY has developed over time to address methodological challenges using behavioural theories covering three of the eight topics identified above.

3.2 Behavioural theories that have influenced the way health state preferences are elicited

Our understanding of what the QALY represents and the methods that can be used to elicit the preferences to quality-adjust the life years has changed over the years. In many cases these developments can be traced back to theories emanating from Economics and Decision Science as well as from within Health Economics. In this section, we illustrate the behavioural theories that have informed the research question of "How to get the numbers?", with the aim to explore how these theories have influenced and reverberated within Health Economics. We draw on the set of publications that cite Torrance (1970), Torrance et al (1972) and/or Fanshel and Bush (1970), identified by the Scopus search. We additionally include systematic review articles that have explored the behavioural aspects of the TTO, and related publications known to the authors but that did not cite the source papers and so were not identified in the Scopus search. What we present below are illustrative, and not exhaustive of theories that have influenced the elicitation of health state preferences. Parallel developments within the other topics identified in the publication cluster analysis, including the development and classification of health states and health indices, were taking place at the same time in Europe and elsewhere, but we do not go into further detail here.

Figure 4 summarises four landmark methodological theories over the past 50 years originating within Economics and Decision Science. These theories link to three of the eight topics identified in the cluster analysis above and which we categorise here as Expected Utility Theory (von Neumann & Morgenstern, 1944), Non-Expected Utility Theory (Kahneman & Tversky, 1979; Loewenstein & Prelec, 1992; Loomes & Sugden, 1982) and Probabilistic Choice Theory (McFadden, 1974). A fourth topic, Order Effects, is defined broadly to encompass behavioural theories around timing/sequence of events (Loewenstein & Prelec, 1993). The top half of this figure documents key references summarising these methodological theories within Economics and Decision Science, visualising these

on a timeline to illustrate the developments over time. The bottom half of Figure 4 summarises examples of when and in what ways these theories were first acknowledged in the health economics literature and how they have led to methodological innovation of the QALY that still reverberate today.

<<Figure 4: Behavioural theories that have influenced HEs over the past 50 years >>

Expected Utility Theory preceded Torrance and Fanshel and Bush (Fanshel & Bush, 1970; Torrance, 1970) – but it was not until after Keeney and Raiffa's influential book on decision theory (Keeney & Raiffa, 1976) that an axiomatic basis for the QALY based under Expect Utility Theory was proposed (Miyamoto & Eraker, 1985; Pliskin et al., 1980). In 1985 the seminal paper by Williams illustrated the use of QALYs to capture the health benefits of coronary heart bypass surgery (Williams, 1985). Within this framework challenges of how to value prolonged period of ill-health, and the notion of maximal endurable time in ill-health was likened to the concept of diminished marginal utility (Sutherland et al., 1982). Under the Expected Utility framework the SG (Gudex, 1994a) and TTO (Gudex, 1994b) were initially explored in the UK valuation set of the EQ-5D-3L (EuroQol Group, 1990). The original UK EQ-5D-3L valuation study used the TTO to develop a value set (Dolan et al., 1996) and this value set was subsequently employed to estimate population norms (Kind et al., 1999). There was experimentation with alternatives to the TTO but these were found to have their own limitations that are still not fully resolved today. For example, the Equivalent Numbers method (Berg, 1973; Bush et al., 1973; Torrance, 1986) later referred to as the person trade-off (PTO) (Nord et al., 1993; Olsen, 1994) allowed for interpersonal comparisons of QALYs but the axiomatic basis of these methods within Expected Utility Theory has been questioned (Doctor et al., 2009). Later developments have included asking respondents to value their own health, termed experienced-based utility, rather than a hypothetical health state to inform the EQ-5D-3L value set (Burstrom et al., 2006).

Throughout the 1990s there was mounting evidence of systematic differences in the values based on the methods used and/or violations of the Expected Utility axioms of the QALY (Doctor et al., 2010), which led to alternative specifications for the axiomatic basis for the QALY (Bleichrodt & Quiggin, 1997). Insights from psychology about 'framing' led to discussions of 'heuristics' in decision making

(Kahneman & Tversky, 1984). In early 2000, the differences between health state values derived from the TTO and SG (Bleichrodt, 2002) and inconsistencies in the TTO method (Spencer, 2003) were found to be explained by Prospect Theory. This viewpoint remains with us today (Abellan-Perpinan et al., 2009) and has led to calls to correct for these potential biases in the tariffs (Lipman et al., 2019).

Another important influence was the discovery of Order Effects and preferences for sequences of events over time (Chapman et al., 1996; Loewenstein & Prelec, 1993). This led to scrutiny about whether it was appropriate to add the utilities from constituent health states when health varies over time (Gafni, 1995; Treadwell, 1998). These ideas also resonated with earlier suggestions to value sequences of health states, as proposed by the Healthy Years Equivalents method (Mehrez & Gafni, 1989). Elicitation procedures could unwittingly introduce order effects, for example, in the valuation of better than dead and worse than dead states (Robinson & Spencer, 2006). The Lead time TTO method (Devlin et al., 2011) aimed to overcome these issues by standardising procedures for the valuation of better than dead and worse than dead states. However, the additive separability that was assumed by these methods has been shown to be violated by interaction effects with earlier periods (Attema et al., 2013b; Pinto-Prades & Rodriguez-Miguez, 2015). In later work, order effects could be seen to influence other dimensions, for example, the order in which the states were valued (Attema et al., 2013a; Chuang & Kind, 2011). The challenges of the best methods to value temporary or recurring states, and the associated order effects, still reverberate within Health Economics today (Boye et al., 2014; Neumann et al., 2018; Ogwulu et al., 2017).

Random Utility Theory (Manski, 2001; McFadden, 1974) and later Discrete Choice Experiments (DCEs) (Louviere et al., 2000) replaced the notion of deterministic choice with probabilistic choice theory and applied to health (Bansback et al., 2012; Ryan & Hughes, 1997). The DCE was readily adapted to replicate the TTO or SG methods on an ordinal scale (Mulhern et al., 2019; Norman et al., 2013) and methods were developed to link these ordinal scales to a cardinal TTO scale (Bansback et al., 2012; Stolk et al., 2010) as well as exploring ways to validate states worse than being dead (Robinson et al., 2015). These developments led to use of the DCE method to value the EQ-5D-5L in some countries (Viney et al., 2014). The UK revaluation of the EQ-5D 5L used both a hybrid between the DCE and TTO methods (Devlin et al., 2018) but this approach has come under scrutiny

(Hernandez-Alava et al., 2018). Research continues to explore the robustness of these methods and comparability with existing TTO valuations (Robinson et al., 2017) alongside research into the econometric methods needed to meaningfully interpret responses (de Bekker-Grob et al., 2019; Feng et al., 2018; Lancsar et al., 2017; Soekhai et al., 2019).

4. Potential directions for future QALY research over the next 50 years

What are gaps in the current literature that need to be explored over the next 50 years? To reflect on the contributions of Fanshel and Bush and Torrance (Fanshel & Bush, 1970; Torrance, 1970) we elicited contributions from researchers who had made theoretical contributions to the QALYs, or undertaken empirical work involving health state measurement and economic evaluation. Contributors were asked to reflect on the development of the concept of the QALY, but to focus on the "current limitations of health measurement and valuation and their application in practice, and what type of research needs to be done that in 50 years' time would seem to be as innovative as Fanshel and Bush and Torrance papers." Contributors were given the opportunity to revise contributions at the draft manuscript stage, but there was no attempt to achieve consensus across the authors and so we have denoted authorship of each contribution. The contributions are presented in no particular order.

4.1 How to improve QALYs using new insights from behavioural economics - Han Bleichrodt, Erasmus School of Economics, The Netherlands

Let me state from the outset that I believe QALYs should be a utility model and, therefore, reflect people's preferences for health. I believe that the simple QALY model proposed by Fanshel and Bush and Torrance (Fanshel & Bush, 1970; Torrance, 1970) serves that purpose surprisingly well. Probably the utility function for life duration could allow for some risk aversion (a power utility function with power equal to 0.75 performs well), but otherwise the model seems adequate at least for chronic health states. For non-chronic health states, we may need to allow for some interactions over time (e.g. (Guerrero & Herrero, 2005)). How best to do so is a topic of interest.

The main challenge for QALYs as I see it, is how to value health states. Even though a lot of progress has been made over the past 50 years (Brazier et al., 2016), I believe that the main methods that we

use, the TTO and the SG, are problematic as they include substantial biases and give utilities that are too high (Bleichrodt, 2002). Consequently, policy based on these methods overemphasizes extending life duration compared with improving quality of life. This overestimation is even more pronounced in chained SG measurements, as in the SF-6D. The way forward is, in my opinion, to use better behavioural theories, in particular prospect theory (Tversky & Kahneman, 1992). In Bleichrodt, Pinto, and Wakker (2001) (Bleichrodt et al., 2001), we showed how prospect theory can be used to improve the SG utilities without requiring additional measurements. Later papers confirmed the advantages of our method over the SG and the TTO (e.g. (Attema et al., 2012; van Osch et al., 2004). Prospect theory may also be fruitfully applied in the estimation of the willingness to pay (WTP) for a QALY. Having a reliable WTP for a QALY is clearly desirable, but current estimates suffer from inconsistencies. I conjecture that prospect theory may help to solve at least some of these inconsistencies. Other areas in which new behavioural theories may improve QALY-based decision making are belief elicitation when probabilities are unknown (an under-researched area in health economics I believe) and the incentivisation of health utility measurement using Bayesian truth serums and markets (Baillon, 2017; Prelec, 2004).

A final topic for future research is the aggregation of QALYs. The current COVID-19 crisis has put the question of how to trade off the health of different people to the fore. In the literature, several alternatives to simply aggregating QALYs have been proposed (e.g. severity of illness, proportional shortfall, fair innings). The question which of these is most consistent with societal values about fairness has, to the best of my knowledge, not yet been satisfactorily answered. Insights from prospect theory can also be useful in addressing this question (Bleichrodt et al., 2004).

The simple QALY model has proved to be very useful both in research and in policy and its first 50 years have been a success story. I expect that insights from behavioural economics will improve it further and will make its next 50 years even more successful.

4.2 What would be the role of the QALY for public health decision making in the next 50 years?

Rhiannon Tudor Edwards, Bangor University, UK

The QALY was designed to capture the health benefits of a health care intervention (Williams, 1985). It was never designed to capture the resulting wider externality benefits from prevention that might occur across a range of sectors of the economy (Edwards & McIntosh, 2019). Recognising this, the National Institute for Health and Care Excellence (NICE) brought out guidance in 2014 proposing the increased use of cost-benefit analysis (CBA) and cost-consequence analysis (CCA) in the evaluation of the cost-effectiveness of public health interventions, rather than the recommended cost-utility analysis (CUA) for health care technologies based on a reference case cost-per-QALY threshold (National Institute for Health and Clinical Excellence, 2012, 2018). However, Owen and Fischer (Owen & Fischer, 2019) show 85% of public health interventions evaluated by NICE fall below the threshold of £20,000-£30,000 per QALY.– This is significant when we only spend 4% of the NHS budget on prevention (Masters et al., 2017).

Eighty percent of chronic health problems, such as premature heart disease, stroke and diabetes, are preventable (World Health Organization, 2020). Fifty percent of all GP appointments, 64% of outpatient appointments and 70% of all inpatient bed stays prior to COVID-19 were for preventable conditions (Department of Health, 2012; The King's Fund, 2021).

In the aftermath of an unprecedented pandemic when virtually unlimited resources were made available by the UK Treasury to uphold the economy through furlough schemes and the NHS received additional funding based on the precautionary principle (Fischer & Ghelardi, 2016), there will be a role in future for the QALY, and cost-per-QALY calculations to reflect on the true opportunity cost of applying this precautionary principle to the management of the pandemic. There will be a role for the QALY in estimating and informing future debate about what society is in fact prepared to pay right across the economy to tackle future pandemics.

The UK Government is planning, at the time of preparing this contribution, a Keynesian programme of widespread investment in the material infrastructure of the UK economy financed by low cost borrowing. Decisions about what projects and programmes to invest in across a wide range of sectors

will require CBA for which the Treasury has established guidelines (Treasury, 2018, 2020). The QALY will have a role, valued at either the NICE threshold of £20,000 to £30,000 or Treasury valuation of £60,000 (Department of Health, 2009).

In terms of prevention at a population level, there is a need for research to capture the total health gains in terms of e.g. the QALY gains from legislation and potential changes in industry and social norms that arise and are enabled through very small changes in behaviour and resultant health benefits across many individuals in the population. Relevant examples include: reducing salt content of bread; reducing sugar content in fizzy drinks; stopping smoking in public places, and introducing minimum alcohol pricing. Research is needed to articulate to government the total health gains in terms of QALYs gained or avoided QALY losses from such population level preventative public health measures across society from housing, education, transport and the environment.

Ongoing work on EQ-5D value sets for QALY calculation both nationally and internationally will be needed to make sure that values for either the three or five level value sets are truly representative of populations across Europe in which 10 years of austerity compounded by the current economic consequences on the COVID-19 pandemic is widening the gap between the wealthy and the poor (Janssen et al., 2019). Any future value sets must be representative of geography, socio-economic status, ethnicity, and possibly health state experience (living with a chronic disease or disability) (Zakaria, 2020). The development of the EuroQol Health and Wellbeing (EQ-HWB) measure may prove very useful in the field of public health economics (Peasgood et al., 2021).

Finally, continued healthy debate between the QALY paradigm and other alternative paradigms such as the capabilities approach keep us challenging our heuristics about well-being and health gain measurement in public policy (Edwards & Lawrence, 2021; Lorgelly et al., 2010).

4.3 Welfarism, non-welfarism and the QALY- Aki Tsuchiya, University of Sheffield, UK

Health economics started as a branch of microeconomics, and microeconomics is (typically) welfarist. Welfarism is the principle under which “the judgment of the relative goodness of alternative states of

affairs must be based exclusively on, and taken as an increasing function of, the respective collections of individual utilities in these states." (Sen, 1979, p.468).

But the Quality-Adjusted Life Year (QALY) as a policy maximand is not welfarist. There were efforts in the 1990s to give the QALY a welfarist foundation – but it was too hard.

Instead, much of the health economics literature assumes that health-related social welfare is maximised when the amount of health across a population, measured in QALYs, is maximised. The QALY is operationalised by adjusting years of life using a population value set, and this reliance on a standard set of values for everybody means that maximising aggregate health measured in QALYs does not generally maximise aggregate individual utility.

Interestingly, however, methods of health state valuation, used to estimate the population value sets, are typically quite welfarist. In a health state valuation study, the respondent is shown two possible health scenarios and asked to indicate their personal preference (e.g. "would you prefer to live in this health state for 10 years and die, or in full health for 5 years and die?"). A less welfarist practice might ask respondents to indicate which of two groups of patients should be given higher priority in a publicly funded health care system.

At the same time, health state valuations do not measure individual utility associated with different health states on an absolute scale, but generate relative adjustment weights for the QALY. Note that the life-year component of the QALY is not subject to valuation – all life years are given the same social value, with a fixed threshold price.

One implication of working with the QALY has been that everything else is assumed to remain the same – if all non-health aspects of life remain the same, then maximising QALYs would improve social welfare. We have analysed technical efficiency, assuming that the exogenous cost-effectiveness threshold is valid, that health care only affects health, and that only health care affects health. But the rest of life does not remain the same, health care spills over into non-health areas, and non-health policies have health implications. The QALY today is focused entirely on health, but it

could become extra-health, using quality-adjustment weights that capture general quality of life and well-being.

Over the past 50 years, the QALY has helped health economics grow out of microeconomics. Health economists have stepped away from welfarism, adopted health as the policy maximand, and developed economic evaluation. The next 50 years, we could re-assess some of what we do with the QALY: think of less welfarist ways of valuing health (and well-being); expand quality-adjustment to go beyond health; and explore ways of analysing allocative efficiency with the QALY. Extra-health economists could give back to economics a non-welfarist welfare economics.

4.4 Challenges of using QALYs in some specific contexts – Oliver Rivero-Arias, University of Oxford, UK

Many health economists (including me) tend to ignore some of the assumptions of the conventional QALY model (e.g. risk neutrality over time) when estimating the health benefits of an intervention (Weinstein et al., 2009). The many advantages of the model and the lack of a convincing alternative are strong arguments to turn a blind eye on such limitations. For many clinical conditions where measuring and valuing health gains with standard approaches do not present a dilemma, this works remarkably well. However, the model presents challenges when employed in areas such as the evaluation of rare diseases and end of life interventions. The inability to accurately capture health gains, relevant non-health benefits, potential spillover health effects on carers and the evidence suggesting that a "QALY may not be a QALY, may not be a QALY" are examples where the QALY fails to impress (Nord et al., 2009).

HTAs organisations, with NICE being the clear example, have implemented the pragmatic solution of increasing the cost-per-QALY threshold representing the opportunity cost of funding a treatment in areas where the conventional QALY struggles. A solution that is not supported with evidence and with an arbitrary selection of thresholds. Solutions that are more elegant are needed including the incorporation of spillover effects, equity and fairness in the measure of health benefits in the denominator of the incremental cost-effectiveness ratio. Although some approaches has been proposed (Al-Janabi et al., 2016; Cookson et al., 2009), their wider application is still limited. Given that HTA organisations are explicitly recognising the shortcomings of conventional QALYs, it is not

difficult to imagine an increase in the number of methodological developments on these lines and their applications in empirical studies in the future.

Another potential solution is to extend the remit of the QALY to incorporate additional elements that society values (Coast et al., 2015). Empirical work has shown us that health is not the only dimension society values when making health care decisions and other elements including functioning and capabilities may be as important as QALYs. It is difficult to foresee a future without the QALY being replaced completed by capabilities well-being measures. Therefore, a key development for the next half century will be the search of appropriate complex analytical techniques to maximise multidimensional outcomes. Rather than aggregating multiple dimensions into a single metric, a framework that simultaneously facilitates the maximisation of QALYs and people's freedoms to achieve elements of well-being is needed. Interestingly, optimisation of multidimensional outcomes was considered 50 years ago by Zabronsky (1967), but it was analytically unattainable at the time, favouring the nurturing of ideas that resulted in the development of a single index (Packer, 1968).

Will the current thrive of methodological innovations incorporating more health and non-health elements into the QALY continue? Will a landmark discovery to optimise multidimensional outcomes occur changing our perspective about making decisions in health? The next 50 years will be a fascinating time for the QALY and certainly an interesting time for the next generation of health economists.

4.5 QALYs in a risky imprecise world - Anne Spencer, University of Exeter, UK

The methods that Torrance and Fanshel and Bush (Fanshel & Bush, 1970; Torrance, 1970) proposed provide an intuitive concept of the opportunity costs and trade-offs which appeal to policy makers.

The momentum to apply these methods to value new classification systems and to develop new tariffs, including condition specific measures and well-being, has narrowed the focus to decision making under certainty. How well are such tariffs able to cope when we face risk and uncertainty arising in the COVID-19 epidemic? Theories of decision making within economics and psychology

encompass risk and uncertainty and were the inspiration for the SG. Some of the earliest studies valuing generic health states framed the decision under certainty, using the TTO (Gudex, 1994b), and under risk using the SG (Gudex, 1994a). Research has shown the ability to apply a risk-based discrete choice experiment to derive utility values for health states under non-expected utility models (Robinson et al., 2015) that readily incorporate worse than dead health states (Robinson & Spencer, 2006). Though it is argued that the SG is a stylised portrayal of the risk involved in medical decision-making, the same criticisms of unrealism apply to the TTO with assumes away the uncertainties about expected life expectancy. As we reassess the usefulness of our QALY tariffs post COVID-19, the need to consider how to incorporate risk and uncertainty within decision-making will become a more pressing issue over the next 50 years.

Do we have well defined preferences for aspects of our lives that we have yet to experience? Moreover, when the choices involve quiet complex questions – how easy are they for us to answer these questions (Peeters & Stiggelbout, 2009)? Adding additional information to the descriptive system does not seem to reduce the discrepancies between experienced and hypothetical (Burstrom et al., 2006) and has led to incorporating people's lived experiences and experienced-based valuations for QALY tariffs (Brazier et al., 2005). Increasing recognition that stated preferences from elicitation studies are imprecise (Butler & Loomes, 2007) is often overlooked. Under conventional approaches the factors that lead to these errors are not systematic and cancel out on average, but under imprecision, people may take 'cues' from the way the question is asked which may lead them to more likely to choose one option over another, and this can lead to systematic differences on average (Pinto-Prades et al., 2018).

I believe that the next 50 years will recognise the need to involve people in meaningful discussion of length and quality of life under uncertainty, acknowledging their imprecision in responses, to guide policy makers.

4.6 Practicalities of eliciting preferences: common problems and potential solutions - Andrew Lloyd, Director of Acaster Lloyd Consulting Ltd., UK

The TTO method is used to elicit the public's view of health states (Drummond et al., 2015). The quality of the TTO method and associated estimation should therefore be judged by the extent to which it really reflects the views of society. Put another way we should judge the quality of TTO valuation data in terms of whether the elicitation or estimation procedures are causing systematic error. Kind and Chuang (Kind & Chuang, 2019) illustrate how there may be problems arising from both elicitation and estimation in national TTO-based value sets.

The EuroQol Group have attempted to address sources of bias or error from the TTO by standardising the methods (Oppe et al., 2014). Computer-based administration of experimental stimuli, use of quality assurance methods and a standardised approach to the statistical estimation methods have all been implemented in recent years. This has certainly reduced sources of error. This work is important because it is perhaps the most comprehensive critical examination of the TTO that has ever been conducted. Methods have moved forward, but challenges remain.

The TTO method asks participants to consider themselves in a defined health state. If we consider the example of the EQ-5D, the health states typically describe a quite significant impact on quality of life leaving the participant with problems completing usual activities, washing & dressing, problems with mobility and the experience of anxiety, depression or pain. Many (or even most) TTO participants have probably not experienced long periods (months and years) with these problems and so it is quite difficult to judge how bad it would be. Participants are not given any information as to why they are experiencing this state. Indeed, it is not clear at all how participants imagine the impact or burden of the states. If the elements of the vignette (mobility, usual activities, pain etc) are not internally consistent with each other then there is some evidence that participants may ignore parts of the vignette (Lloyd & Quadri, 2008). There is also evidence that a process of deliberation can often lead people to change their values, suggesting that their preferences may not be as well established as the method assumes (Robinson & Bryan, 2013). Deliberative processes are time consuming and expensive, but could well produce better quality data.

One significant challenge with the TTO method is the valuation of states worse than dead. The valuation of states worse than dead influences the value of all states because it determines the position of the lower end of the scale. For this reason indirectly it is a very important driver of the cost effectiveness of treatments. In the original UK EQ-5D-3L valuation study the assessment of states worse than dead employed a very different version of the TTO with an arbitrary rescaling of values (Dolan, 1997). More recently the lead time TTO method has been used which avoids the need to change the TTO task. But the lead time task may still lead to framing effects. A high number of states valued at -1.0 is often seen for example in lead time TTO research (Devlin et al., 2018). Such findings raise questions about what the natural distribution of TTO values should look like (Al Sayah et al., 2016).

The valuation of states worse than dead currently assumes that people have continuous preferences for such states – meaning for example that people consider some states as a little worse than dead and other states as a lot worse than dead. This assumption itself may be wrong. It is possible that people accept states as not worth living but simply attach an equal value to all such states. Bernfort and colleagues (Bernfort et al., 2018) have queried the validity of the valuation of states worse than dead through research with people in such states. The conceptualisation of states worse than dead is an important influence on TTO scores but is still quite poorly understood.

4.7 Measuring health status for QALY calculation, are we there yet? Richard Norman, Curtin University, Australia

A central aspect of the QALY approach that has developed over the last 50 years are the instruments we use to describe health. These instruments are designed to have a range of desirable features. First, they need to be able to detect small but important differences in health status, both in cross-sections and (perhaps more importantly for QALY construction) longitudinally. Second, they have to not place undue burden on respondents; in a world where survey respondents frequently face a battery of instruments, it is important to use parsimonious instruments that do not tire or bore respondents into using simplifying heuristics. Finally, it is normally the case that these instruments seek to be generalisable across conditions. Obviously, there is tension between these three issues. In

particular, the desire for sensitivity is often at odds with both parsimony and generalisability, and the range of instruments available for QALY construction reflect different points on the trade-off continuum between the three.

Early work on these instruments included the widely-used Rosser disability/distress scale (Rosser & Kind, 1978; Rosser & Watts, 1972), which in turn was partly informed by the work on the Health-Status Index of Fanshel and Bush (Fanshel & Bush, 1970). Indeed, the 1972 paper by Rosser and Watts conducted at St. Olave's Hospital (Rosser & Watts, 1972) provides a two dimension health system (disability with eight levels, distress with four) which is clearly a forerunner of the instruments we use today. The current landscape has a range of instruments which can be used to describe health within the QALY framework. The EQ-5D is the most widely used of these in most settings. The instrument, developed through a wide international collaboration of researchers and clinicians, is extremely parsimonious, with all three versions (the 3-level, the 5-level and the youth version) consisting of five questions only. However, questions remain about whether it is adequately sensitive in a range of settings. For example, recent work by Shah et al explored aspects of health not captured by the EQ-5D, concluding that sensory deprivation and mental health are inadequately measured in the instrument. These are areas, which are more fully included in other competing instruments, such as the SF-6D, AQoL, HUI-3 and the 15D (Shah et al., 2017).

Moving into the next 50 years, the trade-off between sensitivity, parsimony and generalisability will continue to play a major role in the choice of instruments to describe health in QALY construction. Sensitivity can likely be improved through consideration of health status in specific populations such as the CHU-9D in children (Ratcliffe et al., 2012). Equally, instruments can also be generated using disease-specific measures (in a similar way to deriving the SF-6D from the SF-36 or SF-12), such as the EORTC QLU-C10D (King et al., 2016; King et al., 2018). However, both extensions raise issues around the generalisability of value sets and resultant QALYs. Finally, there is ongoing valuable work considering expanding the instruments we use to consider a broader concept of outcome, such as the ICECAP suite of instruments (Coast et al., 2008a), and the EQ-HWB (Brazier & Tsuchiya, 2015). These extensions are exciting and are likely to form a major trend in the area. However, we

simultaneously must consider their role in health policy, given the outcomes they estimate may be too different from QALYs to allow reasonable comparison.

4.8 Revealing the QALY- Philip Clarke, Oxford University, UK.

The underlying empirical approach to the elicitation of the values for health states has remained largely unchanged over the past 50 years and very much builds on the approach proposed by Fanshel and Bush and Torrance (Fanshel & Bush, 1970; Torrance, 1970). While techniques and statistical methods have become more sophisticated, the basic premise is the same. Patients or the public are asked to state preferences for trade-offs between risk of death or length of life and continuing to experience the health state that is under valuation. This contrasts with the approach taken by economists to elicit a value of life in monetary terms which have involved the use of both stated preference methods such as the contingent valuation (Persson et al., 2001), but also revealed values via, for example, wage-risk trade-off studies. The latter uses hedonic methods to look at the wage premium for jobs with greater occupational risks (Viscusi & Aldy, 2003). Here I outline the scope for empirical deriving values for health states based on revealed values.

It important to recognize that patients do sometimes trade off quality and quantity of life. Take a patient with Parkinson disease (Lang, 2000): they can opt for surgery such as a pallidotomy that can reduce symptoms, but the mortality risk of operation is around 1% (de Bie et al., 2002). The decision to have surgery is akin to a SG in that they are seeking to improve quality of life through an intervention that places them at a risk of death.

Similarly, preventative drug treatments for chronic conditions such as cardiovascular disease can involve choices akin to TTOs. Take blood pressure medications for patient with hypertension, several classes of anti-hypertensive drugs lower mortality (Zanchetti et al., 2015), but often have a range of side effects that may impact on a patient's quality of life. A patient's decision to discontinue taking such medication again may involve trade-offs between life expectancy and current quality of life.

The development of hedonic valuation methods for a QALY would require a substantial research effort and faces a number of challenges. First, trade-offs in real life are more complex than the

stylised decisions that have been used to elicit stated preferences for health states. In particular, interventions such as neurosurgery do not necessarily return the patient to full health. For example, in the case of pallidotomy for Parkinson disease, complications of surgery may include worsening of memory, injury to optic tract and stroke (Metman & Kompoliti, 2010). Hence a patient's decision to opt for surgery is a complex gamble involving both a risk of death and multiple potential future health states.

A key issue will be capturing a wide enough range of trade-offs to identify a functional relationship between quality and quantity of life. One way forward would be to build on the development of multi-attribute utility instruments. For example, using such an instrument to measure the side effects of drugs and the degree to which adverse changes influence the propensity to discontinue use of medications (Ivarsson et al., 2019) may provide a practical way to obtain revealed values for a range of health states. Obtaining such values could greatly deepen our understanding of the value people place on their own quality of life.

5. Conclusion

This article is a celebration of the seminal papers by Torrance, Fanshel and Bush that defined much of the early QALY's development, documenting the history of these papers and charting the methodological debates that influenced its development over the last 50 years. The resilience of the concept of the QALY, its endorsement by policy makers in many countries and its ability to incorporate new methodological influences are hallmarks of its success.

There are a number of limitations to our co-citation analysis and network visualisations. Firstly, these were based on a search of publications citing Torrance (1970), Torrance et al (1972) and/or Fanshel and Bush (1970), and therefore omit publications that do not cite these. We partly overcame this limitation by exploring the development of the behavioural theories underpinning the QALY in more detail, referring to systematic reviews and publications known to the authors. We recommend that future research explores alternative search strategies, for example, based on keywords such as "trade-off". Such searches were beyond the scope of the current paper, requiring more resources to

exclude the large number of unrelated publications, for example, referring to trade-offs in optimisation techniques in computer science. We recommend that future work explores the use of keywords and abstracts to develop network visualisation maps to explore more granular interpretation of the development of research topics and themes over time.

The invited contributions - the many voices of the title - provide an overview for the anticipated research agenda going forward. Topics for future research can be categorised into three broad themes covering the research questions: 1) What the numbers mean?; 2) How to get the numbers?; and 3) What aspects beyond health to incorporate? Some of these research questions were touched on by Torrance in his 2006 paper tantalizingly entitled – things I never got to (Torrance, 2006).

Research into 'what the numbers mean' includes the interpretations of the QALY as a metric of individual utility versus a metric of social good; and whether the QALY should be maximised with little regard to allocative efficiency or distribution. Another challenge is the meaning of negative quality-adjustment weights and the meaning of 'states worse than being dead'. Issues related to how health is described, requiring a balance to be struck between sensitivity, parsimony and generalisability of the descriptive system also belong here. A lingering concern, particularly in the United States, has been that the use of the QALY in economic evaluations are regarded as discriminatory against those who are older or who have a disability, and suggestions such as the Equal Value of Life Years Gained have been proposed to address this (Pearson, 2019).

Research into 'how to get the numbers' need to consider new behavioural theories, and ways of incorporating risk and uncertainty within decision making when probabilities are unknown and preferences are imprecise. Lakdawalla and Phelps propose a Generalized Risk-Adjusted QALY as one way of capturing attitudes to risk as well quality of life in a single index (Lakdawalla & Phelps, 2020). There is also the conundrum of how to value worse than dead states on a scale that is comparable with states better than dead, which George Torrance when interviewed in 2020 suggested was a field that required further research (see George Torrance Interview Transcript in the supplementary material 1).

Research into issues beyond health includes a potential use of QALYs to address allocative efficiency, and to capture outcomes affecting housing, education, transport and the environment. These broader issues may be incorporated through multi-attribute utility and/or revealed preferences. Additional aspects such as equity considerations could be addressed here.

The next 50 years is likely to see developments in all these areas, and more. As researchers remain curious and constantly strive to develop the methodological frameworks, the QALY lets us aim to influence policy makers to remain open to new developments even when these challenge existing valuation methods.

6. References

Abellan-Perpinan, J.M., Bleichrodt, H., & Pinto-Prades, J.L. (2009). The predictive validity of prospect theory versus expected utility in health utility measurement. *J Health Econ,* 28, 1039-1047.

Al-Janabi, H., van Exel, J., Brouwer, W., & Coast, J. (2016). A Framework for Including Family Health Spillovers in Economic Evaluation. *Med Decis Making,* 36, 176-186.

Al Sayah, F., Mladenovic, A., Gaebel, K., Xie, F., & Johnson, J.A. (2016). How dead is dead? Qualitative findings from participants of combined traditional and lead-time time trade-off valuations. *Qual Life Res,* 25, 35-43.

Ali, S., & Ronaldson, S. (2012). Ordinal preference elicitation methods in health economics and health services research: using discrete choice experiments and ranking methods. *Br Med Bull,* 103, 21-44.

Anderson, J.P., Kaplan, R.M., Schneider, A.M., & Schneiderman, L.J. (1987). *James W. Bush, Community and Family Medicine: San Diego*: http://texts.cdlib.org/view?docId=hb6z09p0jh;NAAN=13030&doc.view=frames&chunk.id=div00007&toc.depth=1&toc.id=&brand=calisphere.

Anon (1899). Vaccination in India. *British Medical Journal,* 1, 1341-1341.

Arrow, K.J. (1963). Uncertainty and the Welfare Economics of Medical Care. *The American Economic Review,* 53, 941-973.

Attema, A.E., Bleichrodt, H., & Wakker, P.P. (2012). A direct method for measuring discounting and QALYs more easily and reliably. *Med Decis Making,* 32, 583-593.

Attema, A.E., Edelaar-Peeters, Y., Versteegh, M.M., & Stolk, E.A. (2013a). Time trade-off: one methodology, different methods. *Eur J Health Econ,* 14 Suppl 1, S53-64.

Attema, A.E., Versteegh, M.M., Oppe, M., Brouwer, W.B., & Stolk, E.A. (2013b). Lead time TTO: leading to better health state valuations? *Health Econ,* 22, 376-392.

Baillon, A. (2017). Bayesian markets to elicit private information. *Proc Natl Acad Sci U S A,* 114, 7958-7962.

Bansback, N., Brazier, J., Tsuchiya, A., & Anis, A. (2012). Using a discrete choice experiment to estimate health state utility values. *J Health Econ,* 31, 306-318.

Berg, R.L. (1973). Establishing the values of various conditions of life for a health status index. In R.L. Berg (Ed.), *Health status indexes* pp. 120-134). Chicago, IL: Hospital Research and Education Trust.

Bergner, M. (1985). Measurement of health status. *Medical Care,* 23, 696-704.

Bernfort, L., Gerdle, B., Husberg, M., & Levin, L.A. (2018). People in states worse than dead according to the EQ-5D UK value set: would they rather be dead? *Qual Life Res,* 27, 1827-1833.

Birch, S., & Gafni, A. (1992). Cost effectiveness/utility analyses. Do current decision rules lead us to where we want to be? *J Health Econ,* 11, 279-296.

Bleichrodt, H. (2002). A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Econ,* 11, 447-456.

Bleichrodt, H., Diecidue, E., & Quiggin, J. (2004). Equity weights in the allocation of health care: the rank-dependent QALY model. *J Health Econ,* 23, 157-171.

Bleichrodt, H., Pinto, J.L., & Wakker, P.P. (2001). Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science,* 47, 1498-1514.

Bleichrodt, H., & Quiggin, J. (1997). Characterizing QALYs under a general rank dependent utility model. *Journal of Risk and Uncertainty,* 15, 151-165.

Boye, K.S., Matza, L.S., Feeny, D.H., Johnston, J.A., Bowman, L., & Jordan, J.B. (2014). Challenges to time trade-off utility assessment methods: when should you consider alternative approaches? *Expert Review of Pharmacoeconomics & Outcomes Research,* 14, 437-450.

Brazier, J., Akehurst, R., Brennan, A., Dolan, P., Claxton, K., McCabe, C., et al. (2005). Should patients have a greater role in valuing health states? *Appl Health Econ Health Policy,* 4, 201-208.

Brazier, J., Ratcliffe, J., Saloman, J., & Tsuchiya, A. (2016). *Measuring and Valuing Health Benefits for Economic Evaluation (2 ed.)*. Oxford: Oxford University Press.

Brazier, J., & Tsuchiya, A. (2015). Improving Cross-Sector Comparisons: Going Beyond the Health-Related QALY. *Appl Health Econ Health Policy,* 13, 557-565.

Bult, J.R., Hunink, M.G.M., Tsevat, J., & Weinstein, M.C. (1998). Heterogeneity in the Relationship Between the Time Tradeoff and Short Form-36 for HIV-Infected and Primary Care Patients. *Medical Care,* 36, 523-532.

Burstrom, K., Johannesson, M., & Diderichsen, F. (2006). A comparison of individual and social time trade-off values for health states in the general population. *Health Policy,* 76, 359-370.
Bush, J.W., Chen, M.M., & Patrick, D.L. (1973). Health status index in cost effectiveness: Analysis of PKU program. In R.L. Berg (Ed.), *Health status indexes* pp. 172-208). Chicago, IL: Hospital Research and Education Trust.
Butler, D.J., & Loomes, G.C. (2007). Imprecision as an account of the preference reversal phenomenon. *American Economic Review,* 97, 277-297.
Chapman, G.B., Bergus, G.R., & Elstein, A.S. (1996). Order of information affects clinical judgment. *Journal of Behavioral Decision Making,* 9, 201-211.
Chiang, C.L. (1965). An Index of Health: Mathematical Models. *Vital Health Stat 1*, 1-19.
Chuang, L.H., & Kind, P. (2011). The effect of health state selection on the valuation of EQ-5D. *Med Decis Making,* 31, 186-194.
Coast, J., Flynn, T.N., Natarajan, L., Sproston, K., Lewis, J., Louviere, J.J., et al. (2008a). Valuing the ICECAP capability index for older people. *Soc Sci Med,* 67, 874-882.
Coast, J., Kinghorn, P., & Mitchell, P. (2015). The development of capability measures in health economics: opportunities, challenges and progress. *Patient,* 8, 119-126.
Coast, J., Smith, R.D., & Lorgelly, P. (2008b). Welfarism, extra-welfarism and capability: the spread of ideas in health economics. *Soc Sci Med,* 67, 1190-1198.
Cookson, R., Drummond, M., & Weatherly, H. (2009). Explicit incorporation of equity considerations into economic evaluation of public health interventions. *Health Econ Policy Law,* 4, 231-245.
Coons, S.J., Rao, S., Keininger, D.L., & Hays, R.D. (2000). A comparative review of generic quality-of-life instruments. *Pharmacoeconomics,* 17, 13-35.
Craig, B.M., Busschbach, J.J., & Salomon, J.A. (2009). Keep it simple: ranking health states yields values similar to cardinal measurement approaches. *J Clin Epidemiol,* 62, 296-305.
Crystal, R.A., & Brewster, A.W. (1966). Cost Benefit and Cost Effectiveness Analyses in the Health Field: An Introduction. *Inquiry,* 3, 3-13.
de Bekker-Grob, E.W., Swait, J.D., Kassahun, H.T., Bliemer, M.C.J., Jonker, M.F., Veldwijk, J., et al. (2019). Are Healthcare Choices Predictable? The Impact of Discrete Choice Experiment Designs and Models. *Value in Health,* 22, 1050-1062.
de Bie, R.M., de Haan, R.J., Schuurman, P.R., Esselink, R.A., Bosch, D.A., & Speelman, J.D. (2002). Morbidity and mortality following pallidotomy in Parkinson's disease: a systematic review. *Neurology,* 58, 1008-1012.
Department of Health (2009). *Quantifying health impacts of government policies: A how-to guide to quantifying the health impacts of government policies* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/216003/dh_120108.pdf.
Department of Health (2012). *Long-term conditions compendium of information*. Leeds: Department of Health https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/216528/dh_134486.pdf.
Devlin, N.J., Shah, K.K., Feng, Y., Mulhern, B., & van Hout, B. (2018). Valuing health-related quality of life: An EQ-5D-5L value set for England. *Health Econ,* 27, 7-22.
Devlin, N.J., Tsuchiya, A., Buckingham, K., & Tilling, C. (2011). A uniform time trade off method for states better and worse than dead: feasibility study of the 'lead time' approach. *Health Econ,* 20, 348-361.
Doctor, J.N., Bleichrodt, H., & Lin, H.J. (2010). Health utility bias: a systematic review and meta-analytic evaluation. *Med Decis Making,* 30, 58-67.
Doctor, J.N., Miyamoto, J., & Bleichrodt, H. (2009). When are person tradeoffs valid? *J Health Econ,* 28, 1018-1027.
Dolan, P. (1997). Modeling valuations for EuroQol health states. *Medical Care,* 35, 1095-1108.
Dolan, P., Gudex, C., Kind, P., & Williams, A. (1996). The time trade-off method: results from a general population study. *Health Econ,* 5, 141-154.
Donabedian, A. (1972). Models for Organizing the Delivery of Personal Health Services and Criteria for Evaluating Them. *The Milbank Memorial Fund Quarterly,* 50, 103-154.
Drummond, M. (1991). *Australian Guidelines for Cost-effectiveness Studies of Pharmaceuticals: The Thin End of the Boomerang?* York: University of York
Drummond, M.F., Sculpher, M., Claxton, K., Stoddart, G.L., & Torrance, G. (2015). *Methods for the economic evaluation of health care programmes (4 ed.)*. Oxford: Oxford University Press.

Edwards, R.T., & Lawrence, C.L. (2021). 'What you see is all there is': The importance of heuristics in cost-benefit analysis (CBA) and social return on investment (SROI) in the evaluation of public health interventions. *Applied Health Economics and Health Policy*.
Edwards, R.T., & McIntosh, E. (2019). *Applied Health Economics for Public Health Practice and Research*. Oxford: Oxford University Press.
Enthoven, A. (2019). How Systems Analysis, Cost-Effectiveness Analysis, or Benefit-Cost Analysis First Became Influential in Federal Government Program Decision-Making. *Journal of Benefit-Cost Analysis,* 10, 146-155.
EuroQol Group (1990). EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy,* 16, 199-208.
Evans, D., Freund, D., Dittus, R., Robertson, J., & Henry, D. (1990). *The use of economic analysis as a basis for inclusion of pharmaceutical products on the Pharmaceutical Benefits Scheme*. Australia, Canberra: Department of Health Housing and Community Services
Fanshel, S. (1972). A meaningful measure of health for epidemiology. *Int J Epidemiol,* 1, 319-337.
Fanshel, S., & Bush, J.W. (1970). Health-Status Index and Its Application to Health-Services Outcomes. *Operations Research,* 18, 1021-&.
Feldstein, M.S. (1963). Operational research and efficiency in the health service. *Lancet,* 1, 491-493.
Feng, Y., Devlin, N.J., Shah, K.K., Mulhern, B., & van Hout, B. (2018). New methods for modelling EQ-5D-5L value sets: An application to English data. *Health Econ,* 27, 23-38.
Fischer, A.J., & Ghelardi, G. (2016). The Precautionary Principle, Evidence-Based Medicine, and Decision Theory in Public Health Evaluation. *Front Public Health,* 4, 107.
Foster, R.B., & Hoeber, F.P. (1955). Cost-effectiveness analysis for strategic decisions. *Journal of the Operations Research Society of America,* 3, 482-493.
Gafni, A. (1995). Time in health: can we measure individuals' "pure time preferences"? *Med Decis Making,* 15, 31-37.
Gafni, A., Birch, S., & Mehrez, A. (1993). Economics, health and health economics: HYEs (healthy-years equivalent) versus QALYs (quality-adjusted live-year). *J Health Econ,* 12, 325-339.
Gilson, B.S., Gilson, J.S., Bergner, M., Bobbit, R.A., Kressel, S., Pollard, W.E., et al. (1975). The sickness impact profile. Development of an outcome measure of health care. *Am J Public Health,* 65, 1304-1310.
Gold, M.R., Stevenson, D., & Fryback, D.G. (2002). HALYS and QALYS and DALYS, Oh My: similarities and differences in summary measures of population Health. *Annu Rev Public Health,* 23, 115-134.
Green, C., Brazier, J., & Deverill, M. (2000). Valuing health-related quality of life. A review of health state valuation techniques. *Pharmacoeconomics,* 17, 151-165.
Grogono, A.W., & Woodgate, D.J. (1971). Index for measuring health. *Lancet,* 2, 1024-1026.
Grosse, S.D. (2008). Assessing cost-effectiveness in healthcare: history of the $50,000 per QALY threshold. *Expert Rev Pharmacoecon Outcomes Res,* 8, 165-178.
Gudex, C. (1994a). *Standard gamble user manual: Props and self-completion method*. York: University of York
Gudex, C. (1994b). *Time trade-off user manual: Props and self-completion method.* York: University of York
Guerrero, A.M., & Herrero, C. (2005). A semi-separable utility function for health profiles. *J Health Econ,* 24, 33-54.
Guess, F.M. (1965). *A Cost/Effectiveness computer procedure for optimum allocation of fallout shelter system funds under uniform or variable risk assumptions*. Durham, North Carolina: Research Triangle Institute
Hernandez-Alava, M., Pudney, S., & Wailoo, A. (2018). *Quality review of a proposed EQ-5D-5L value set for England*: Universities of Sheffield and York
Ivarsson, B., Hesselstrand, R., Radegran, G., & Kjellstrom, B. (2019). Health-related quality of life, treatment adherence and psychosocial support in patients with pulmonary arterial hypertension or chronic thromboembolic pulmonary hypertension. *Chronic respiratory disease,* 16, 1479972318787906.
Janssen, M.F., Szende, A., Cabases, J., Ramos-Goni, J.M., Vilagut, G., & Konig, H.H. (2019). Population norms for the EQ-5D-3L: a cross-country analysis of population surveys for 20 countries. *Eur J Health Econ,* 20, 205-216.
Kahneman, D., & Tversky, A. (1979). Prospect Theory - Analysis of Decision under Risk. *Econometrica,* 47, 263-291.
Kahneman, D., & Tversky, A. (1984). Choices, Values, and Frames. *American Psychologist,* 39, 341-350.

Kaplan, R.M., & Ernst, J.A. (1983). Do Category Rating Scales Produce Biased Preference Weights for a Health Index? *Medical Care,* 21, 193-207.
Keeney, R.L., & Raiffa, H. (1976). *Decisions with multiple objectives, preferences and value trade-offs*. London: Wiley.
Kind, P., & Chuang, L.H. (2019). "A picture is worth a thousand words" : what can we learn from published 5L valuation studies? . Proceedings of the 36th Meeting of the EuroQol Group. Brussels, Belgium.
Kind, P., Hardman, G., & Macran, S. (1999). *UK population norms for EQ-5D*: Centre for Health Economics, University of York https://EconPapers.repec.org/RePEc:chy:respap:172chedp.
King, M.T., Costa, D.S., Aaronson, N.K., Brazier, J.E., Cella, D.F., Fayers, P.M., et al. (2016). QLU-C10D: a health state classification system for a multi-attribute utility measure based on the EORTC QLQ-C30. *Qual Life Res,* 25, 625-636.
King, M.T., Viney, R., Simon Pickard, A., Rowen, D., Aaronson, N.K., Brazier, J.E., et al. (2018). Australian Utility Weights for the EORTC QLU-C10D, a Multi-Attribute Utility Instrument Derived from the Cancer-Specific Quality of Life Questionnaire, EORTC QLQ-C30. *Pharmacoeconomics,* 36, 225-238.
Klarman, H.E. (1965). *The Economics of Health*. New York ; London: Columbia University Press.
Klarman, H.E. (1982). The road to cost-effectiveness analysis. *Milbank Mem Fund Q Health Soc,* 60, 585-603.
Lakdawalla, D.N., & Phelps, C.E. (2020). Health technology assessment with risk aversion in health. *J Health Econ,* 72, 102346.
Lancsar, E., Fiebig, D.G., & Hole, A.R. (2017). Discrete Choice Experiments: A Guide to Model Specification, Estimation and Software. *Pharmacoeconomics,* 35, 697-716.
Lang, A.E. (2000). Surgery for Parkinson disease: A critical evaluation of the state of the art. *Arch Neurol,* 57, 1118-1125.
Lipman, S.A., Brouwer, W.B.F., & Attema, A.E. (2019). QALYs without bias? Nonparametric correction of time trade-off and standard gamble weights based on prospect theory. *Health Econ,* 28, 843-854.
Llewellyn-Thomas, H., Sutherland, H.J., Tibshirani, R., Ciampi, A., Till, J.E., & Boyd, N.F. (1984). Describing health states. Methodologic issues in obtaining values for health states. *Medical Care,* 22, 543-552.
Llewellyn-Thomas, H.A., Sutherland, H.J., & Thiel, E.C. (1993). Do Patients' Evaluations of a Future Health State Change When They Actually Enter That State? *Medical Care,* 31, 1002-1012.
Lloyd, A., & Quadri, N. (2008). Comparing alternative 5-level versions of the EQ-5D in patients and general population in the UK. Proceedings of the 25th Meeting of the EuroQol Group. Milan, Italy.
Loewenstein, G., & Prelec, D. (1992). Anomalies in Intertemporal Choice - Evidence and an Interpretation. *Quarterly Journal of Economics,* 107, 573-597.
Loewenstein, G.F., & Prelec, D. (1993). Preferences for Sequences of Outcomes. *Psychological Review,* 100, 91-108.
Loomes, G., & Sugden, R. (1982). Regret Theory - an Alternative Theory of Rational Choice under Uncertainty. *Economic Journal,* 92, 805-824.
Lorgelly, P.K., Lawson, K.D., Fenwick, E.A., & Briggs, A.H. (2010). Outcome measurement in economic evaluations of public health interventions: a role for the capability approach? *Int J Environ Res Public Health,* 7, 2274-2289.
Louviere, J.J., Hensher, D.A., & Swait, J.D. (2000). *Stated choice methods : analysis and application*. New York: Cambridge University Press.
MacKillop, E., & Sheard, S. (2018). Quantifying life: Understanding the history of Quality-Adjusted Life-Years (QALYs). *Soc Sci Med,* 211, 359-366.
Manski, C.F. (2001). Daniel McFadden and the econometric analysis of discrete choice. *Scandinavian Journal of Economics,* 103, 217-229.
Masters, R., Anwar, E., Collins, B., Cookson, R., & Capewell, S. (2017). Return on investment of public health interventions: a systematic review. *J Epidemiol Community Health,* 71, 827-834.
McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In M. Balch, & S. Wu (Eds.), *Essays on economic behavior under uncertainty*. Amsterdam: North Holland.
McHorney, C.A. (1999). Health status assessment methods for adults: past accomplishments and future challenges. *Annu Rev Public Health,* 20, 309-335.
Mehrez, A., & Gafni, A. (1989). Quality-adjusted life years, utility theory, and healthy-years equivalents. *Med Decis Making,* 9, 142-149.

Metman, L.V., & Kompoliti, K. (2010). *Encyclopedia of movement disorders*. Kidlington, Oxford, UK ; San Diego, CA: Academic Press.
Miyamoto, J.M., & Eraker, S.A. (1985). Parameter estimates for a QALY utility model. *Med Decis Making,* 5, 191-213.
Mulhern, B., Norman, R., Street, D.J., & Viney, R. (2019). One Method, Many Methodological Choices: A Structured Review of Discrete-Choice Experiments for Health State Valuation. *Pharmacoeconomics,* 37, 29-43.
Mushkin, S.J. (1958). Toward a definition of health economics. *Public health reports (Washington, D.C. : 1896),* 73, 785-793.
National Institute for Health and Clinical Excellence (2012). *Methods for the development of NICE public health guidance (third edition)*: NICE https://www.nice.org.uk/process/pmg4/resources/methods-for-the-development-of-nice-public-health-guidance-third-edition-pdf-2007967445701.
National Institute for Health and Clinical Excellence (2018). *Developing NICE guidelines: The manual*: NICE https://www.nice.org.uk/process/pmg20/chapter/introduction-and-overview.
Neumann, P.J., Kim, D.D., Trikalinos, T.A., Sculpher, M.J., Salomon, J.A., Prosser, L.A., et al. (2018). Future Directions for Cost-effectiveness Analyses in Health and Medicine. *Medical Decision Making,* 38, 767-777.
Nord, E., Daniels, N., & Kamlet, M. (2009). QALYs: some challenges. *Value in Health,* 12 Suppl 1, S10-15.
Nord, E., Richardson, J., & Macarounas-Kirchmann, K. (1993). Social evaluation of health care versus personal evaluation of health states. Evidence on the validity of four health-state scaling instruments using Norwegian and Australian surveys. *Int J Technol Assess Health Care,* 9, 463-478.
Norman, R., Cronin, P., & Viney, R. (2013). A pilot discrete choice experiment to explore preferences for EQ-5D-5L health states. *Appl Health Econ Health Policy,* 11, 287-298.
Novick, D. (1968). The Origin and History of Program Budgeting. *California Management Review,* 11, 7-12.
Ogwulu, C.B., Jackson, L.J., Kinghorn, P., & Roberts, T.E. (2017). A Systematic Review of the Techniques Used to Value Temporary Health States. *Value in Health,* 20, 1180-1197.
Olsen, J.A. (1994). Persons vs years: two ways of eliciting implicit weights. *Health Econ,* 3, 39-46.
Oppe, M., Devlin, N.J., van Hout, B., Krabbe, P.F., & de Charro, F. (2014). A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value in Health,* 17, 445-453.
Owen, L., & Fischer, A. (2019). The cost-effectiveness of public health interventions examined by the National Institute for Health and Care Excellence from 2005 to 2018. *Public Health,* 169, 151-162.
Packer, A.H. (1968). Applying Cost-Effectiveness Concepts to the Community Health System. *Operations Research,* 16, 227-253.
Patrick, D.L., Bush, J.W., & Chen, M.M. (1973). Toward an operational definition of health. *J Health Soc Behav,* 14, 6-23.
Pearson, S.D. (2019). Why the Coming Debate Over the QALY and Disability Will be Different. *J Law Med Ethics,* 47, 304-307.
Peasgood, T., Mukuria, C., Carlton, J., Connell, J., Devlin, N., Jones, K., et al. (2021). 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? *Eur J Health Econ,* 22, 1067-1081.
Peeters, Y., & Stiggelbout, A.M. (2009). Valuing health: does enriching a scenario lead to higher utilities? *Med Decis Making,* 29, 334-342.
Persson, U., Norinder, A., Hjalte, K., & Gralen, K. (2001). The value of a statistical life in transport: Findings from a new contingent valuation study in Sweden. *Journal of Risk and Uncertainty,* 23, 121-134.
Pinto-Prades, J.L., & Rodriguez-Miguez, E. (2015). The lead time tradeoff: the case of health states better than dead. *Med Decis Making,* 35, 305-315.
Pinto-Prades, J.L., Sanchez-Martinez, F.I., Abellan-Perpinan, J.M., & Martinez-Perez, J.E. (2018). Reducing preference reversals: The role of preference imprecision and nontransparent methods. *Health Econ,* 27, 1230-1246.
Pliskin, J.S., Shepard, D.S., & Weinstein, M.C. (1980). Utility-Functions for Life Years and Health-Status. *Operations Research,* 28, 206-224.
Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science,* 306, 462-466.

Ratcliffe, J., Stevens, K., Flynn, T., Brazier, J., & Sawyer, M. (2012). An assessment of the construct validity of the CHU9D in the Australian adolescent general population. *Qual Life Res,* 21, 717-725.
Robinson, A., & Spencer, A. (2006). Exploring challenges to TTO utilities: valuing states worse than dead. *Health Econ,* 15, 393-402.
Robinson, A., Spencer, A., & Moffatt, P. (2015). A framework for estimating health state utility values within a discrete choice experiment: modeling risky choices. *Med Decis Making,* 35, 341-350.
Robinson, A., Spencer, A.E., Pinto-Prades, J.L., & Covey, J.A. (2017). Exploring Differences between TTO and DCE in the Valuation of Health States. *Med Decis Making,* 37, 273-284.
Robinson, S., & Bryan, S. (2013). Does the process of deliberation change individuals' health state valuations? An exploratory study using the person trade-off technique. *Value in Health,* 16, 806-813.
Rodrigues, S.P., van Eck, N.J., Waltman, L., & Jansen, F.W. (2014). Mapping patient safety: a large-scale literature review using bibliometric visualisation techniques. *BMJ Open,* 4, e004468.
Rosser, R.M., & Kind, P. (1978). A scale of valuations of states of illness: is there a social consensus? *Int J Epidemiol,* 7, 347-358.
Rosser, R.M., & `Watts, V.C. (1972). The measurement of hospital output. *Int J Epidemiol,* 1, 361-368.
Ryan, M., & Hughes, J. (1997). Using conjoint analysis to assess women's preferences for miscarriage management. *Health Econ,* 6, 261-273.
Sen, A. (1979). Utilitarianism and Welfarism. *Journal of Philosophy,* 76, 463-489.
Shah, K.K., Mulhern, B., Longworth, L., & Janssen, M.F. (2017). Views of the UK General Public on Important Aspects of Health Not Captured by EQ-5D. *Patient,* 10, 701-709.
Sintonen, H. (1981). An approach to measuring and valuing health states. *Soc Sci Med Med Econ,* 15, 55-65.
Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science,* 24, 265-269.
Soekhai, V., de Bekker-Grob, E.W., Ellis, A.R., & Vass, C.M. (2019). Discrete Choice Experiments in Health Economics: Past, Present and Future. *Pharmacoeconomics,* 37, 201-226.
Spencer, A. (2003). The TTO method and procedural invariance. *Health Econ,* 12, 655-668.
Stalmeier, P.F., Bezembinder, T.G., & Unic, I.J. (1996). Proportional heuristics in time tradeoff and conjoint measurement. *Med Decis Making,* 16, 36-44.
Stiggelbout, A.M., Kiebert, G.M., Kievit, J., Leer, J.W., Stoter, G., & de Haes, J.C. (1994). Utility assessment in cancer patients: adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. *Med Decis Making,* 14, 82-90.
Stinnett, A.A., & Paltiel, A.D. (1996). Mathematical programming for the efficient allocation of health care resources. *J Health Econ,* 15, 641-653.
Stolk, E.A., Oppe, M., Scalone, L., & Krabbe, P.F. (2010). Discrete choice modeling for the quantification of health states: the case of the EQ-5D. *Value in Health,* 13, 1005-1013.
Sutherland, H.J., Llewellyn-Thomas, H., Boyd, N.F., & Till, J.E. (1982). Attitudes toward quality of survival. The concept of "maximal endurable time". *Med Decis Making,* 2, 299-309.
Taheri, S., Asadi, S., Nilashi, M., Ali Abumalloh, R., Ghabban, N.M.A., Mohd Yusuf, S.Y., et al. (2021). A literature review on beneficial role of vitamins and trace elements: Evidence from published clinical studies. *J Trace Elem Med Biol,* 67, 126789.
The King's Fund (2021). *Long-term conditions and multi-morbidity*: https://www.kingsfund.org.uk/projects/time-think-differently/trends-disease-and-disability-long-term-conditions-multi-morbidity.
Thurstone, L.L. (1928). An Experimental Study of Nationality Preferences. *The Journal of General Psychology,* 1, 405-425.
Torrance, G.W. (1970). *A generalized cost-effectiveness model for the evaluation of health programs*. Ontario, Canada: Faculty of Business, McMaster University https://macsphere.mcmaster.ca/handle/11375/5559.
Torrance, G.W. (1986). Measurement of health state utilities for economic appraisal. *J Health Econ,* 5, 1-30.
Torrance, G.W., Feeny, D.H., Furlong, W.J., Barr, R.D., Zhang, Y., & Wang, Q. (1996). Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. *Medical Care,* 34, 702-722.
Torrance, G.W., Thomas, W.H., & Sackett, D.L. (1972). A utility maximization model for evaluation of health care programs. *Health services research,* 7, 118-133.

Treadwell, J.R. (1998). Tests of preferential independence in the QALY model. *Med Decis Making,* 18, 418-428.
Treasury, H.M. (2018). *The Green Book: Central Government Guidance on Appraisal and Evaluation*. London: The Stationary Office https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/685903/The_Green_Book.pdf.
Treasury, H.M. (2020). *Central Magenta Book: Central Government guidance on evaluation*. London: The Stationary Office.
Tsevat, J., Goldman, L., Lamas, G.A., Pfeffer, M.A., Chapin, C.C., Connors, K.F., et al. (1991). Functional status versus utilities in survivors of myocardial infarction. *Medical Care,* 29, 1153-1159.
Tsevat, J., Goldman, L., Soukup, J.R., Lamas, G.A., Connors, K.F., Chapin, C.C., et al. (1993). Stability of time-tradeoff utilities in survivors of myocardial infarction. *Med Decis Making,* 13, 161-165.
Tversky, A., & Kahneman, D. (1992). Advances in Prospect-Theory - Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty,* 5, 297-323.
van Osch, S.M., Wakker, P.P., van den Hout, W.B., & Stiggelbout, A.M. (2004). Correcting biases in standard gamble and time tradeoff utilities. *Med Decis Making,* 24, 511-517.
Viney, R., Norman, R., Brazier, J., Cronin, P., King, M.T., Ratcliffe, J., et al. (2014). An Australian discrete choice experiment to value eq-5d health states. *Health Econ,* 23, 729-742.
Viscusi, W.K., & Aldy, J.E. (2003). The value of a statistical life: A critical review of market estimates throughout the world. *Journal of Risk and Uncertainty,* 27, 5-76.
von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, N.J.: Princeton University Press.
Walsh, J. (1963). Civil Defense: Housing Reverses Direction and Approves Fallout Shelter Program, Sequel Pending. *Science,* 141, 1264-1265.
Waltman, L., van Eck, N.J., & Noyons, E.C.M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics,* 4, 629-635.
Weinstein, M.C., Torrance, G., & McGuire, A. (2009). QALYs: the basics. *Value in Health,* 12 Suppl 1, S5-9.
Williams, A. (1985). Economics of coronary artery bypass grafting. *Br Med J (Clin Res Ed),* 291, 326-329.
Williams, A. (2005). Discovering the QALY, or how Rachel Rosser changed my life. In A. Oliver (Ed.), *Personal Histories in Health Research* pp. 191-206). London: The Nuffield Trust.
World Health Organization (2020). *Overview - Preventing chronic diseases: a vital investment* https://www.who.int/chp/chronic_disease_report/part1/en/index11.html.
Wright, D.R., Wittenberg, E., Swan, J.S., Miksad, R.A., & Prosser, L.A. (2009). Methods for measuring temporary health States for cost-utility analyses. *Pharmacoeconomics,* 27, 713-723.
Zakaria, F. (2020). *Ten lessons for a post-pandemic world*. London: Penguin Books.
Zanchetti, A., Thomopoulos, C., & Parati, G. (2015). Randomized controlled trials of blood pressure lowering in hypertension: a critical reappraisal. *Circ Res,* 116, 1058-1073.

Figure 1: Uptake of publications over time of Torrance et al (1972) or Fanshel and Bush (1970) over time

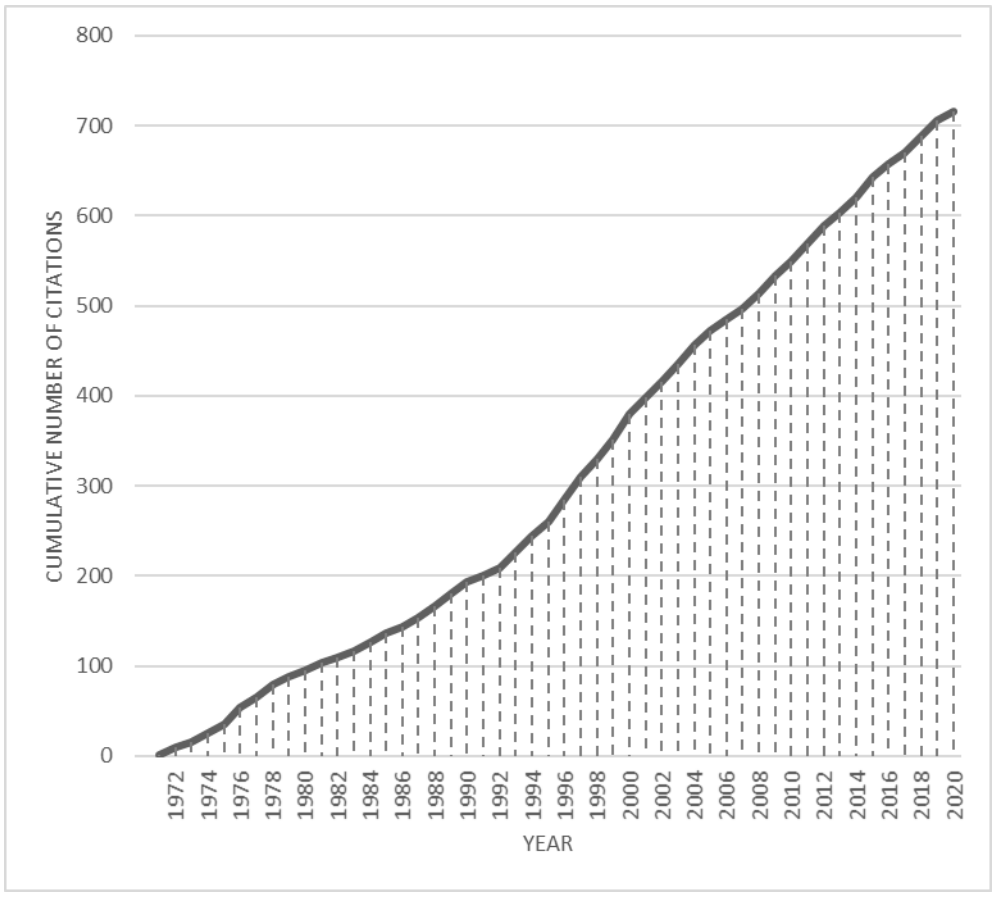
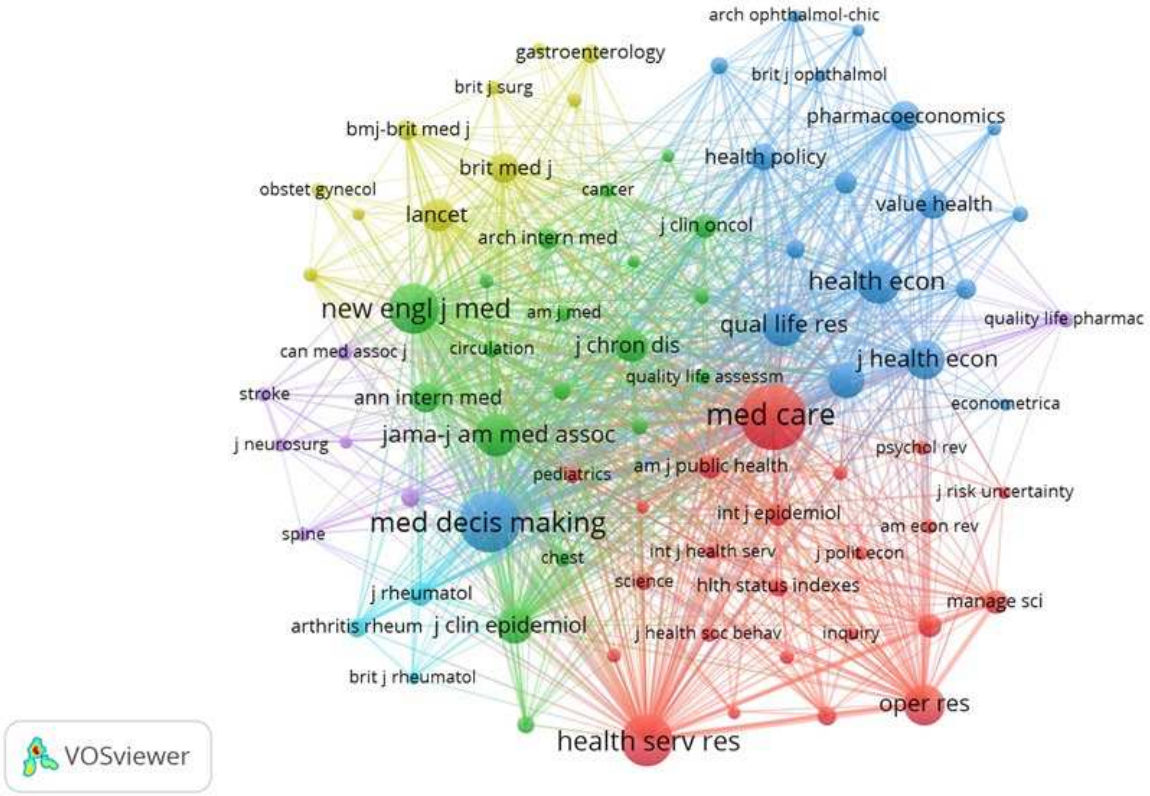
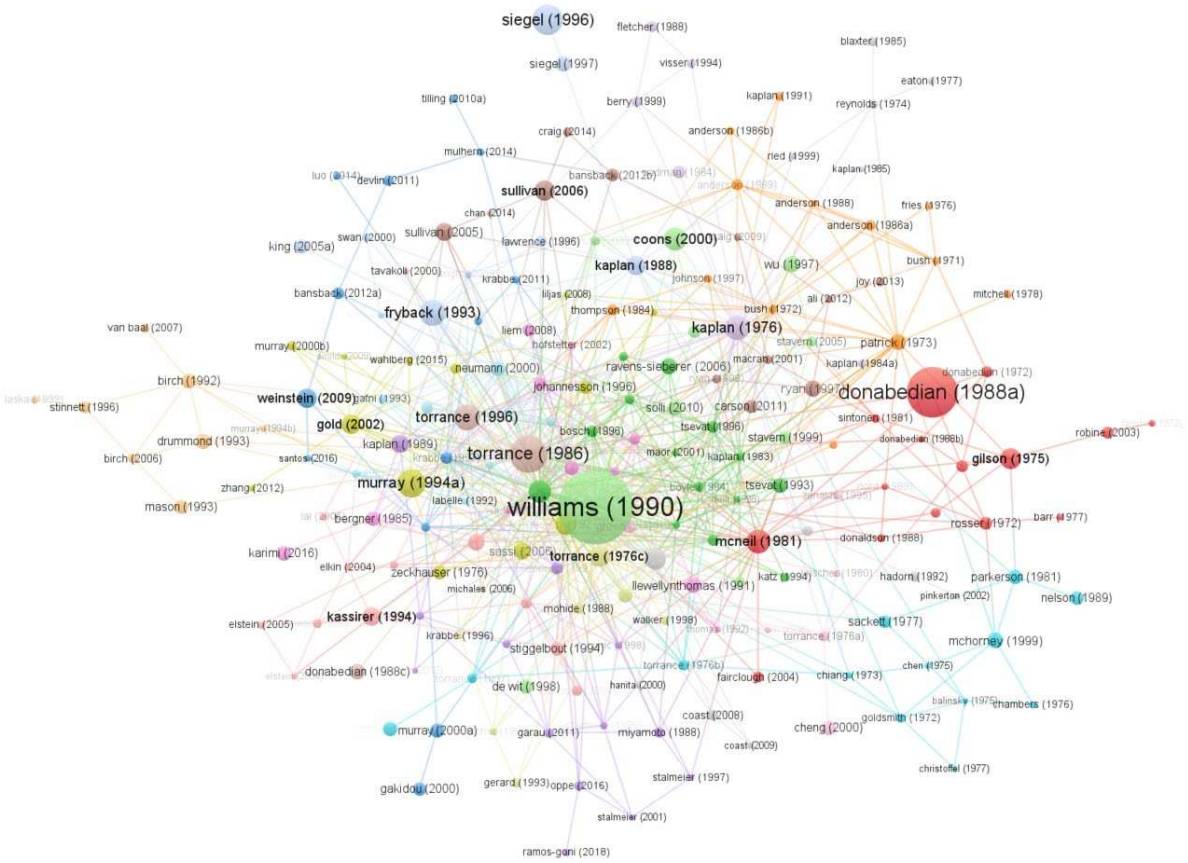


Figure 2: Journal co-citation network visualization map



Each journal is represented by a node on the map. Larger journal nodes cover a larger proportion of the total citations and/or co-citations. Links between the journal nodes indicate co-citation. The more frequently journals co-cite one another, the closer together the journal nodes are positioned on the map, forming clusters of journals highlighted by VOSviewer in colour.

Figure 3: Publication co-citation network visualization map



Each publication is represented by a node and labelled by the first author. Larger nodes and labels cover a larger proportion of the total citations and/or co-citations. For some items the label may not be displayed due to the small number of citations. The more frequently publications co-cite one another, the closer together the publication nodes are positioned on the map, forming clusters of publications highlighted by VOSviewer in colour. Publications within the same cluster are assigned the same colour. Williams 1990 refers to the reference: EuroQol Group. EuroQol - a new facility for the measurement of health-related quality of life. Health Policy, 1990. 16: 199-208.

Figure 4: Behavioural theories that have influenced HEs over the past 50 years