Applied Health Economics and Health Policy
https://doi.org/10.1007/s40258-024-00919-8

REVIEW ARTICLE

The Short Form 6 Dimensions (SF-6D): Development and Evolution

Clara Mukuria[1] · Donna Rowen[1] · Brendan Mulhern[2] · Emily McDool[1] · Samer Kharroubi[3] · Jakob B. Bjorner[4] · John E. Brazier[1]

Accepted: 10 September 2024

Abstract
This paper considers the development and evolution of the short-form 6 dimensions (SF-6D), a generic preference-weighted measure consisting of a health classification with accompanying value set that was developed from one of the widest used health related quality of life measures, the SF-36 health survey. This enabled health state utility values to be directly generated from SF-36 and SF-12 data for a range of purposes, including to produce quality adjusted life years for use in economic evaluation of healthcare interventions across a range of different conditions and treatments. This paper considers the rationale for the development of the measure, the development process, performance and how the SF-6D has evolved since its conception. This includes the development of an updated version, SF-6D version 2 (SF-6Dv2), which was generated to deal with some criticisms of the first version, and now includes a standalone version for inclusion in studies without relying on use of SF-36 or SF-12. Valuation methods have also evolved, from standard gamble in-person interviews to online discrete choice experiment surveys. International work related to the SF-6Dv1 and SF-6Dv2 is considered. We also consider recommendations for use, highlighting key psychometric evidence and reimbursement agency recommendations.

1 Introduction

The short-form 6 dimensions (SF-6D) is a health classification that was developed from the SF-36 health survey (SF-36, formerly referred to as short form 36), which is one of the most widely used health related quality of life (HRQoL) measures internationally (e.g. see Haraldstad et al. [1], Pequeno et al. [2] and Siette et al. [3]) [4–7]. The SF-6D has been valued to generate utilities that reflect members of the public's preferences for health on the zero to one utility scale (dead to full health) where values below zero indicate a health state that is considered worse than being dead. The SF-6D utilities can then be combined with length of life to generate quality adjusted life years (QALYs), an overall metric reflecting both quality of life and length of life [8]. QALYs are used in comparative analysis undertaken as part of cost effectiveness analysis (CEA) of healthcare interventions to inform decision making [9].

Key Points for Decision Makers

The Short form 6 dimensions (SF-6D) is a health classification system that was developed from profile generic measures of health-related quality of life, the SF-36 health survey (SF-36) and SF-12 health survey (SF-12), which enabled utilities to be generated from these measures.

An innovative approach was taken to develop the SF-6D, and this has evolved over time to reflect advances in generating new instruments and value sets which have been applied in the creation of SF-6D version 2 (SF-6Dv2) including a standalone version, the SF-6Dv2 health utility survey.

There are now several SF-6D value sets available from different countries around the world and these are likely to increase.

✉ Clara Mukuria
c.mukuria@sheffield.ac.uk

1 Division of Population Health, School of Medicine and Population Health, University of Sheffield, Sheffield, UK

2 Centre for Health Economics Research and Evaluation, University of Technology Sydney, Sydney, Australia

3 Department of Nutrition and Food Sciences at the American University of Beirut, Beirut, Lebanon

4 QualityMetric, IQVIA, Johnston, USA

Published online: 26 October 2024

The SF-6D enables utilities to be generated from any study that has used the SF-36 or SF-12 without new data collection or additional questionnaires. The SF-6D can be generated for all versions of the SF-36 and SF-12. This paper aims to describe the development of the SF-6D, how the SF-6D has evolved over time and to consider when the SF-6D and different versions should be used. This paper provides an important overview for both current and future users of the measures and those interpreting results from the measures across a range of different applications. In this paper, the first version is referred to as SF-6Dv1 to distinguish it from subsequent versions. The term SF-6D refers to both versions.

2 Initial Development of the SF-6Dv1

2.1 Why was the SF-6Dv1 Developed?

The SF-6Dv1 was developed from the SF-36v1 [4, 7]. The SF-36 is a profile measure that covers eight health domains: physical functioning, bodily pain, role limitations which relate to physical health, role limitations related to emotional problems, mental health, social functioning, energy/fatigue and general health perceptions [10]. In total, seven of the SF-36 domains are included in the SF-6D (role limitation domains are combined, and general health is excluded). The SF-36 and the related RAND-36 health survey 1.0 were initially developed in 1990 drawing on evidence generated in the Medical Outcome Study, a large-scale study of patients in the USA, which aimed to assess health care use and the health and wellbeing of participants [10, 11]. The two versions have the same domains and items, though item phrasing varies for some items and the measures are scored differently and were distributed by different organisations. We focus here on the SF-36.

The SF-36 was developed with a standard 4 week recall period and an acute 1 week recall period version [12] (a 24-h recall period version is now also available). As a profile measure, the SF-36 provides a score for each of the eight health domains. Principal components analysis of the eight subscales from the SF-36 in different datasets indicated that a large proportion of the variation could be explained by two distinct physical and mental health components and therefore two component summary scores, physical and mental component summary scores (PCS and MCS), were developed which relied on weighted subscale scores [12, 13]. In scoring PCS and MCS, each subscale is weighted so that the resulting principal components capture the largest possible part of the subscale variation. Further, PCS and MCS were scored using T scores (mean = 50, standard deviation = 10) on the basis of the US general population in 2009. The SF-12, a shorter version of the SF-36 was also developed on the back of evidence that the precision of PCS and MCS could be maintained with fewer questions. The work to develop the SF-12 was based on identifying which items could be used to generate PCS and MCS scores that were similar to the scores from the SF-36 [12].

Evidence on the validity and reliability of the SF-36 has been shown in many different populations and much work was done to translate the measure into different languages [14, 15]. Many studies have found that PCS and MCS are sufficient for tracking changes in different groups [16]. This resulted in wide use of the SF-36 in clinical trials and observational studies. Therefore, an approach to generate utilities from it was desirable. While subscale and summary scores are useful in identifying whether the burden of poor HRQoL lies either at the specific domain level or whether it is physical, mental health or both, these scores are less useful for economic analyses for two reasons. The first is that although CEA to inform decision-making can potentially use a single domain or either PCS or MCS as the outcome measure, none of these single scores provide an overall single assessment of HRQoL [9]. Without a single overall score, it can be difficult to assess overall change, especially where there are simultaneous improvements and deteriorations for different domains. Secondly, the profile scores are not preference-based therefore they do not provide information on which of the domains have the largest impact based on individuals' preferences for use in QALY estimation. The SF-6Dv1 classification system was therefore developed to allow utilities to be generated to estimate QALYs. These utilities could then be generated for any dataset containing the SF-36 [4, 7] or the SF-12 [5].

2.2 Who Developed the SF-6Dv1?

Development of the SF-6Dv1 was led by Prof John Brazier at the University of Sheffield and was initiated as a result of discussions with users of the SF-36 at meetings of the International Quality of Life Assessment Group (IQOLA) in the early 1990s [17]. The first attempt was undertaken as part of a PhD by Brazier with colleagues Tim Usherwood, Rosemary Harper and Kate Thomas at the University of Sheffield who provided advice on the design and conduct of the study. The final design, interviews and all analyses were undertaken by John Brazier [4]. This was a small-scale study using a convenience sample and provided a proof of concept that it was feasible to derive a preference-based measure from the SF-36. This resulted in a proposal to undertake a larger scale study using a representative sample of the UK population to provide the valuation data that was funded by GlaxoWelcome in 1999. Co-investigators included Jeniffer

Roberts and Mark Deverill, with advice from John Ware and Barbara Gandek at Tufts University [5, 7].

2.3 How was the SF-6Dv1 Developed?

The development of the SF-6Dv1 classification system and valuation is described in detail in Brazier et al. [7] and Brazier and Roberts [5]. In brief, the aim was to identify a set of questions from SF-36v1 that were amenable to valuation. Selecting questions was necessary as it was not considered feasible or desirable to value all the questions. As the SF-36v1 has more than one question for most of its domains, Brazier and colleagues proposed a set of criteria to inform the selection. The criteria included selecting: (1) only one item for closely related items to avoid redundancy; (2) where possible, negatively phrased items instead of positively phrased ones as they were more relevant in the context of health care services, which are designed to reduce the negative impact of ill health; and (3) items that were most preferred by people (where this was available) drawing from evidence on international translation work undertaken for the SF-36v1. The item selection process was also informed by the factor analysis that was used to develop the SF-12v1 and this helped to identify how each item contributed to each domain [18].

The SF-36v1 has one domain related to general health perception which was not included in the SF-6D as it was considered 'illogical' to include this in a classifier aimed at generating preferences for the components that make up overall health [4]. Initial work used 20 items from the SF-36v1 to construct a health classification [4] but the final version of SF-6Dv1 used 11 items, 8 that were drawn from the SF-12v1, with an additional 3 from the SF-36v1. The six dimensions included questions covering physical functioning (combining three questions to cover the severity range of questions in this dimension), role limitation (combined two items related to physical and emotional health), mental health (combining two items for depression and anxiety), social functioning, pain and energy/fatigue(each covered by one item) [7]. Some of the questions in SF-36v1 (and SF-12v1) related to mental health, social functioning and vitality had six response levels including 'a good bit of the time' which was not included in the classifier. Instead, these responses were reallocated randomly to adjacent levels. The SF-6Dv1 classifier response levels ranged from four to six and combined with the dimensions, describes 18,000 health states. Individuals who have completed all 11 questions from the SF-36v1 can be assigned to the SF-6Dv1 classifier.

2.4 How was the SF-6Dv1 Valued?

The valuation survey of SF-6Dv1 in the United Kingdom (UK) selected 49 health states out of 18,000 using an orthogonal array which was the minimum number required to enable an additive model to be estimated to enable the generation of utilities for all 18,000 SF-6Dv1 health states. An additional 200 states were selected across mild, moderate and severe states using a stratified sampling method to enable the estimation of more complex models [7]. In an interviewer-administered exercise, respondents ranked five health states along with the worst state and immediate death. This was followed by valuation of the five intermediate states using standard gamble (SG) where participants were asked to choose between the certain prospect of living in the state or the uncertain prospect of the best state defined by the SF-6D or the worst state using props developed at McMaster [19]. The probability in the uncertain prospect was varied using a ping-pong approach (high probability followed by low then high) until the respondent was indifferent between the certain and uncertain prospect. Separately, the worst state was valued against "immediate death" with the SG task depending on whether the respondent thought the worst state was better or worse than dead. This final step allowed the five states to be placed on the dead to full health QALY scale. There were 836 interviews conducted in a nationally representative sample for age and sex but 225 were excluded for either failing to value the worst state, failing to value more than two or more states or valuing all the states the same. Additive models with interactions were assessed using ordinary least squares (OLS) and random or fixed effects models to account for multiple responses for each individual. Mean level models were also estimated.

The recommended model restricted the constant to one and used the mean level model estimates [7]. This model had logical inconsistencies for some of the dimensions, meaning that as health deteriorated utility increased. This created issues for use in decision making, and subsequently a revised model was recommended for use that was logically consistent, meaning that as health deteriorated utility did not increase [5]. The final value set ranged from 0.301 to 1, with no values considered to be worse than dead. This value set enables SF-6Dv1 utilities to be generated from SF-36v1 (any recall period) where there is no missing data for the questions used to generate the classifier.

2.5 How was the SF-6Dv1 Made Available?

SF-6Dv1 was licensed both by QualityMetric and University of Sheffield where the SF-6Dv1 was free for non-commercial use (including academic, publicly funded healthcare systems and not-for-profit organisations) with a cost payable for commercial use (note that licensing requests to use SF-6D are now only made via QualityMetric, see Sect. 5.3).

3 Evolution of the SF-6D

3.1 Estimating SF-6Dv1 from SF-12

SF-6Dv1 was initially developed from the SF-36v1. However, the SF-12v1 was also available and it had the added advantage of being brief without substantial loss of information compared with the SF-36v1 [20]. It was therefore considered useful to consider an approach to generate utility values from the SF-12v1. Similar steps to those applied to the SF-36v1 were applied to the SF-12v1, i.e. the development of a classification system, valuation of a subset of health states and then generation of utility values for the whole measure [5]. Using this approach, a new version of the SF-6Dv1 referred to as SF-6Dv1 (SF-12) was developed and published in 2004 [5]. It has the same dimensions as SF-6Dv1 and draws on seven items from SF-12v1 which are also in the SF-6Dv1, and it defines 7500 states. The dimensions were the same across the two versions, but physical functioning and pain were simplified in SF6Dv1 (SF-12) with two physical functioning items used instead of the three used from the SF-36, and one level of pain was removed (pain that does not interfere with your normal work). Given the overlap between SF-6Dv1 and SF-6Dv1 (SF-12), the same valuation survey was used with a subset of the states (241) used to model the results for SF-6Dv1 (SF-12). Mean level models were estimated. There were logical inconsistencies, and these were removed by combining adjacent levels. The SF-6Dv1 (SF-12) coefficients were larger than the SF-6Dv1 coefficients for all dimensions apart from for physical functioning and pain [5].

3.2 Modifications to Align with Developments in the SF-36

The SF-36v1 was revised in 1998 to improve the measure following evidence from early use of the measures [20]. Development of the new version, the SF-36v2, included reduction of response levels from six to five for items which included 'a good bit of the time' which was dropped, and an increase in response levels for items in two domains (role physical and role emotional) from two to five. Similar changes were implemented in the SF-12 to generate SF-12v2. Users of SF-36v2 (and the related SF-12v2) were still interested in generating SF-6D utilities. The main adjustment needed to score the SF-6Dv1 from the SF-36v2 or SF-12v2 is the collapsing of the five response categories for role physical and role emotional into two levels. Five response levels were already in use for the SF-6Dv1 classifier with random reallocation to adjacent levels for items that included 'a good bit of the time'. This allows SF-6Dv1 utilities to be generated for all versions of SF-36 and SF-12.

3.3 Generating SF-6Dv1 When Data is Missing in SF-36v1 or SF-12v1

Since SF-6D requires responses to a subset of SF-36 or SF-12 items, it can be generated when there are missing responses to other items outside of this subset. The general principle for handling missing data for the SF-6Dv1 is that missing data can be allowed for an item, if the item response in the particular situation has no impact on the overall score. For example, in the physical functioning dimension where there are three items, if the response to the item PF10 ('bathing and dressing yourself') is 1 ('yes, limited a lot') the overall physical function dimension score will be 5 regardless of the response to the two other physical functioning items. Thus, for this particular situation, these two items can be allowed to be missing.

3.4 Developments in Valuation—Bayesian Approaches

Health state values pose a significant challenge for conventional statistical modelling procedures owing to their nature, namely: skewed, truncated, non-continuous and hierarchical. While attempts to model these data have shown some success with instruments including EQ-5D, SF-6D, and Health Utilities Index 2 (HUI2) [7, 21, 22], concerns persisted regarding the size of prediction errors. More specifically, SF-6Dv1 exhibits non-monotonicity, where certain better states are assigned lower values than worse ones, alongside systematic patterns in prediction errors, including overestimation of poor health state values and underestimation of good health state values. Moreover, these models have limitations in capturing the effects of covariates on health state values.

Traditionally, the conventional statistical models used in previous analyses have been frequentist. An alternative to modelling health state valuation data is a nonparametric Bayesian approach. This approach offers enhanced flexibility and realistic inference compared with standard frequentist models in addition to being more flexible in capturing of covariate effects. Kharroubi et al. [23–25] introduced a nonparametric Bayesian model to model SG data generated in previous studies, notably applied to UK SF-6Dv1 resulting in a different SF-6Dv1 UK value set [25]. The differences between non-parametric Bayesian and parametric frequentist models are potentially important. For the UK SF-6Dv1, the differences in average health state values between the two models ranged from 0.01 for the mildest state through to 0.25 for the worst health state, with an average of 0.11 averaged across the 249 states that were valued [25]. This Bayesian model has been applied to other health state measures including HUI2 [26] and EQ-5D [27].

3.5 SF-6Dv1 International Developments

SF-36 and SF-12 are used internationally and differences between countries mean that it is necessary to generate preferences that are population specific. Many health technology assessment (HTA) agencies recommend the use of value sets using the preferences of their country population [28]. Therefore, utilities for the SF-6Dv1 have been generated for countries other than the UK. Furthermore, SF-6D value sets are accepted by a number of reimbursement agencies around the world [29, 30].

The UK value set for the SF-6Dv1 has been highly influential, and the SG valuation approach has also been used to estimate country specific value sets across three continents. This includes Asian value sets in China (Hong Kong) [31], Japan [32] and Lebanon [33], European value sets in Portugal [34] and Spain [35, 36], and a South American value set in Brazil [37].

The SG valuation method, used for the SF-6Dv1, is often regarded as being theoretically superior as a health state valuation method owing to its basis in expected utility theory, but has been criticised for a number of reasons including practical concerns (see [9] for an overview). These include concerns about respondent understanding of the tasks owing to the complexity of the iterative probability trade-off. Another key feature of SG value sets is the relatively mild values assigned to poor health states as a result of respondent risk aversion. The valuation task also required a 2-stage chained process with states being valued against full health and the worst state, followed by valuation of the worst state in comparison with both full health and dead. It has been suggested that this increases the likelihood of higher values owing to risk aversion. In recent years the use of discrete choice experiment (DCE) methods has increased in popularity [38, 39], and three SF-6Dv1 value sets using DCE approaches have been developed in Australia [40], the Netherlands [41] and the USA [42]. Other methods, including a probability lottery equivalent adaptation of SG [36], have also been used to develop a Spanish SF-6Dv1 (SF-12) value set [35].

There has been more variation in the modelling approaches used in international studies [39], including exploration of an inverse probability weighting technique [43]. Bayesian modelling approaches have also been applied for other countries including China (Hong Kong) [25], Japan [32] and Lebanon [44]. Bayesian modelling has also been extended to handle combined datasets such, as UK/China, UK/Japan and UK/Lebanon SF-6D data [45–47], incorporating richer structures for covariate effects. The objective of combining datasets is to demonstrate a powerful approach for analysing data from two distinct nationalities or ethnic groups, aiming to discern and potentially estimate underlying utility functions more efficiently. A key innovation in this analysis involves utilizing the covariate framework of a Bayesian model to represent the differences between the two countries. This approach offers dual benefits. Firstly, in scenarios where ample data from both countries are available, the Bayesian model recognizes the primary differences in how individuals from each country value health [45, 46]. Secondly, in situations where one country has plenty of data while the other country's data are limited, a combined analysis may yield more accurate estimations of the latter's population utility functioning compared with separate analyses [47]. By leveraging insights from the first country, this analysis enables us to reduce sample sizes in the second country, achieving comparable accuracy to that of a full-scale study. This approach will be hugely important in countries with smaller settings and/or low- and middle-income countries (LMICs) lacking the capacity to conduct large-scale evaluation exercises, thus facilitating the development of localized value sets [47].

4 SF-6D Version 2 (SF-6Dv2)

4.1 Rationale for Development

The SF-6Dv1 has received criticism on the basis of both the classification system and valuation technique. It has been argued that there is ambiguity around the ordering of severity levels in the physical functioning dimension, including 'a lot' of limitations in moderate activities and 'a little' limitation in bathing and dressing [7]. The role dimension has also raised concerns owing to the 'floor effect' whereby a high proportion of individuals report the lowest severity level [48] and owing to insensitivity as a result of four levels being associated with only two unique utility decrements [6]. Additionally, the vitality dimension is positively framed while the other dimensions are negatively framed and there are concerns that this contrast in framing may have caused confusion amongst participants in valuation studies [6].

As described in Sect. 3.5, concerns have been raised about the valuation methods utilised for the SF-6Dv1 which led to issues with the estimated range of values. The UK value set in particular suffered from high values for severe health states. In the value set, some adjacent severity levels were combined for some dimensions to have the same utility decrement to avoid a logical inconsistency in health state values, and this resulted in a reduced number of utility values and decreasing sensitivity [9, 49].

4.2 SF-6Dv2 Classification

To address the concerns with SF-6Dv1 that were outlined above, a new revised version of SF-6D, the SF-6Dv2, was developed from the SF-36v2 and published in 2020 [6]. The SF-6Dv2 classification system was based on new

psychometric analyses of two data sets: a large inpatient sample, the Health Outcome Data Repository ($n = 49{,}029$) recently discharged hospital patients [50] and an international sample from four countries (UK, Canada, Australia and USA), the multi instrument comparison dataset ($n = 5331$) of 'healthy public' respondents and respondents with a range of self-reported health conditions [51].

The analyses for the SF-6Dv2 used 30 items to identify the best possible classification system (excluding items on general health perception and health transition) [6]. Exploratory and confirmatory factor analyses identified a 6-factor model with factors similar to the six dimensions identified for the SF-6Dv1: physical functioning, role limitations, pain, vitality, social functioning and mental health. Positively formulated items for vitality and mental health (e.g. full of life and happy) were excluded from this model since these items tended to form their own factor.

Rasch model analyses were used to evaluate item fit, item measurement properties, and differential item functioning. Results from these analyses were used to inform item selection and development of the classification system. The items selected included three items on physical functioning (vigorous activities, moderate activities and bathing and dressing), two items on role limitations (accomplishing less owing to either physical or emotional health problems), one item on pain (severity of bodily pain), one item on vitality (worn out), one item on social functioning (limitations in social activities) and two items on mental health (depressed and very nervous). For the dimensions with one item (pain, vitality and social functioning), the classification system is based directly on the item response categories. For dimensions with two items, the classification is based on the item with the worst score [for example, if a person answers being very nervous 'a little of the time' (2) and depressed 'most of the time' (4), the overall mental health score is 4]. For the physical functioning dimension, where 64 response combinations are possible (27 if missing responses are not allowed), results from Rasch model analyses were used to assign results into five easily interpretable categories (see Table 1) [6].

The final SF-6Dv2 classification system is presented in Table 1 and compared with SF-6Dv1. Generally, the SF-6Dv2 classification system is more concisely worded than SF-6Dv1 and all items are formulated negatively. For physical functioning, the number of levels is reduced from six to five, since Rasch analysis showed that one SF-6Dv1 category (5: health limits you a little in bathing and dressing) overlapped with another category (4: health limits you a lot in moderate activities). For role limitations, the four SF-6Dv1 categories are expanded to five categories with a clearer rank order. In the social functioning dimension, categories are slightly simplified, but otherwise unchanged. The pain dimension has changed from interference to severity. For the mental health dimension, the indicator of anxiety has changed from tense to very nervous and the indicator for depression has changed from downhearted and low to depressed. Finally, the vitality dimension has changed from positively worded energy to worn out.

4.3 Valuation Approaches and Value Set Development

The UK value set was published alongside the updated classification system, and used a protocol based on 360 DCE tasks including a duration attribute presented as both health state pairs ($n = 300$), and triplets including two health states, and a third 'immediate death' option ($n = 60$) [49]. Data were collected online from 3014 respondents who were representative of the UK population in age and sex in 2015. The UK value set produced using the protocol ranges from 1 (for the best health state) to − 0.574 (for the worst), with 15.2% of health states valued as worse than dead (below 0).

Other value sets for the SF-6Dv2 have been developed internationally using a variety of preference elicitation approaches, and value sets for the SF-6Dv2 continue to be produced worldwide (Table 2 provides a summary). The UK DCE protocol has been used in Australia [52] from 3001 respondents which resulted in a wider range than the UK value set (1 to − 0.685, with 20% worse than dead). A value set for Iran has also been published using the protocol [53], and ranges from 1 to − 0.796. Across all three value sets using the protocol, pain has the largest overall decrement (from best to worst level) followed by mental health, physical functioning and social functioning. Vitality and role functioning have the smallest overall decrements, but the order differs between countries.

A series of studies to produce value sets using a variety of approaches have been conducted in mainland China. Wu et al. [54] used both DCE with a duration attribute and time trade-off (TTO) to value SF-6Dv2 in a general population sample in mainland China and recommend the value set derived from the TTO data as although both methods performed well, there were logical inconsistencies in the DCE data. The same team also estimated SF-6Dv2 values using Best Worst Scaling (BWS) methods [55].

The original DCE valuation protocol has generated value sets with face validity. One criticism of the protocol is that the DCE triplet tasks were not generated using an established DCE design method but were selected by the research team from the pool of 300 choice set pairs on the basis of the severity of the health states. To rectify this, an updated protocol has been developed that uses an established DCE design approach to generate three blocks of tasks. These include a 'core' set of 304 choice set pairs (design 1), 76 choice set pairs including health states that are commonly experienced in the general population (design 2), and 76

Table 1 Comparison of SF-6Dv2 and SF-6D health classification system

	SF-6Dv2	SF-6Dv1
Physical functioning	1. Limited in vigorous activities not at all 2. Limited in vigorous activities a little 3. Limited in moderate activities a little 4. Limited in moderate activities a lot 5. Limited in bathing and dressing a lot	1. Your health does not limit you in vigorous activities 2. Your health limits you a little in vigorous activities 3. Your health limits you a little in moderate activities 4. Your health limits you a lot in moderate activities 5. Your health limits you a little in bathing and dressing 6. Your health limits you a lot in bathing and dressing
Role limitations	1. Accomplish less than you would like none of the time 2. Accomplish less than you would like a little of the time 3. Accomplish less than you would like some of the time 4. Accomplish less than you would like most of the time 5. Accomplish less than you would like all of the time	1. No problems with your work or other daily activities as a result of your physical health or any emotional problems 2. Limited in the kind of work or other activities as a result of your physical health 3. Accomplish less than you would like as a result of emotional problems 4. Limited in the kind of work or other activities as a result of your physical health and accomplish less than you would like as a result of emotional problems
Social functioning	1. Social activities are limited none of the time 2. Social activities are limited a little of the time 3. Social activities are limited some of the time 4. Social activities are limited most of the time 5. Social activities are limited all of the time	1. Your health limits your social activities none of the time 2. Your health limits your social activities a little of the time 3. Your health limits your social activities some of the time 4. Your health limits your social activities most of the time 5. Your health limits your social activities all of the time
Pain	1. No pain 2. Very mild pain 3. Mild pain 4. Moderate pain 5. Severe pain 6. Very severe pain	1. You have no pain 2. You have pain but it does not interfere with your normal work 3. You have pain that interferes with your normal work a little bit 4. You have pain that interferes with your normal work moderately 5. You have pain that interferes with your normal work quite a bit 6. You have pain that interferes with your normal work severely
Mental health	1. Depressed or very nervous none of the time 2. Depressed or very nervous a little of the time 3. Depressed or very nervous some of the time 4. Depressed or very nervous most of the time 5. Depressed or very nervous all of the time	1. You feel tense or downhearted and low none of the time 2. You feel tense or downhearted and low a little of the time 3. You feel tense or downhearted and low some of the time 4. You feel tense or downhearted and low most of the time 5. You feel tense or downhearted and low all of the time
Vitality	1. Worn out none of the time 2. Worn out a little of the time 3. Worn out some of the time 4. Worn out most of the time 5. Worn out all of the time	1. You have a lot of energy all of the time 2. You have a lot of energy most of the time 3. You have a lot of energy some of the time 4. You have a lot of energy a little of the time 5. You have a lot of energy none of the time

Permission to use the measures should be obtained from QualityMetric

SF-6Dv1 short form 6 dimensions version 1, *SF-6Dv2* short form 6 dimensions version 2

triplets choice sets including health states and immediate death. This has been used to develop a US value set for the SF-6Dv2 [56]. Other SF-6Dv2 value set studies using the protocol are expected to emerge over time.

In New Zealand, an adaptive DCE approach [potentially all pairwise RanKings of all possible alternatives (PAPRIKA)] has been used to estimate an SF-6Dv2 value set [57]. A number of studies have also explored valuation of the SF-6Dv2 using different valuation methods in specific populations. For example, Dufresne et al. [58] used both DCE and TTO approaches to estimate values for a population group with food allergies. Toure et al. [59] use DCE and TTO to estimate value sets for samples of people with breast and colorectal cancer. Kouakou et al. [60] also use DCE and TTO approaches, and in addition generate willingness to pay estimates using contingent valuation and DCE for a general public sample in Quebec, Canada.

4.4 Comparing SF-6Dv1 and SF-6Dv2

Differences in the two classification systems are set out in Sect. 4.2. With variation in the number of levels in some dimensions (physical functioning and role limitation), the SF-6Dv1 describes 18,000 health states, whereas the SF-6Dv2 describes 18,750. As may be expected, there is evidence of strong correlations between the dimensions of the SF-6Dv1 and SF-6Dv2 [61]. The two classification systems

Table 2 SF-6Dv1 and SF-6Dv2 general population country and region-specific value sets

Country	SF-6Dv1		SF-6Dv2	
	Value set reference	Method	Value set reference	Method
Australia	Norman et al. [103]	DCE	Mulhern et al. [52]	DCE
Brazil	Cruz et al. [37]	SG		
Canada			Kouakou et al. [60] (Quebec)	DCE and TTO
China	Lam et al. [31] (Hong Kong)	SG	Wu et al. [54]	TTO, DCE, BWS
Iran			Darroudi et al. [53]	DCE
Japan	Brazier et al. [32]	SG		
Lebanon	Kharroubi et al. [33]	SG		
Netherlands	Jonker et al. [41] (SF-12 version)	DCE		
New Zealand			Sullivan et al. [57]	DCE (PAPRIKA)
Portugal	Ferreira et al. [34]	SG		
Spain	Martínez-Pérez [35] (SF-12 version)	PLE		
United Kingdom	Brazier and Roberts [5] Kharroubi et al. [23]	SG	Mulhern et al. [49]	DCE
USA	Craig et al. [42]	DCE		

SF-6Dv1 6Dv2 short form 6 dimensions version 1, *SF-6Dv2* short form 6 dimensions version 2, *DCE* discrete choice experiment, *DCE (PAPRIKA)* discrete choice experiment (potentially all pairwise rankings of all possible alternatives), *SG* standard gamble, *PLE* probability lottery equivalent

have been found to lead to variation in the described levels of impairment, particularly for vitality and role limitations where considerable changes were made; in a large cross-country sample, the SF-6Dv2 role dimension is found to be less likely to suffer from floor effects, relative to the SF-6Dv1, while the SF-6Dv2 attracts a higher proportion of 'best' level responses in the vitality dimension [62]. There are also differences in possible utility values generated by the measures; for the UK, the SF-6Dv2 utility values range from − 0.574 to 1 whereas the SF-6Dv1 values range from 0.301 to 1 [5, 49]. A comparison of the psychometric performance of the two versions in the UK findings show that the SF-6Dv2 successfully discriminates between patient groups and in some conditions, the SF-6Dv2 outperforms the SF-6Dv1 (e.g. diabetes, arthritis) [61] and research is expected to emerge as SF-6Dv2 gains further use.

4.5 Research Developments

4.5.1 Standalone SF-6Dv2

The SF-6D was not designed to be completed as a standalone measure. However, there has been continued interest from users for an option to use the classifier rather than either the SF-36 or SF-12. The development of the SF-6Dv2 classification system prompted the development of a six-item short form, the SF-6Dv2 health utility survey (SF-6Dv2 HUS) from which the SF-6Dv2 can be scored [63]. This form was developed and evaluated using think-aloud and cognitive debriefing interviews of two survey forms: form A used a question and response format (e.g. for bodily pain, the question was 'during the past 4 weeks, how much bodily pain have you had?' with responses: 'none', 'very mild pain', 'mild pain', 'moderate pain', 'severe pain' and 'very severe pain'). Form B used a heading and five or six statements for each dimension describing the health levels (e.g. for bodily pain, the heading was 'pain in the past 4 weeks' and the statements were: 'you had no bodily pain', 'you had very mild bodily pain', 'you had mild bodily pain', 'you had moderate bodily pain', 'you had severe bodily pain' and 'you had very severe bodily pain'). Participants were randomized to answer either form A or form B first followed by the other form. Generally, participants evaluated both forms as easy to answer, but preferred form A. In separate evaluations of each dimension, participants liked elements of the form B physical functioning dimension, so these elements were included in the final form [63].

This is an important update as it allows for SF-6Dv2 values to be calculated when only utilities, and not the full SF-36 or SF-12 scores, are required, therefore reducing the response burden on patients. The SF-6Dv2 HUS has been translated into 20 languages and further languages are expected.

Two studies have explored the consistency of responses to different versions of the SF-6Dv2. Poder et al. [64] tested three versions of the SF-6Dv2 (derived from the SF-36, a version with 10 questions, and another with six questions) and found relatively good consistency in respondents' answers. Ameri et al. [65] found some

evidence of differences based on the version used in a breast cancer population, leading to small differences in values. This is useful evidence given the development of the standalone version, but it should be noted that these studies did not use the licenced version developed by Broderick et al. [63].

4.5.2 Estimating SF-6Dv2 Values from the SF-12

Currently the SF-6Dv2 cannot be generated using SF-12 data since some of the items required to generate responses for the SF-6Dv2 classification are not included in the SF-12. This is recognised as a limitation, and therefore research is ongoing to enable the prediction of SF-6Dv2 scores from SF-12 through the use of statistical approaches to mapping.

5 Recommendations for Use

5.1 Where and When to Use SF-6D?

The SF-36 and SF-12 are widely used with evidence of validity in different populations (see Sect. 5.2). This offers the opportunity for SF-6D to be used in those contexts for both the 4 week and 1 week recall period versions. However, given that the SF-6D uses only a subset of SF-36 or SF-12 items, direct evidence of the psychometric properties of whichever version of the SF-6D is used is useful to support validity of use in different populations. Where possible, drawing on evidence from systematic reviews across several studies on psychometric performance of the SF-6D in populations of interest is recommended rather than relying on single studies [66].

The version of the SF-36 or SF-12 that is used will determine which version of the SF-6D is appropriate. SF-6Dv1 can be generated from the SF-36 or SF-12 version 1 and 2 (Fig. 1). SF-6Dv2 can be generated from SF-36v2 or SF-6Dv2 HUS. Given the improvements made in developing the SF-6Dv2, it is recommended that users of the SF-36v2 generate SF-6Dv2 utilities where possible, e.g. in the UK where value sets are available for both SF-6D versions. Future work will aim to enable different versions to be linked so that current and previous SF-36 or SF-12 data can be used to generate similar versions of the SF-6D.

The use of SF-6Dv1 and SF-6Dv2 as a classification system directly, rather than obtaining this information by administering SF-36 or SF-12 is possible, with the SF-6Dv2 HUS developed and tested formally [63]. The decision of whether to collect data using the classification system or SF-12 or SF-36 data should take into account both response burden and whether the wider level of information generated from instead using SF-36 or SF-12 over the classification system is advantageous.

The availability of country-specific value sets may also determine which version of SF-6D to use. There are more official value sets available for SF-6Dv1 ($n = 9$) than SF-6Dv2 ($n = 5$) (Table 2) but more SF-6Dv2 value sets are expected over time. The SF-6Dv2 valuation protocol uses DCE which can be implemented online which may increase the generation of SF-6Dv2 value sets as it is more cost effective to undertake valuation surveys online rather than in person.

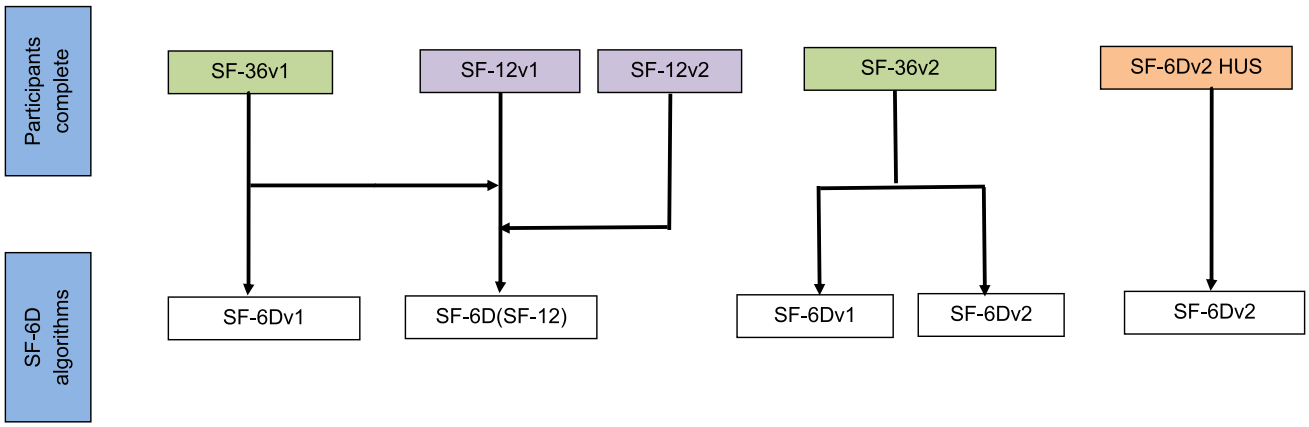

Fig. 1 Selecting the SF-6D version

In the context of CEA, different HTA/reimbursement agencies may recommend different generic preference-based measures. The SF-6D is named as an example of a preference-based measure for use in 11 of 14 guidelines that recommend the use of QALYs without recommending specific measures [29]. In England and Wales, the recommended measure is the EQ-5D but where this is not shown to be psychometrically valid, the alternative would be to use a valid generic preference-based measure such as SF-6D [67].

5.2 Psychometric Validity of SF-6D

Generic measures, such as SF-6D (and the SF-36 or SF-12 that it is derived from) are necessarily restricted to core dimensions of HRQoL that may apply across different health conditions to enable wide use and comparability. Generic measures are designed to assess HRQoL either directly (e.g. pain), or indirectly (e.g. the impact of pain on role functioning). Psychometric assessment of the measures is important as it ensures that the measures are relevant in the populations that they are used in terms of face validity for patients, content that reflects the impact of conditions and/or interventions, and the ability to detect differences in groups with known differences or over time where change has occurred [68]. When assessing condition-specific measures, there may sometimes be gold standard measures or objective assessments against which measures can be assessed but this is not the case for generic measures. There is also no gold standard for subjective health. For preference-based measures, there is an added level of complexity as preferences come from members of the public and there is no gold standard against which to assess the resultant utilities. Therefore we use psychometric assessment to support hypotheses about how utilities would be expected to differ in terms of the ability to distinguish between groups with known differences and show responsiveness which are important for CEA [68].

Evidence on SF-36/SF-12 from systematic reviews may be used to assess where the content of the measures has been found to be valid, which is a useful starting point for identifying where SF-6D could be valid. However, the generalizability of evidence on SF-36/SF-12 is limited given that SF-6D only uses a subset of these items and is scored using preferences. Given the extremely wide use of SF-36/SF-12, it is not possible to review all the available evidence for these measures. We recommend that users who wish to use the SF-6D identify or conduct systematic literature reviews to assess the validity of these measures in their target populations. For example, there is evidence of the validity of the SF-36 and/or SF-12 in different physical health populations e.g. breast cancer [69], prostate cancer [70], traumatic brain injury [71], rheumatoid arthritis [72], systemic lupus erythematosus [73], ulcerative colitis [74], hip and knee disorders [75] and haemophilia [76]. SF-36 has also been found to be valid in mental health populations e.g. in an alcohol dependent population [77] and for depression and anxiety [78]. However, systematic reviews have found mixed evidence in some populations e.g. stroke [79], hearing loss [80], vision loss [81], schizophrenia [82] and personality disorders [83].

As with the SF-36/SF-12, it is not possible to review all the evidence related to SF-6D (mostly SF-6Dv1) and other measures in the context of this paper. Here we highlight evidence from a review of reviews [84] for SF-6Dv1 and emerging papers for SF-6Dv2, the latter based on a brief scoping review undertaken for this manuscript in March 2024.

In their review of reviews on the psychometric performance of generic preference-based measures across different conditions, Finch et al. [84] found 30 systematic reviews some of which included SF-6Dv1 ($n = 12$). SF-6Dv1 showed evidence of being valid for urinary incontinence, visual disorders and depression but was mixed for diabetes and spinal cord injury in terms of known group validity. However, there was evidence of responsiveness for spinal cord injury, systemic lupus erythematosus and depression and anxiety. Overall, however, there were fewer studies assessing the psychometric performance of SF-6Dv1 compared with EQ-5D in this review. In comparative studies with EQ-5D, SF-6Dv1 has fewer ceiling effects than EQ-5D but higher floor effects for more severe conditions [85].

A search was undertaken on PubMed and Google Scholar with the search term 'SF-6Dv2/SF-6D version 2/SF-6Dvs2' covering the period from 2020 to 2024. Studies are reported here if they assessed psychometric validity. The search identified seven studies that assessed the performance of SF-6Dv2. Most of the published SF-6Dv2 evidence relates to studies undertaken in mainland China using the Chinese SF-6Dv2 value set. The SF-6Dv2 has been found to be valid in a Chinese general population sample ($n = 19{,}177$) [86] and in a study involving university staff and students ($n = 291$ and 183, respectively) [87] showing no evidence of ceiling effects for utilities and the ability to discriminate between chronic condition groups or having a disease/symptoms or injury. In Chinese patients ($n = 117$) with Pompe disease (PD), a rare genetic metabolic myopathy, the SF-6Dv2 was found to have no ceiling effects and it was able to discriminate between groups on the basis of severity of disabilities [85]. For cancer, SF-6Dv2 was able to discriminate between known groups in Chinese patients with lymphoma ($n = 200$) and there was evidence of responsiveness ($n = 78$) [88]. Evidence from other countries in general population and patient samples support these findings. In a UK general population and mixed patient sample ($n = 7392$), the SF-6Dv2 performed well in terms of known-group validity and successfully distinguished disease severity and between the disease and healthy groups [61]. SF-6Dv2 showed a statistically

significant treatment benefit in two trials, and non-significant benefits in two other trials in an analysis in randomised trials of an obesity treatment, thus supporting the responsiveness of the SF-6Dv2 [89]. In an Iranian breast cancer population ($n = 416$) there was convergence of some of the SF-6Dv2 dimensions and utilities (UK value set) with a breast cancer-specific measure, and SF-6Dv2 utilities discriminated between progressive stage of diagnosis and severe treatment strategies [90]. Therefore, the emerging evidence on the SF-6Dv2 suggests it can discriminate between groups with known differences and that it can capture change over time.

In addition to psychometric evidence, minimal important differences (MIDs) have been estimated for generic measures including SF-6Dv1. MID estimates can be used to understand the clinical significance of changes in HRQoL scores. They are determined using both distribution (e.g. effect size of approximately 0.5 standard deviations) and anchor-based methods (e.g. global ratings of change) [91]. They are therefore dependent on the clinical population where they are estimated, for example, MIDs for SF-6Dv1 were found to range from 0.01 to 0.1 across 11 patient groups [92]. No MIDs have yet been estimated for SF-6Dv2.

5.3 Licensing of SF-6Dv1, SF-6Dv2, and the SF-6Dv2 HUS

The SF-6D is licensed by QualityMetric (https://www.qualitymetric.com/health-surveys/sf-6dv2-license-request/). The licence is provided for free for non-commercial use while commercial use attracts a fee.

The measures are widely licenced, with 600 commercial licences to date since 2022, of which 351 are from SF-6Dv1 and 249 from SF-6Dv2, and 153 academic SF-6D licences since 2020.

It is important to note that the SF-6D is derived from either the SF-36 (https://www.qualitymetric.com/health-surveys/the-sf-36v2-health-survey/), the SF-12 (https://www.qualitymetric.com/health-surveys/the-sf-12v2-pro-health-survey/), or the SF-6Dv2 HUS (https://www.qualitymetric.com/health-surveys/sf-6dv2-health-utility-survey/). Thus, a licence must be obtained for the relevant measure. The measures are available for paper and pencil administration, electronic administration including single item presentation for handheld devices, and as interview administered versions. These licences are also managed by QualityMetric. When requesting a license for the SF-36 and SF-12, users can request to also get access to the SF-6D algorithm to calculate the SF-6Dv1 and SF-6Dv2 scores.

6 Discussion

6.1 Development and Evolution

The development of the SF-6D was led by a perceived need for a way to generate utilities from the SF-36. This work led to an approach that enabled a classifier to be derived and valued without requiring new data collection in the target population as the SF-6D algorithm could be applied both prospectively and retrospectively. The innovative approach used in the development work has been formalised and replicated for use in condition-specific measures [93]. As the SF-36 has evolved, there has been a need to update the SF-6D to ensure that SF-6D utilities can still be derived from the different SF-36 versions. This included adaptation to be able to be generated from the short version SF-12 and changes to the wording which resulted in SF-36v2 and SF-12v2. The SF-6Dv1 algorithm was therefore adapted to enable SF-6Dv1 to be derived from both version 1 and 2 of the SF-36 and SF-12.

The UK SF-6Dv1 valuation approach has been replicated in other countries but some have also used different valuation methods, such as DCE with duration to value SF-6Dv1 owing to concerns with using SG as a valuation method. This switch reflects wider innovations within the health state valuation field where DCE has gained traction over the last decade. There were also innovations in modelling the health state valuation data using nonparametric Bayesian approaches rather than frequentist approaches. This approach can be used to combine data from different countries which would be particularly valuable for countries with limited resources to conduct large-scale valuation studies.

Criticisms with both the classifier and valuation method used for SF-6Dv1 have led to the development of SF-6Dv2. The SF-6Dv2 was based on reviewing all the relevant items and identifying the best performing ones while also improving the wording to minimise the occurrence of levels that were difficult to distinguish in valuation. In addition, an international DCE valuation protocol has been developed and refined that can be used online for cost-effective health state valuation. As a result, SF-6Dv2 utilities are different from SF-6Dv1 utilities and this has an implication for comparability. Future work should consider linking utilities from the two versions in a similar way to the approach undertaken for the EQ-5D-3L and EQ-5D-5L [94, 95].

6.2 Comparison to Other Generic Measures

As a generic measure, the SF-6D includes dimensions that are similar to those included in measures, such as EQ-5D including those related to physical functioning, mental health and pain [96]. Like other generic measures, SF-6D

focuses primarily on physical health with only two dimensions that relate to mental health (mental health and role limitations). Notably fatigue is one dimension included in SF-6D that is not included directly in the most commonly used generic measure, the EQ-5D. SF-6D does not include impairments, such as vision, hearing and dexterity that are included in measures, such as the Health Utilities Index (HUI) [97] but instead focuses on the impact of these limitations on higher level dimensions. Even where there is overlap in dimensions, there are differences in the number of questions, wording of questions related to similar constructs, response options, recall periods, valuation methods and available value sets which means that measures are not directly comparable [96].

SG was used to value SF-6Dv1 and has also been used for the HUI to transform visual analogue scale (VAS) values to utilities [97]. However, although SG is often regarded as theoretically superior owing to its basis in expected utility theory, it is complex and has been criticised as noted in Sect. 3.5 and other measures have used either VAS or TTO. For example, EQ-5D-3L has mainly been valued using TTO (see e.g. [7, 21, 22]). DCE has grown in usage for health state valuation [98] and the development of an international valuation protocol for SF-6Dv2 using DCE [49] reflects this as does the use of both TTO and DCE in valuing EQ-5D-5L [99]. There are now 12 country specific value sets that are available for SF-6D but these are fewer than those available for the most commonly used generic measure, the EQ-5D.

As noted in Sect. 5.2, although there is a lot of evidence on psychometric validity of the SF-36/12 from which the SF-6D is derived, the evidence for SF-6D is comparatively less compared with EQ-5D. The available psychometric evidence shows that SF-6D measures are valid in some populations but when compared to EQ-5D-3L, SF-6Dv1 has fewer ceiling effects than EQ-5D but higher floor effects for more severe conditions [84]. Higher floor effects may impede the ability of SF-6Dv1 to distinguish levels of severity and lower the responsiveness of the measure for those with the poorest health. There is little evidence comparing the performance of SF-6Dv2 and EQ-5D-5L, and addressing this evidence gap would be informative. Separately, the availability of MID estimates means that SF-6D scores have applicability in research and clinical practice but these need to be estimated in each clinical population of interest.

Compared with other commonly used generic measures, SF-6D has the advantage of being linked to profile measures of HRQoL, the SF-36 and the SF-12. The SF-36 and SF-12 measures are widely used internationally providing a large pool of data that can be used to inform resource allocation via CEA. This includes generation of norm data from general population samples which can be useful when undertaking modelling and there are examples of norm data for both SF-6Dv1 [100–105] and SF-6Dv2 [106, 107].

The direct link between the profile and preference-based measure also provides a rich resource of different types of information from the same measure in a single population which can be useful in providing a fuller understanding of the impact of health conditions and interventions and a rich suite of data describing the experience of patients and to general population preferences. As the SF-36 and SF-12 are commonly used in trials, this does not require the inclusion of a different measure to generate utility values. The link to a profile measure allows the SF-6D classifiers to use a combination of questions for some dimensions, e.g. physical functioning has two or three questions which uses more relevant information compared with other generic measures which rely on a single question. This is lower than those used in the AQoL-8D which relies on all the questions in the valuation survey, but higher than in some generic measures, such as the EQ-5D [96]. SF-6D is often generated through the collection of SF-36 and SF-12 data, this involves higher respondent burden owing to the larger number of items that are administered than EQ-5D where the dimensions are asked directly. Longer measures may lead to respondent fatigue, potentially impacting the accuracy and reliability of the response and they may limit practicality in some clinical and research settings where simplicity and brevity are prioritized. For studies collecting new data, the SF-6Dv2 HUS offers an alternative that overcomes these concerns.

6.3 Conclusions

The evolution of SF-6D from SF-6Dv1 to SF-6Dv2 has reflected the use of state-of the art methods for valuation and development of classifiers to be employed to generate a second version that resolves some concerns with the first version. The generation of international value sets to reflect preferences from different country populations enables own country preferences to be used to inform CEA and potentially resource allocation decisions. Whilst the SF-6Dv2 can be used as a standalone version, thus addressing concerns of patient burden, derivation of the SF-6D from SF-36 and SF-12 data has the advantage that a profile measure can be used in a study to capture data on a larger number of items on HRQoL and is still able to generate utilities. The SF measures continue to be widely used internationally both commercially and for non-commercial studies, and evidence is emerging over time on the performance of SF-6Dv2 and its comparability to SF-6Dv1. We envisage that SF-6D will remain a key measure internationally primarily for use in cost utility analysis, but also in a variety of other applications, such as clinical settings and population health assessment, in the future.

Data availability This article does not analyse data.

Declarations

Funding The authors did not receive support from any organisation for the submitted work.

Conflicts of interest J.E.B, B.M., D.R. and J.B.B. were involved in the development of SF-6D and/or SF-6Dv2; C.M., D.R., J.E.B, B.M., J.B.B. and S.K. have been involved in valuation studies of the SF-6D and/or SF-6Dv2; C.M., D.R. and J.E.B. are based at the University of Sheffield which owns the copyright for the algorithms for the SF-6D and SF-6Dv2; D.R., B.M. and J.E.B. receive developer royalties from commercial licenses of SF-6Dv2; C.M., D.R., B.M. and J.E.B. are members of the EuroQol Group; E.M. has no conflicts of interest to declare.

Author contributions This was an invited paper. C.M., D.R., B.M., S.K., E.M. and J.B.B. conceptualized the study; C.M., D.R., B.M., S.K., E.M., J.B.B. and J.E.B. carried out literature review; C.M., D.R., B.M., S.K., E.M., J.B.B. and J.E.B. carried out writing—original draft preparation; C.M., D.R., B.M., S.K., J.B.B. and J.E.B. carried out writing—review and editing.

References

1. Haraldstad K, et al. A systematic review of quality of life research in medicine and health sciences. Qual Life Res Int J Qual Life Asp Treat Care Rehabil. 2019;28(10):2641–50.
2. Pequeno NPF, et al. Quality of life assessment instruments for adults: a systematic review of population-based studies. Health Qual Life Outcomes. 2020;18(1):208–208.
3. Siette J, et al. Systematic review of 29 self-report instruments for assessing quality of life in older adults receiving aged care services. BMJ Open. 2021;11(11):e050892–e050892.
4. Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 health survey. J Clin Epidemiol. 1998;51(11):1115–28.
5. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care. 2004;42(9):851–9.
6. Brazier JE, et al. Developing a new version of the SF-6D health state classification system from the SF-36v2: SF-6Dv2. Med Care. 2020;58(6):557–65.
7. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002;21(2):271–92.
8. Weinstein MC, Torrance G, McGuire A. QALYs: the basics. Value Health. 2009;12(Suppl 1):S5–9.
9. Brazier J, Ratcliffe J, Saloman J, Tsuchiya A. Measuring and valuing health benefits for economic evaluation. In: Oxford Medicine Online. Oxford, Oxford University Press; 2017.
10. Ware JE, Sherbourne CD. The MOS 36-ltem short-form health survey (SF-36). Med Care. 1992;30(6):473–83.
11. Hays RD, Sherbourne CD, Mazel RM. The rand 36-item health survey 1.0. Health Econ. 1993;2(3):217–27.
12. Ware JE. SF-36 health survey: manual and interpretation guide. Boston: The Health Institute, New England Medical Center; 1993.
13. McHorney CA, Ware JE, Rachel Lu JF, Sherbourne CD. The MOS 36-ltem short-form health survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med Care. 1994;32(1):40–66.
14. Gandek B, Ware JE. Methods for validating and norming translations of health status questionnaires. J Clin Epidemiol. 1998;51(11):953–9.
15. Keller SD, et al. Use of structural equation modeling to test the construct validity of the SF-36 health survey in ten countries. J Clin Epidemiol. 1998;51(11):1179–88.
16. Frendl DM, Ware JE Jr. Patient-reported functional health and well-being outcomes with drug therapy: a systematic review of randomized trials using the SF-36 health survey. Med Care. 2014;52(5):439–45.
17. Ware JE, Gandek B. Overview of the SF-36 health survey and the International Quality Of Life Assessment (IQOLA) Project. J Clin Epidemiol. 1998;51(11):903–12.
18. Ware JE, et al. Evaluating translations of health status questionnaires: methods from the IQOLA Project. Int J Technol Assess Health Care. 1995;11(3):525–51.
19. Furlong W, et al. Guide to design and development of health-state utility instrumentation. Hamilton: Centre for Health Economics and Policy Analysis (CHEPA), McMaster University; 1992.
20. Ware JE. SF-36 health survey update. Spine. 2000;25(24):3130–9.
21. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35(11):1095–108.
22. McCabe C, et al. Using rank data to estimate health state utility models. J Health Econ. 2006;25(3):418–31.
23. Kharroubi SA, Brazier JE, Roberts J, O'Hagan A. Modelling SF-6D health state preference data using a nonparametric Bayesian method. J Health Econ. 2007;26(3):597–612.
24. Kharroubi SA, O'Hagan A, Brazier JE. Estimating utilities from individual health preference data: a nonparametric Bayesian method. J R Stat Soc Ser C Appl Stat. 2005;54(5):879–95.
25. Kharroubi SA, Brazier JE, McGhee S. Modeling SF-6D Hong Kong standard gamble health state preference data using a non-parametric bayesian method. Value Health. 2013;16(6):1032–45.
26. Kharroubi SA, McCabe C. Modeling HUI 2 health state preference data using a nonparametric Bayesian method. Med Decis Mak. 2008;28(6):875–87.
27. Kharroubi SA, Daher CA. Modelling a preference-based index for EQ-5D using a non-parametric Bayesian method. Qual Life Res. 2018;27:2841–50.
28. Rowen D, Azzabi Zouraq I, Chevrou-Severac H, van Hout B. International regulations and recommendations for utility data for health technology assessment. Pharmacoeconomics. 2017;35(Suppl 1):11–9.
29. Kennedy-Martin M, et al. Which multi-attribute utility instruments are recommended for use in cost-utility analysis? A review of national health technology assessment (HTA) guidelines. Eur J Health Econ HEPAC Health Econ Prev Care. 2020;21(8):1245–57.
30. ISPOR. Pharmacoeconomic guidelines around the world. 2024 April 2024. https://www.ispor.org/heor-resources/more-heor-resources/pharmacoeconomic-guidelines. Accessed 14 Aug 2024

31. Lam CLK, Brazier J, McGhee SM. Valuation of the SF-6D health states is feasible, acceptable, reliable, and valid in a Chinese population. Value Health. 2008;11(2):295–303.
32. Brazier JE, et al. Estimating a preference-based index from the Japanese SF-36. J Clin Epidemiol. 2009;62(12):1323–31.
33. Kharroubi SA, Mukuria C, Dawoud D, Rowen D. Estimating the SF-6Dv1 value set for a population-based sample in Lebanon. Value Health Reg Issues. 2024;42:1–10.
34. Ferreira LN, et al. A Portuguese value set for the SF-6D. Value Health. 2010;13(5):624–30.
35. Martínez-Pérez J-E, Abellán-Perpiñán J-M, Sánchez-Martínez F-I, Ruiz-López J-J. A Spanish value set for the SF-6D based on the SF-12 v1. Eur J Health Econ. 2024. https://doi.org/10.1007/s10198-023-01657-9.
36. Abellán Perpiñán JM, Sánchez Martínez FI, Martínez Pérez JE, Méndez I. Lowering the "floor" of the SF-6D scoring algorithm using a lottery equivalent method. Health Econ. 2011;21(11):1271–85.
37. Cruz LN, et al. Estimating the SF-6D value set for a population-based sample of Brazilians. Value Health. 2011;14(5):S108–14.
38. Mulhern B, Norman R, Street DJ, Viney R. One method, many methodological choices: a structured review of discrete-choice experiments for health state valuation. Pharmacoeconomics. 2018;37(1):29–43.
39. Wang L, Poder TG. A systematic review of SF-6D health state valuation studies. J Med Econ. 2023;26(1):584–93.
40. Norman R, et al. Valuing SF-6D health states using a discrete choice experiment. Med Decis Mak. 2013;34(6):773–86.
41. Jonker MF, Donkers B, de Bekker-Grob EW, Stolk EA. Advocating a paradigm shift in health-state valuations: the estimation of time-preference corrected QALY tariffs. Value Health. 2018;21(8):993–1001.
42. Craig BM, Pickard AS, Stolk E, Brazier JE. US valuation of the SF-6D. Med Decis Mak Int J Soc Med Decis Mak. 2013;33(6):793–803.
43. Méndez I, Perpiñán JMA, Martínez FIS, Pérez JEM. Inverse probability weighted estimation of social tariffs: an illustration using the SF-6D value sets. J Health Econ. 2011;30(6):1280–92.
44. Kharroubi SA. A Bayesian nonparametric approach for modeling SF-6D health state utility scores. Value Health Reg Issues. 2022;27:1–11.
45. Kharroubi SA, Brazier JE, McGhee S. A comparison of Hong Kong and United Kingdom SF-6D health states valuations using a nonparametric Bayesian method. Value Health. 2014;17(4):397–405.
46. Kharroubi SA. A comparison of Japan and UK SF-6D health-state valuations using a non-parametric Bayesian method. Appl Health Econ Health Policy. 2015;13(4):409–20.
47. Kharroubi SA. Modeling SF-6D health utilities: is Bayesian approach appropriate? Int J Environ Res Public Health. 2021;18(16):8409.
48. Ferreira PL, Ferreira LN, Pereira LN. How consistent are health utility values? Qual Life Res. 2008;17(7):1031–42.
49. Mulhern BJ, Bansback N, Norman R, Brazier J. Valuing the SF-6Dv2 classification system in the United Kingdom using a discrete-choice experiment with duration. Med Care. 2020;58(6):566–73.
50. Currie CJ, et al. The Routine collation of health outcomes data from hospital treated subjects in the Health Outcomes Data Repository (HODaR): descriptive analysis from the first 20,000 subjects. Value Health. 2005;8(5):581–90.
51. Richardson J, Khan MA, Iezzi A, Maxwell A. Cross-national comparison of twelve quality of life instruments: MIC paper 1 background, questions, instruments. Research Paper. Melbourne, Australia; 2012.
52. Mulhern B, Norman R, Brazier J. Valuing SF-6Dv2 in Australia using an international protocol. Pharmacoeconomics. 2021;39(10):1151–62.
53. Daroudi R, et al. Valuing the SF-6Dv2 in the capital of Iran using a discrete choice experiment with duration. Qual Life Res. 2024;33(7):1853-1863.
54. Wu J, et al. Valuation of SF-6Dv2 health states in China using time trade-off and discrete-choice experiment with a duration dimension. Pharmacoeconomics. 2021;39(5):521–35.
55. Osman AMY, Wu J, He X, Chen G. Eliciting SF-6Dv2 health state utilities using an anchored best-worst scaling technique. Soc Sci Med. 2021;279: 114018.
56. Rendas-Baum R, et al. HTA163 development of SF-6Dv2 health utility weights in the United States. Value Health. 2022;25(12):S328.
57. Sullivan T, et al. Creating an SF-6Dv2 social value set for New Zealand. Soc Sci Med. 2024;354: 117073.
58. Dufresne É, et al. SF-6Dv2 preference value set for health utility in food allergy. Allergy. 2020;76(1):326–38.
59. Touré M, Pavic M, Poder TG. Second version of the short form 6-dimension value set elicited from patients with breast and colorectal cancer. Med Care. 2023;61(8):536–45.
60. Kouakou CRC, He J, Poder TG. Estimating the monetary value of a quality-adjusted life-year in Quebec. Eur J Health Econ. 2024;25(5):787–811.
61. McDool E, Mukuria C, Brazier J. A comparison of the SF-6Dv2 and SF-6D UK utility values in a mixed patient and healthy population. Pharmacoeconomics. 2021;39(8):929–40.
62. Whitehurst DGT, Brazier JE, Viney R, Mulhern BJ. The SF-6Dv2: how does the new classification system impact the distribution of responses compared with the original SF-6D? Pharmacoeconomics. 2020;38(12):1283–8.
63. Broderick L, et al. Development of the SF-6Dv2 health utility survey: comprehensibility and patient preference. J Patient Rep Outcomes. 2022;6(1):47–47.
64. Poder TG, Fauteux V, He J, Brazier JE. Consistency between three different ways of administering the short form 6 dimension version 2. Value Health. 2019;22(7):837–42.
65. Ameri H, Safari H, Poder T. Exploring the consistency of the SF-6Dv2 in a breast cancer population. Expert Rev Pharmacoecon Outcomes Res. 2020;21(5):1017–24.
66. Prinsen CAC, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res Int J Qual Life Asp Treat Care Rehabil. 2018;27(5):1147–57.
67. National Institute for Health and Care Excellence. NICE health technology evaluations: the manual. London: NICE; 2022.
68. Brazier J, Deverill M. A checklist for judging preference-based measures of health related quality of life: learning from psychometrics. Health Econ. 1999;8(1):41–51.
69. Treanor C, Donnelly M. A methodological review of the Short Form Health Survey 36 (SF-36) and its derivatives among breast cancer survivors. Qual Life Res. 2014;24(2):339–62.
70. Hamoen EHJ, et al. Measuring health-related quality of life in men with prostate cancer: a systematic review of the most used questionnaires and their validity. Urol Oncol Semin Original Investig. 2015;33(2):69.e19-69.e28.
71. Polinder S, et al. Health-related quality of life after TBI: a systematic review of study design, instruments, measurement properties, and outcome. Popul Health Metr. 2015;13:4–4.
72. Linde L, et al. Health-related quality of life: validity, reliability, and responsiveness of SF-36, EQ-15D, EQ-5D, RAQoL, and HAQ in patients with rheumatoid arthritis. J Rheumatol. 2008;35(8):1528–37.
73. Strand V, Simon LS, Meara AS, Touma Z. Measurement properties of selected patient-reported outcome measures for use in randomised controlled trials in patients with systemic lupus

erythematosus: a systematic review. Lupus Sci Med. 2020;7(1): e000373.
74. Yarlas A, et al. Psychometric validation of the SF-36® Health Survey in ulcerative colitis: results from a systematic literature review. Qual Life Res. 2017;27(2):273–90.
75. Waal JMVD, et al. The impact of non-traumatic hip and knee disorders on health-related quality of life as measured with the SF-36 or SF-12. A systematic review. Quality Life Res. 2005;14(4):1141–55.
76. Szende A, et al. Health-related quality of life assessment in adult haemophilia patients: a systematic review and evaluation of instruments. Haemophilia. 2003;9(6):678–87.
77. McPherson A, Martin CR. A review of the measurement properties of the 36-item short-form health survey (SF-36) to determine its suitability for use in an alcohol-dependent population. J Psychiatr Ment Health Nurs. 2012;20(2):114–23.
78. Brazier J, et al. A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. Health Technol Assess (Winch, Engl). 2014;18(34):vii–188.
79. Cameron LJ, et al. Self-reported quality of life following stroke: a systematic review of instruments with a focus on their psychometric properties. Qual Life Res. 2021;31(2):329–42.
80. Nordvik Ø, et al. Generic quality of life in persons with hearing loss: a systematic literature review. BMC Ear Nose Throat Disord. 2018;18:1–1.
81. Purola P, Koskinen S, Uusitalo H. Impact of vision on generic health-related quality of life—a systematic review. Acta Ophthalmol. 2023;101(7):717–28.
82. Papaioannou D, Brazier J, Parry G. How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in schizophrenia? A systematic review. Value Health J Int Soc Pharmacoecon Outcomes Res. 2011;14(6):907–20.
83. Papaioannou D, Brazier J, Parry G. How to measure quality of life for cost-effectiveness analyses of personality disorders: a systematic review. J Pers Disord. 2013;27(3):383–401.
84. Finch AP, Brazier JE, Mukuria C. What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. Eur J Health Econ HEPAC Health Econ Prev Care. 2018;19(4):557–70.
85. Xu RH, Luo N, Dong D. Measurement properties of the EQ-5D-3L, EQ-5D-5L, and SF-6Dv2 in patients with late-onset Pompe disease. Eur J Health Econ. 2024. https://doi.org/10.1007/s10198-024-01682-2
86. Xie S, et al. Comparison of the measurement properties of SF-6Dv2 and EQ-5D-5L in a Chinese population health survey. Health Qual Life Outcomes. 2022;20(1):96–96.
87. Zhou HJ, et al. Psychometric performance of EQ-5D-5L and SF-6DV2 in measuring health status of populations in Chinese university staff and students. BMC Public Health. 2023;23(1):2314–2314.
88. Zhang A, et al. Psychometric performance of EQ-5D-5L and SF-6Dv2 in patients with lymphoma in China. Eur J Health Econ. 2024. https://doi.org/10.1007/s10198-024-01672-4
89. Bjorner JB, Larsen S, Lübker C, Holst-Hansen T. The improved health utility of once-weekly subcutaneous semaglutide 2.4 mg compared with placebo in the STEP 1–4 obesity trials. Diabetes Obes Metab. 2023;25(8):2142–50.
90. Nahvijou A, Safari H, Ameri H. Psychometric properties of the SF-6Dv2 in an Iranian breast cancer population. Breast Cancer. 2021;28(4):937–43.
91. Norman GR, Sridhar FG, Guyatt GH, Walter SD. Relation of distribution-and anchor-based approaches in interpretation of changes in health-related quality of life. Med Care. 2001;39(10):1039–47.
92. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res. 2005;14:1523–32.
93. Brazier JE, et al. Developing and testing methods for deriving preference-based measures of health from condition-specific measures (and other patient-based measures of outcome). Health Technol Assess. 2012;16(32)):1-114. https://doi.org/10.3310/hta16320.
94. Hernández Alava M, Pudney S, Wailoo A. Estimating the relationship between EQ-5D-5L and EQ-5D-3L: results from a UK Population Study. Pharmacoeconomics. 2023;41(2):199–207.
95. van Hout BA, Shaw JW. Mapping EQ-5D-3L to EQ-5D-5L. Value Health. 2021;24(9):1285–93.
96. Brazier J, Ara R, Rowen D, Chevrou-Severac H. A review of generic preference-based measures for use in cost-effectiveness models. Pharmacoeconomics. 2017;35(S1):21–31.
97. Furlong WJ, Feeny DH, Torrance GW, Barr RD. The Health Utilities Index (HUI®) system for assessing health-related quality of life in clinical studies. Ann Med. 2001;33(5):375–84.
98. Wang H, Rowen DL, Brazier JE, Jiang L. Discrete choice experiments in health state valuation: a systematic review of progress and new trends. Appl Health Econ Health Policy. 2023;21(3):405–18.
99. Stolk E, et al. Overview, update, and lessons learned from the international EQ-5D-5L valuation work: version 2 of the EQ-5D-5L valuation protocol. Value Health. 2019;22(1):23–30.
100. van den Berg B. SF-6D population norms. Health Econ. 2012;21(12):1508–12.
101. Wong CKH, Mulhern B, Cheng GHL, Lam CLK. SF-6D population norms for the Hong Kong Chinese general population. Qual Life Res. 2018;27(9):2349–59.
102. Shiroiwa T, et al. Japanese population norms for preference-based measures: Eq-5d-3l, Eq-5d-5l, and Sf-6d. Value Health. 2015;18(7):A738.
103. Norman R, Church J, van den Berg B, Goodall S. Australian health-related quality of life population norms derived from the SF-6D. Aust N Z J Public Health. 2013;37(1):17–23.
104. Ferreira PL, Ferreira LN, Pereira LN. SF-6D Portuguese population norms. Eur J Health Econ. 2014;16(3):235–41.
105. Ciconelli RM, et al. Brazilian urban population norms derived from the health-related quality of life SF-6D. Qual Life Res. 2015;24(10):2559–64.
106. Xie S, Wu J, Xie F. Population Norms for SF-6Dv2 and EQ-5D-5L in China. Appl Health Econ Health Policy. 2022;20(4):573–85.
107. Poder TG, Carrier N. Quebec health-related quality of life population norms in adults using the SF-6Dv2. Med Care. 2022;60(7):545–54.