



This is a repository copy of *A comparison of the SF-6Dv2 and SF-6D UK utility values in a mixed patient and healthy population.*

White Rose Research Online URL for this paper:
<https://eprints.whiterose.ac.uk/174667/>

Version: Accepted Version

Article:

McDool, E. orcid.org/0000-0002-3530-7921, Mukuria, C. and Brazier, J. (2021) A comparison of the SF-6Dv2 and SF-6D UK utility values in a mixed patient and healthy population. *PharmacoEconomics*, 39 (8). pp. 929-940. ISSN 1170-7690

<https://doi.org/10.1007/s40273-021-01033-6>

This is a post-peer-review, pre-copyedit version of an article published in *PharmacoEconomics*. The final authenticated version is available online at: <http://dx.doi.org/10.1007/s40273-021-01033-6>.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

A comparison of the SF-6Dv2 and SF-6D UK utility values in a mixed patient and healthy population

SF-6Dv2 and SF-6D UK utility values compared

Emily McDool*, Clara Mukuria* and John Brazier*

*The School of Health and Related Research (SCHARR), The University of Sheffield, Sheffield, UK

April 2021

Abstract

Objective: This paper describes the first evaluation of the construct validity and performance of the newly developed preference-based measure of health, the SF-6D version 2 (SF-6Dv2).

Method: Utilising data from the Multi-Instrument Comparison (MIC) project (n=7,932), we explored the descriptive differences in utility values between the SF-6Dv2 and the SF-6D and evaluated the known group validity of both measures by testing the statistical significance of differences in utility values and calculating the effect sizes across known groups. The convergent validity of the SF-6Dv2 was explored by examining whether the SF-6Dv2 is related to alternative validated measures, including the EQ-5D-5L and AQL-8D.

Results: Differences between the utilities values of the SF-6Dv2 and SF-6D were evident; utilities were generally lower for the SF-6Dv2 with larger standard deviations resulting in larger absolute differences between groups. The SF-6Dv2 performed well in terms of known-group validity and successfully distinguished disease severity and between disease and healthy groups, outperforming the SF-6D in some but not all groups. Convergent validity analyses indicated strong associations between the SF-6Dv2 and the SF-6D, EQ-5D-5L and AQL-8D utilities.

Conclusions: The psychometric performance of the SF-6Dv2 is favourable with respect to known group validity and convergent validity, but does not seem to have improved, compared to the SF-6D. However, the new method of valuation has had a substantial impact on the size of absolute differences in utility values which could impact QALY results. The economic evaluation of health interventions may therefore be influenced by the choice of the SF-6Dv2 over the SF-6D.

Corresponding author: Emily McDool Email: e.m.mcdool@sheffield.ac.uk

Funding: This work is funded by SF-6D licensing.

Conflicts of interest: Emily McDool has no conflicts of interest. Clara Mukuria has no conflicts of interest. John Brazier is a developer of the SF-6D, for which commercial users pay a licence fee to the University of Sheffield.

Availability of data: The Multi Instrument Comparison (MIC) Study is held by Monash University and is free to all users. Data requests can be made via AQL

Code availability: Stata version 16

Authors contributions: John Brazier and Clara Mukuria were involved in the study conception and framing of the paper. Emily McDool performed the analysis and prepared the first draft of the manuscript. All authors were involved in the interpretation of the data, and review and critical revision of the content. All authors read and approved the final manuscript.

Key points:

- The SF-6D version 2 (SF-6Dv2) was developed to address concerns with the SF-6D and to use the SF-6Dv2 in economic evaluation, a UK utility value set has been developed.
- The SF-6Dv2 is able to successfully distinguish between groups with known differences, such as disease and healthy groups and outperforms the SF-6D in some but not all cases.
- The SF-6Dv2 is also strongly correlated with other health related quality of life instruments.

1 Introduction

The quality adjusted life year (QALY) can be used to measure outcomes in the economic evaluation of health interventions. A QALY combines the value of health-related quality-of-life (HRQoL) with the value of length of life into a single index number. Typically, the quality element of a QALY is derived from preference-based measures of health which generate a utility value of HRQoL. One such preference-based instrument is the SF-6D which was developed from the SF-36 [1]. The SF-6D describes health on 6 dimensions, including: physical functioning, role limitations, social functioning, pain, mental health and vitality on between 4 and 6 severity levels, thereby describing 18,000 health states. With the development of country specific value sets, including in the UK, the SF-6D has become one of the most widely adopted generic preference-based measures of health in economic evaluation [2].

However, the SF-6D has been criticised on the basis of the ordering of severity levels of the physical functioning dimension, as the disparity between “a lot” of limitations in moderate activities and “a little” limitation in bathing and dressing lacks clarity. In addition, there are concerns surrounding the framing of dimensions which may cause confusion during valuation; the vitality dimension is positively framed while the other dimensions are negatively framed [2]. The sensitivity of the role dimension has also raised concerns due to claims of a “floor effect” whereby many patients report the lowest severity level [3]. Furthermore, the valuation methods adopted in the SF-6D have faced concerns. Firstly, as the cognitively demanding standard gamble approach was used to elicit values and secondly, as a 2-stage valuation process was adopted this may have caused the impact of risk aversion to be double counted [2]. In response to these concerns, a new version of the SF-6D has been developed (SF-6Dv2) which addresses the problems with the descriptive system [2]. The SF-6Dv2 describes 18,750 possible health states. Moreover, a discrete choice experiments (DCE) with duration was used to value the SF-6Dv2 as opposed to standard gamble which was used for the SF-6D [4]. Although standard gamble is an appropriate valuation approach, concerns have been raised with how well respondents understand the task.

The development of a new instrument raises questions in regard to its psychometric validity and the implications of its use in economic evaluation to inform Health Technology Assessment (HTA) [5]. An instrument’s validity refers to the extent to which it measures the construct it is intended to measure. One type of construct validity is convergent validity which refers to the strength of the association between the instrument and other measures of related concepts. A second form of construct validity is known group validity which refers to the degree to which an instrument is able to differentiate between groups where differences are expected a priori. At present,

there is limited evidence on the performance of the SF-6Dv2 including compared with that of the SF-6D and whether it addresses concerns with the latter measure. Whitehurst et al provides the starting point for comparative work by analysing how the classification system of the SF-6Dv2 impacts the distribution of responses compared with the SF-6D [5]. The study identifies that the two classification systems lead to variation in the described levels of impairment, particularly for vitality and role limitations. These differences along with different valuation approaches employed for the two versions are likely to have an impact on the utility values. It is clear that the SF-6Dv2 has resulted in a wider scoring range (-0.574 to 1) when compared with the SF-6D (0.301 to 1) [4]. Given that utility values are applied in the calculation of QALYs in economic evaluation, it is important to assess the impact of using SF-6Dv2 compared to SF-6D. Therefore, the aim of this study was to compare the psychometric performance of the SF-6Dv2 to the SF-6D in order to provide evidence of the comparability of and differences in the updated measure and the original measure for HTA. The specific objectives of the paper were to assess the relative performance of the SF-6Dv2 and SF-6D and explore the descriptive differences between utility values (from the UK value set) of the two instruments. In addition, the known-groups validity and the convergent validity of the SF-6Dv2 was evaluated. This evidence is likely to be of interest to agencies who recommend the use of the SF-6D [6].

2 Methods

2.1 Data source

Data from the Multi-Instrument Comparison (MIC) project were used. The MIC study aimed to compare several health-related quality of life (HRQoL) and wellbeing instruments across multiple disease areas (asthma, cancer, depression, diabetes hearing problems, arthritis, heart disease, COPD and stroke). Data were collected online in six countries (Australia, Canada, Germany, Norway, UK and United States). Additionally, the MIC included a sample of 'healthy' individuals with no self-reported health problems. We used data from seven disease areas, excluding the COPD and stroke groups due to small samples that were only drawn from Australia. Excluding these groups provides a total sample of 7,932 individuals.

2.2 Measures

2.2.1 SF-6Dv2 and SF-6D

Participants in the MIC study completed the SF-36 version 2 (SF-36v2) which is a measure of HRQoL. Both the SF-6Dv2 and SF-6D are derived from the SF-36v2. The SF-6D classification system is derived from 11 of the 36 items from the SF-36v2 [1]. The SF-6D describes health on 6 dimensions (physical functioning, role limitations, social functioning, pain, mental health and vitality). Each dimension has between four and six severity levels thereby describing 18,000 health states. A subset was valued using standard gamble by a UK sample resulting in utility values ranging from 0.301 to 1.

The SF-6Dv2 is similarly derived from the SF-36v2 but there are multiple sources of variation from the SF-6D [2]. Firstly, different questions from the SF-36v2 were used in SF-6Dv2 compared to SF-6D in the role limitations, pain and vitality dimensions (see Online Resource A). There were also differences in the number of

severity levels for the physical functioning dimension, which has 6 levels in the SF-6D and 5 levels in the SF-6Dv2, and the role limitation dimension, which has 4 levels in the SF-6D and 5 levels in the SF-6Dv2. Furthermore, the level descriptions of the SF-6Dv2 were simplified and made more consistent for clarity [2]. The SF-6Dv2 describes 18,750 possible health states. Online Resource A provides further comparisons of the classification systems of the SF-6Dv2 and SF-6D. DCE with duration was used to obtain utility values for the SF-6Dv2 value set [4]. Utilities of the SF-6Dv2 range from -0.574 to 1.

2.2.2 *Other measures*

Participants in MIC also completed other generic preference-based measures including the EQ-5D-5L [7] and the AQoL-8D [8]. The EQ-5D-5L has five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Respondents are able to choose between five levels ranging from “no problems” to “extreme problems / unable to”. The EQ-5D-5L can describe 3,125 health states which are each associated with a utility score from the cross walk [9] to the UK EQ-5D-3L value set which ranges from -0.594 to 1. The AQoL is a 35-item, 8-dimension multi-attribute utility instrument. The dimensions include independent living, pain, senses, mental health, happiness, coping, relationships and self-worth. The AQoL-8D is an extension of two earlier instruments, the ‘AQoL’ (AQoL-4D) and AQoL-6D. Health states defined by the AQoL-8D are associated with a global utility score which ranges from 0.105 to 1.

Participants with a health condition also completed an appropriate and validated condition-specific measure. The Depression, Anxiety and Stress (DASS-21), for example, is a validated measure for those with depression developed by Lovibond and Lovibond who also provide clinical cut-offs to define severity where DASS-21 scores are multiplied by 2 to acquire the total score and therefore range from 0 to 42. We utilise the depression domain total score cut offs: <14 , ≥ 14 to < 21 and ≥ 21 to define mild, moderate and severe groups, which are accepted clinical cut-offs. [10].

Clinical cut-offs were only available and therefore used for the DASS-21 measure of depression. Clinical cut-offs are not available for the remaining condition specific measures, through by nature, the scores of the measures inherently reflect condition severity. In the absence of clinical cut-offs, severity cut-off points were guided by the distribution of responses, the correlation of scores with alternative condition specific measures and indicators and mean scores amongst these correlated indicators. For example, for asthma, we assessed the correlation of the asthma specific measure with specific questions surrounding breathing difficulties and examined the mean and median scores of the measure used to define severity by the extent and severity of breathing difficulties. Furthermore, the MIC data also provides responses to a question about disease severity for each condition, (rated on a five-point scale from very mild to very severe) in the German population only. Bearing in mind the single population that responds to the question, we utilise these responses to ensure that the severity cut offs are broadly correlated with self-reported severity for each condition.

The Asthma Quality of Life Questionnaire (AQLQ-Sydney) [11] is an asthma-specific measure of HRQoL completed by MIC adult respondents with asthma. The asthma global score is calculated as the mean of all 20 items which range from 0 to 5 and is provided in the MIC. We define mild, moderate and severe groups as those with global scores of <1.5 , ≥ 1.5 to ≤ 2.5 and > 2.5 respectively.

The European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire Cancer 30 (EORTC QLQ-C30) is a questionnaire developed by Aaronson et al to assess the quality of life of cancer patients [12]. The calculation of a global mean score from all items is not recommended [13]. As opposed to using the full measure, severity is therefore defined by the response to a single question in the EORTC QLQ-C30 reflecting overall quality of life, albeit not cancer specific: 'How would you rate your quality of life health during the past week?' Responses range from 1 (very poor) to 7 (excellent). Responses of >5 , ≥ 4 to ≤ 5 and <4 are defined as mild, moderate and severe.

Diabetes-39 is an instrument used to assess the quality of life of individuals with diabetes [14]. The MIC provides the additive global score which is the sum of all 39 items each with 7 levels of severity where higher scores indicate greater severity. The score therefore ranges from 39 to 273. We define whole scale scores of <115 , ≥ 115 to ≤ 160 and >160 , as mild, moderate and severe diabetes respectively.

The Abbreviated Profile of Hearing Aid Benefit (APHAB) [15] is used to assess hearing aid benefit and produces scores for aided and unaided performance for the hearing loss sample. We focus on unaided scores given the larger sample size; unaided responses reflect a patient's difficulty when not using amplification. Severity of hearing loss is defined by the global unaided score which is equal to the mean of all item scores in the ease of communication, background noise and reverberation subscale scores, and ranges from 1 to 99. Scores of <40 , ≥ 40 to ≤ 70 and > 70 define mild, moderate and severe hearing loss respectively.

The Arthritis Impact Measurement Scale 2 – Short Form (AIMS2-SF) is a measure of health specifically for individuals with osteoarthritis [16]. The physical, symptom, affect and social subscale scores are summed to give the additive total score which ranges from 38 to 115. Based on the distribution of the total score and correlations with the pain and discomfort dimension of the E5-5D-5L, we define total scores of >95 , ≥ 82 to ≤ 95 and <82 as mild, moderate and severe arthritis respectively.

The MacNEW heart disease health-related quality of life questionnaire is used to assess how daily activities and functioning are affected by coronary heart disease and its treatment [17]. The global score of the measure was calculated as the mean of all 27 items that each comprise of 7 levels of severity. Scores of >5 , ≥ 4 to ≤ 5 and <4 denote mild, moderate and severe heart disease respectively.

2.3 Analysis

2.3.1 Descriptive statistics of utility values

Statistical analyses were used to compare the measures in terms of the utility values they generated. Means, standard deviations and minimum and maximum values were estimated for each utility measure and presented for the sample overall, and by condition. The distribution of utility values of the SF-6Dv2 and SF-6D was examined and compared using density plots (kernel density) for the full dataset as well as by the different condition groups.

Bland-Altman plots were employed to identify whether there was evidence of differences in values across the entire severity scale. These plots describe agreement between two quantitative measures and are presented as a scatter plot in which the X axis represents mean of SF-6Dv2 and SF-6D utility scores and the Y axis shows the difference between SF-6Dv2 and SF-6D values [18].

2.3.2 *Known-group validity: Discrimination across different groups*

Known-group validity was examined by assessing the ability of the SF-6D and SF-6Dv2 to discriminate between respondents with different characteristics. This included the ability to distinguish whether a respondent had an existing health condition, as opposed to being in the 'healthy' group, and by severity within condition groups. We also examined groups based on age as health is expected to deteriorate with age. Based on the distribution of age, four age categories (18-44; 45-54; 55-64 and 65+) were used. In addition, since data on smoking and education is also provided in the MIC data, we also examine groups based on these observable characteristics as a relationship between these characteristics and health has been found elsewhere [19].

To evaluate the relative performance of the two SF-6D versions, mean differences and effect sizes (Cohen's D) were estimated in order to compare groups. The effect size is estimated by dividing the difference in mean scores between two adjacent sub-groups by the pooled standard deviation of scores. Effect sizes of ≥ 0.2 to < 0.5 , ≥ 0.5 to < 0.8 , and ≥ 0.8 , denote small, medium, and large effect sizes respectively [20]. To confirm differences, the statistical significance of differences between the known groups were tested using a T-test to compare two groups or an overall F-test from an Analysis of variance (ANOVA) when comparing across groups where there were more than two groups. In order to confirm the findings of the known-group validity analyses, we employ further tests suggested by Janssen et al including the ratio of the area under the receiver-operator characteristics curve (AUROC) and F-statistics from ANOVA of the SF-6Dv2 and SF-6D for the groups being compared [21].

2.3.3 *Convergent/divergent validity*

Ideally, we would like to assess the convergent validity of the SF-6Dv2 against a gold standard validated measure of health that captures all potential health outcomes. Such a gold standard external validator does not exist for HRQoL. We therefore examined the relationship between the SF-6Dv2 and the original SF-6D as well as how it was related to EQ-5D-5L and AQoL-8D using correlations. Spearman rank correlations were used to assess the association between the dimensions of the measures since the data is ordinal. Pearson correlation coefficients were to compare (continuous) utilities across measures. Correlations were assessed as: ≥ 0.5 strong, < 0.5 to ≥ 0.3 moderate and < 0.3 weak [20].

Higher correlations were expected between similar dimensions across the measures since they attempt to capture the same aspects of health, for instance, the physical functioning dimension of the SF-6Dv2 should be highly correlated with this same dimension measured by the SF-6D. In a similar manner, high correlations are expected between: SF-6Dv2 physical functioning and EQ-5D mobility; SF-6Dv2 pain and EQ-5D pain / discomfort; SF-6Dv2 mental health and EQ-5D anxiety/depression; SF-6Dv2 mental health and AQoL mental health; SF-6Dv2

pain and AqoL pain. Furthermore, the existing literature has highlighted common dimensions and identified high correlations between these dimensions and the SF-6D dimensions [22][23]

3 Results

3.1 Descriptive statistics of utility values

In the MIC data, there were 436 unique utility values for the SF-6D which is 2.4% of all 18,000 states, and 11% of all 4,000 unique possible values that the SF-6D may take [4]. Comparatively, there were 961 unique values reported for the SF-6Dv2 which is 5.1% of all 18,750 states described, and 6.4% of all 15,000 possible values. There were 160 unique values that were less than 0 in the SF-6Dv2 (i.e. 17% of unique values defined in the MIC data).

<insert table 1>

With the exception of the ‘healthy’ and hearing problem sub-samples, the mean SF-6Dv2 utility values were lower than the SF-6D utility values (Table 1). The largest difference in the SF-6D and SF-6Dv2 scores by disease group was observed in the depression group, while the smallest difference was in the hearing problems group where the difference between the SF-6D and SF-6Dv2 was statistically insignificant in contrast to all other groups and in the full sample. The full utility value range was observed for all the groups for the SF-6Dv2 (-0.574 to 1) and for the SF-6D (0.301 to 1) with the exception of arthritis (Table 1).

The SF-6Dv2 and SF-6D utility scores were negatively skewed with a mean of 0.67 and 0.71 and a median of 0.77 and 0.7 respectively (Figure 1). The distribution of scores amongst the disease group sub-samples were similar to the full sample for some, but not all disease groups. Depression and diabetes disease groups in particular had a ‘flatter’ distribution of SF-6Dv2 utility scores in contrast to the SF-6D distribution which was negatively skewed (Figure 1). The Bland-Altman plot showed evidence of agreement between the utility values of the measures particularly for higher utility score (i.e. individuals reporting ‘better’ or full health) (Figure 2). However, there was evidence of disagreement between values with low average scores i.e. those with ‘low’ health and with an average utility score of approximately less than 0.4.

3.2 Known-group validity: Discrimination across different groups

The mean values of SF-6Dv2 and SF-6D utility values by each known group are presented in Table 2 (and presented graphically in Online Resource B). The results from the analyses indicated that both the SF-6Dv2 and SF-6D had the ability to distinguish whether a respondent has an existing health condition, as opposed to being in the ‘healthy’ group (Table 2). The difference in means within groups was larger in the SF-6Dv2 than the SF-6D and all differences were statistically significant. The mean difference was greatest in the depression group (relative to the ‘healthy’ group) and smallest in the hearing problems group. Effect sizes overall were generally large, with the exception of hearing problems where the effect size was 0.403 for the SF-6Dv2 and 0.453 for the SF-6D, and for asthma, where the effect size is equal to 0.75 for the SF-6Dv2 (Table 2). Owing to larger

standard deviations across groups, SF-6Dv2 effect sizes were generally smaller than the SF-6D, with the exception of diabetes and arthritis.

<insert Table 2>

Both the SF-6Dv2 and SF-6D distinguished well between levels of severity within disease groups (Table 2). The significant differences in means were consistently larger in the SF-6Dv2 than the SF-6D across all disease groups. In the majority of diseases, the difference in mean SF-6Dv2 scores were greatest in the moderate vs severe group as opposed to the mild vs moderate (Table 2). The exception to this was the hearing problem group. This was not always the case for the SF-6D utility values and for some groups, the two versions did not agree on the relative order of the differences in means between the severity levels. For example, within the cancer group, the difference in the mean value of the SF-6Dv2 for the moderate vs severe was 0.304 and greater than mild vs moderate, equal to 0.163, but in the SF-6D the largest difference in means was observed in mild vs moderate groups (Table 2). The standard deviation of the SF-6Dv2 increases with severity across all disease groups; this pattern was not observed in the SF-6D for all diseases. With the exception of hearing problems, effect sizes were medium to large for both versions. In all but the hearing group, larger effect sizes were observed for the SF-6D than the SF-6Dv2 when comparing the mild vs moderate severity groups. The effect sizes of the SF-6Dv2 are greater than the SF-6D when comparing moderate to severe (with exception to hearing and asthma groups). The results also demonstrate that both the SF-6D and SF-6Dv2 were able to distinguish between other characteristics including age, education and smoking status with SF-6Dv2 showing larger differences than SF-6D although effect sizes were small (Table 2). These findings are confirmed by the analysis of the AUROC and F-statistic ratios (available in Online Resource B).

3.3 Convergent validity

Table 3 provides the results from the convergent validity analyses. The SF-6Dv2 and SF-6D utility values were strongly correlated ($\rho=0.84$) and equivalent dimensions had very strong correlations (0.74 to 0.93) with the exception of vitality which had a correlation of 0.51 (Table 3). The SF-6Dv2 utility values were also strongly correlated with the EQ-5D utility score ($\rho=0.811$). Strong correlations are observed in the domains that we expect to be highly correlated, including between the SF-6Dv2 pain and EQ-5D pain and discomfort dimensions (0.764), the SF-6Dv2 mental health and EQ-5D anxiety and depression domains (0.628) and the SF-6Dv2 physical functioning and EQ-5D mobility (0.514). Further strong correlations are observed between the SF-6Dv2 physical functioning and EQ-5D usual activities and pain / discomfort domains. Weak correlations were identified between all SF-6Dv2 dimensions and EQ-5D self-care while the SF-6Dv2 mental health did not strongly correlate with any EQ-5D dimension, other than anxiety/ depression (Table 3). SF-6Dv2 utilities were similarly strongly correlated with AQoL utilities (0.762) and strong correlations were observed across the majority of domains (56%). As expected, strong correlations were observed for equivalent dimensions on mental health and pain, -0.697 and -0.820 respectively. SF-6Dv2 social functioning, mental health and vitality domains were strongly correlated with the majority of the AQoL domains. The AQoL sense dimension was weakly correlated with all SF-6Dv2 domains, while the SF-6Dv2 physical functioning domain was the dimension with the fewest strong correlations with AQoL dimensions (Table 3).

<insert Table 3>

4 Discussion and conclusion

This paper utilised the MIC data to compare the psychometric performance of the SF-6Dv2 in terms of known group validity and convergent validity in a mixed sample of healthy individuals and individuals with specific conditions (asthma, cancer, depression, diabetes, hearing problems, arthritis and heart disease). Overall, we identified large differences in the absolute utility values and observed an increased range and variance of the SF-6Dv2 relative to the SF-6D. SF-6Dv2 utility values were generally lower than those from the SF-6D and across disease groups, but were higher for 'healthy' samples. Utility values in the health conditions ranged from 0.757 to 0.446 for SF-6Dv2 and 0.75 to 0.601 for SF-6D with hearing having the highest utility values while depression had the lowest values for both measures. The wider range of values in the SF-6Dv2 reflects the wider range of values in the value set based on DCE with duration which ranges between -0.574 and 1 compared to the SG value set for the SF-6D which ranges between 0.301 and 1. This wider range generates large differences between those with and without a condition for the SF-6Dv2 which is likely to have an impact on QALY estimates. Although there were larger differences between utility values in the two versions for the conditions groups, ranking across them was mainly consistent (hearing, asthma, heart disease, cancer, diabetes, arthritis and depression), though in the SF-6D ranking diabetes precedes cancer, the difference is small and therefore most likely insignificant. The standard deviations were larger for the SF-6Dv2 which led to marginally smaller effect sizes for most conditions, compared to the SF-6D. Larger standard deviations will have an impact on uncertainty which will be evident in the sensitivity analysis of economic evaluations.

Similarly, although both measures were able to discriminate across severity groups, SF-6Dv2 had larger differences compared to SF-6D. For example, the difference between utility values for moderate and severe condition-specific levels of severity ranged between 0.289 and 0.362 for arthritis, cancer, depression and heart disease, whereas none of the differences were more than 0.11 for SF-6D. In some cases, there was a larger difference between utility values for the mild and moderate levels of severity for the SF-6D e.g. for cancer, diabetes, arthritis and heart disease although these differences were not large. This did not occur for the SF-6Dv2 apart from in hearing where, although both versions were able to discriminate between severity levels, the differences were not as large as for the other conditions. This reflects the generic nature of the SF-36v2 which does not include hearing as a dimension. These differences may reflect both the changes in choice of items for the classifier and the changes to the valuation approach. The items selected for SF-6Dv2 aimed to be consistent especially with regards to role limitations (which included a more severe level of problems), vitality (which was previously a positively framed item) and physical functioning (where a severity level was removed). Whereas changes to the role limitations and vitality dimensions may have resulted in lower values, the change made to the physical functioning dimension is less likely to explain the observed differences.

The convergent validity analysis showed strong associations between the SF-6Dv2 and the SF-6D utilities and across equivalent dimensions. The exception was vitality which is not unexpected since a negatively framed item was used in the SF-6Dv2 compared to a positively framed one in the SF-6D. In addition, the SF-6Dv2 was strongly associated with the EQ-5D utility score and the AQoL-8D utility score. Strong correlations were

identified between dimensions where expected, for instance between the SF-6Dv2 pain domain and the pain domains of both the EQ-5D and AQL and the SF-6Dv2 mental health and EQ-5D anxiety / depression and AQL mental health domains.

Newer versions of measures that aim to address previous limitations in the descriptive systems and valuation methods will have an impact on the utility values derived from preference-based measures e.g. the changes in EQ-5D from the 3L to the 5L [24]. The differences in the SF-6D and SF-6Dv2 stem from, firstly, improvements made to the classification system of the SF-6D, which have resulted in variation in the described levels of impairment, although the evidence suggests that these changes have made little difference to the psychometric performance. And secondly, the valuation approach employed which was adapted to address concerns with the SF-6D. Given the little difference in effect sizes and F-statistics, the likely greatest contributing factor to the differences in utility values between the SF-6Dv2 and the SF-6D is the change in valuation method to DCE with duration which has been shown to generate lower utility values. The DCE with duration results make the SF-6Dv2 more consistent with the EQ-5D and other measures of preference-based measures of health in terms of the range and values. This would seem to suggest that the variant of standard gamble with the chained procedure created an artificial floor effect which is consistent with the finding of Tsuchiya et al. [25] who re-valued the SF-6D states using TTO and identified that values were closer to the EQ-5D.

The study benefits from assessment of the SF-6D and SF-6Dv2 in a large international sample across multiple conditions and a number of other commonly used generic preference-based measures. The availability of condition specific measures to judge the performance was also useful as it provides information on how well the two versions discriminate across severity groups. However, most of the measures do not have clinically accepted cut-offs. The approach to define groups based on distribution and an external measure of severity available for one of the countries aimed to provide a more objective choice for the cut-offs but these cut-offs were sample dependent. Furthermore, although known group and convergent validity are assessed, the data were cross-sectional therefore responsiveness and test re-test ability of the SF-6Dv2 could not be undertaken. Future work should examine the validity of the measure across different time points, settings and populations. The final impact of the changes from SF-6D to SF-6Dv2 would also need further testing in the context of trials as the impact may vary depending on the condition and the impact of the intervention [26]. This study provides evidence of comparison across the two versions of the SF-6D which is important for decision-makers looking to use either measure.

This study shows that the development of the new preference-based measure of health, the SF-6Dv2, and the production of its UK value set provided multiple sources of potential deviation from the original SF-6D utility values which result in larger absolute differences between patient and healthy samples and between different patient severity groups. Due to these differences, the selection of the SF-6Dv2 over the SF-6D is likely to influence the QALY results and this has potential implications for the economic evaluation of health interventions. Future work should assess the impact in the context of economic evaluations.

Tables and figures

Table 1 SF-6Dv2 and SF-6D Utility values by 'healthy' & disease groups

	Mean	SD	Median	Min	Max	N	T-value (P value)
Healthy							
SF-6Dv2	0.831	0.156	0.864	-0.193	1		-12.85
SF-6D	0.800	0.108	0.810	0.398	1	1760	(0.000)
Difference	0.031						
Asthma							
SF-6Dv2	0.677	0.279	0.770	-0.574	1		4.7
SF-6D	0.707	0.130	0.696	0.319	1	856	(0.000)
Difference	-0.03						
Cancer							
SF-6Dv2	0.632	0.306	0.734	-0.574	1		7.2
SF-6D	0.686	0.133	0.673	0.301	1	772	(0.000)
Difference	-0.054						
Depression							
SF-6Dv2	0.446	0.320	0.497	-0.574	1		20.05
SF-6D	0.601	0.107	0.601	0.301	1	917	(0.000)
Difference	-0.155						
Diabetes							
SF-6Dv2	0.630	0.316	0.739	-0.574	1		9.45
SF-6D	0.696	0.141	0.691	0.301	1	924	(0.000)
Difference	-0.066						
Hearing problems							
SF-6Dv2	0.757	0.228	0.820	-0.574	1		-1.45
SF-6D	0.750	0.119	0.753	0.334	1	831	(0.152)
Difference	0.007						
Arthritis							
SF-6Dv2	0.570	0.313	0.697	-0.480	1		13.7
SF-6D	0.670	0.126	0.647	0.319	0.958	929	(0.000)
Difference	-0.100						
Heart							
SF-6Dv2	0.655	0.300	0.752	-0.574	1		6.8
SF-6D	0.700	0.133	0.696	0.301	1	943	(0.000)
Difference	-0.045						
Full sample							
SF-6Dv2	0.668	0.296	0.765	-0.574	1		19.9
SF-6D	0.711	0.137	0.700	0.301	1	7932	(0.000)
Difference	-0.043						

Table 2 Known group validity

	Groups	N	SF-6Dv2					SF-6D				
			Mean	SD	Diff. in means	Test stat. (P-value)	Effect Size	Mean	SD	Diff. in means	Test stat. (P-value)	Effect Size
Health condition	Healthy	1,760	0.831	0.156				0.800	0.108			
	Asthma	856	0.677	0.279	0.153	18.01** (0.00)	0.750	0.707	0.130	0.093	19.214*** (0.00)	0.801
	Cancer	772	0.632	0.306	0.199	21.558*** (0.000)	0.931	0.686	0.133	0.114	22.688*** (0.00)	0.979
	Depression	917	0.446	0.320	0.385	41.781*** (0.000)	1.702	0.601	0.107	0.199	45.284*** (0.000)	1.844
	Diabetes	924	0.630	0.316	0.200	21.970*** (0.000)	0.893	0.696	0.141	0.104	21.319*** (0.000)	0.866
	Hearing problems	831	0.757	0.228	0.073	9.569*** (0.000)	0.403	0.750	0.119	0.051	10.766*** (0.000)	0.453
	Arthritis	929	0.570	0.313	0.260	28.762*** (0.000)	1.166	0.670	0.126	0.131	28.058*** (0.000)	1.138
	Heart disease	943	0.655	0.300	0.175	19.963*** (0.000)	0.806	0.700	0.133	0.100	21.146*** (0.000)	0.853
Asthma severity	Mild	507	0.767	0.197				0.751	0.114			
	Moderate	212	0.619	0.305	0.144	89.939*** (0.000)	0.501	0.674	0.128	0.076	102.643*** (0.000)	0.650
	Severe	137	0.449	0.340	0.170		0.593	0.598	0.109	0.076		0.647
Cancer severity	Mild	232	0.824	0.174				0.797	0.107			
	Moderate	342	0.661	0.228	0.163	188.182*** (0.000)	0.626	0.679	0.102	0.119	266.963*** (0.000)	1.159
	Severe	198	0.358	0.348	0.304		1.166	0.569	0.097	0.110		1.073
Depression severity	Mild	260	0.663	0.233				0.682	0.105			
	Moderate	200	0.516	0.248	0.147	194.234***	0.554	0.610	0.076	0.072	167.429***	0.791

			SF-6Dv2					SF-6D				
	Groups	N	Mean	SD	Diff. in means	Test stat. (P-value)	Effect Size	Mean	SD	Diff. in means	Test stat. (P-value)	Effect Size
	Severe	457	0.227	0.306	0.289	(0.000)	0.850	0.552	0.091	0.090	(0.000)	0.637
Diabetes severity	Mild	460	0.785	0.192				0.775	0.112			
	Moderate	232	0.587	0.285	0.198	194.533*** (0.000)	0.694	0.658	0.118	0.117	254.887*** (0.000)	1.033
	Severe	232	0.368	0.354	0.219		0.767	0.577	0.111	0.081		0.716
Hearing problems severity	Mild	287	0.798	0.185				0.763	0.110			
	Moderate	225	0.734	0.236	0.064	12.387*** (0.000)	0.274	0.737	0.120	0.026	8.962*** (0.000)	0.215
	Severe	104	0.680	0.271	0.053		0.229	0.708	0.133	0.029		0.237
Arthritis severity	Mild	354	0.785	0.121				0.769	0.093			
	Moderate	326	0.577	0.230	0.208	380.781*** (0.000)	0.854	0.653	0.085	0.116	432.164*** (0.000)	1.283
	Severe	249	0.256	0.334	0.321		1.317	0.551	0.094	0.101		1.118
Heart disease severity	Mild	598	0.799	0.152				0.768	0.100			
	Moderate	216	0.542	0.252	0.257	512.042*** (0.000)	1.204	0.621	0.079	0.147	486.703*** (0.000)	1.691
	Severe	129	0.180	0.230	0.362		1.441	0.517	0.080	0.104		1.191
Age groups	Below 45	2,589	0.665	0.288				0.707	0.134			
	45-54	1,672	0.622	0.333	-0.043	45.869*** (0.000)	-0.147	0.693	0.145	-0.013	43.199*** (0.000)	-0.097
	55-64	1,976	0.652	0.307	-0.029		-0.100	0.706	0.141	-0.012		-0.089

	Groups	N	SF-6Dv2					SF-6D				
			Mean	SD	Diff. in means	Test stat. (P-value)	Effect Size	Mean	SD	Diff. in means	Test stat. (P-value)	Effect Size
	65+	1,695	0.736	0.241	-0.084		-0.287	0.743	0.123	-0.038		-0.277
Education	Below degree	5,689	0.653	0.303	-0.052	-6.997*** (0.000)	-0.174	0.705	0.138	-0.022	-6.409*** (0.000)	-0.160
	Degree	2,243	0.705	0.276				0.727	0.134			
Smoker	Non- smoker	6,009	0.690	0.281	-0.091	11.772*** (0.000)	-0.308	0.721	0.136	-0.038	10.687*** (0.000)	-0.280
	Smoker	1,923	0.599	0.331				0.683	0.138			

Notes: Test statistic provided is a t-value for the comparison of two groups and the F-statistic from ANOVA when comparing across groups (by age and within disease by severity)

Effect sizes of ≥ 0.2 to < 0.5 , ≥ 0.5 to < 0.8 , and ≥ 0.8 , denote small, medium, and large effect sizes respectively

N is representative of the sample size for both the SF-6Dv2 and SF-6D

Table 3 Convergent validity

	SF-6Dv2	Physical functioning	Role limitation	Social functioning	Pain	Mental health	Vitality
SF-6D Index	0.843***	-0.599***	-0.776***	-0.757***	-0.629***	-0.658***	-0.613***
Physical Functioning	-0.611***	0.863***	0.504***	0.419***	0.549***	0.245***	0.348***
Role limitation	-0.656***	0.434***	0.743***	0.542***	0.434***	0.451***	0.433***
Social functioning	-0.679***	0.393***	0.581***	0.870***	0.429***	0.496***	0.475***
Pain	-0.733***	0.569***	0.538***	0.499***	0.818***	0.312***	0.428***
Mental health	-0.644***	0.219***	0.436***	0.486***	0.261***	0.927***	0.542***
Vitality	-0.597***	0.418***	0.513***	0.470***	0.417***	0.432***	0.512***
EQ-5D Index	0.811***	-0.606***	-0.599***	-0.542***	-0.728***	-0.447***	-0.504***
Mobility	-0.452***	0.514***	0.388***	0.318***	0.468***	0.167***	0.262***
Self-care	-0.238***	0.239***	0.195***	0.196***	0.225***	0.117***	0.138***
Usual activities	-0.539***	0.509***	0.479***	0.412***	0.478***	0.269***	0.356***
Pain / discomfort	-0.631***	0.538***	0.435***	0.375***	0.764***	0.231***	0.358***
Anxiety / depression	-0.544***	0.240***	0.435***	0.457***	0.302***	0.628***	0.454***
AQoL Index	0.762***	-0.473***	-0.627***	-0.622***	-0.556***	-0.648***	-0.617***
Independent living	0.648***	-0.643***	-0.560***	-0.503***	-0.571***	-0.325***	-0.420***
Happiness	0.583***	-0.292***	-0.485***	-0.491***	-0.336***	-0.595***	-0.515***
Mental Health	0.642***	-0.275***	-0.499***	-0.530***	-0.376***	-0.697***	-0.569***
Coping	0.667***	-0.416***	-0.573***	-0.526***	-0.428***	-0.559***	-0.587***
Relationships	0.636***	-0.368***	-0.542***	-0.572***	-0.385***	-0.572***	-0.509***
Self-worth	0.589***	-0.279***	-0.494***	-0.512***	-0.329***	-0.606***	-0.509***
Pain	0.702***	-0.585***	-0.488***	-0.457***	-0.820***	-0.287***	-0.409***
Senses	0.309***	-0.284***	-0.251***	-0.246***	-0.270***	-0.219***	-0.223***

Notes: Pearson's correlations provided for continuous scores; Spearman's correlations provided for ordinal data domains

All correlations are significant at the 1% significant level $p < 0.01$ (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$)

Correlations are assessed as: ≥ 0.5 Strong, < 0.5 to ≥ 0.3 Moderate and < 0.3 Weak Cohen 1992.

Figure 1 SF-6Dv2 and SF-6D scores of the full sample and by disease groups

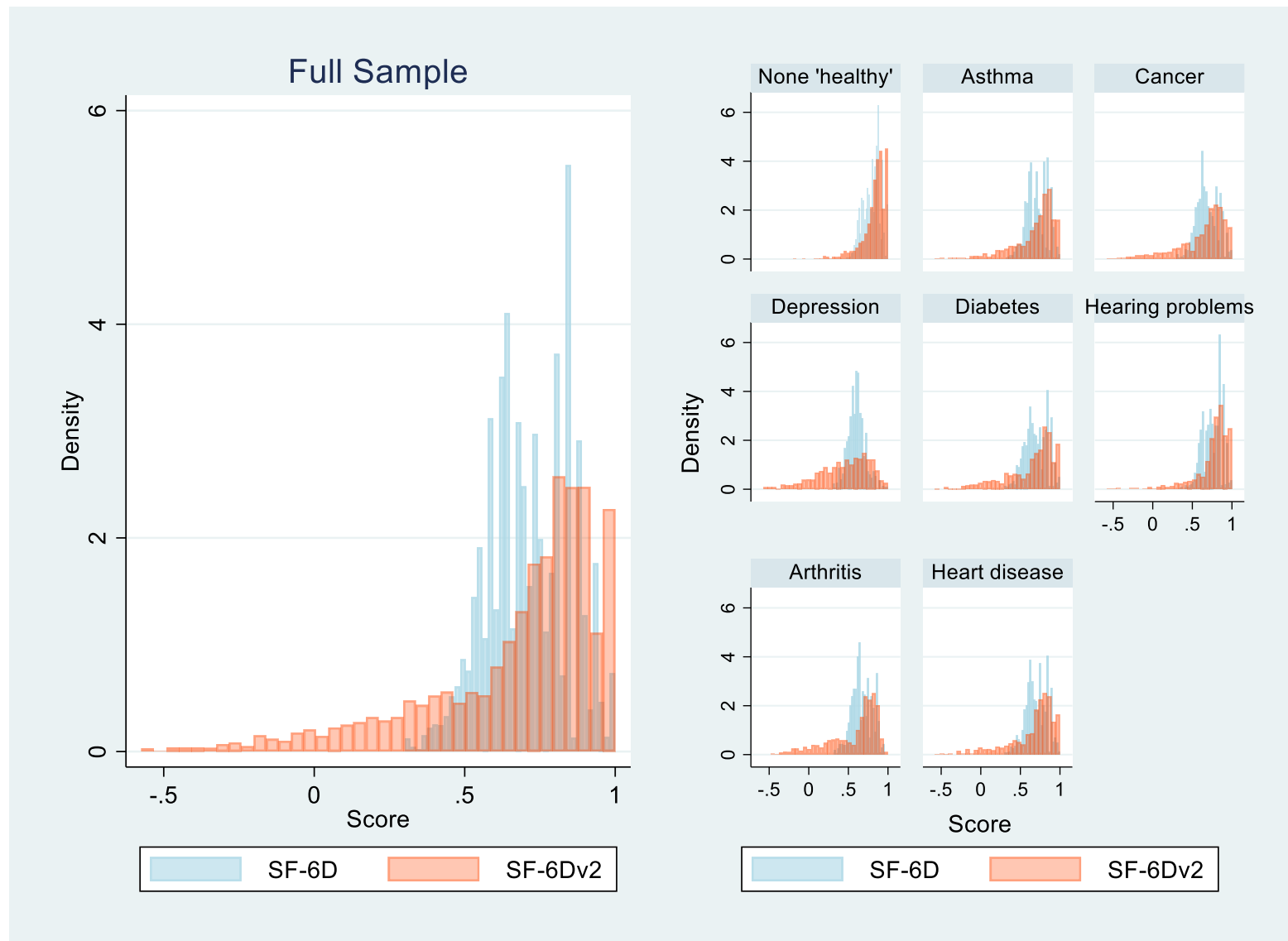
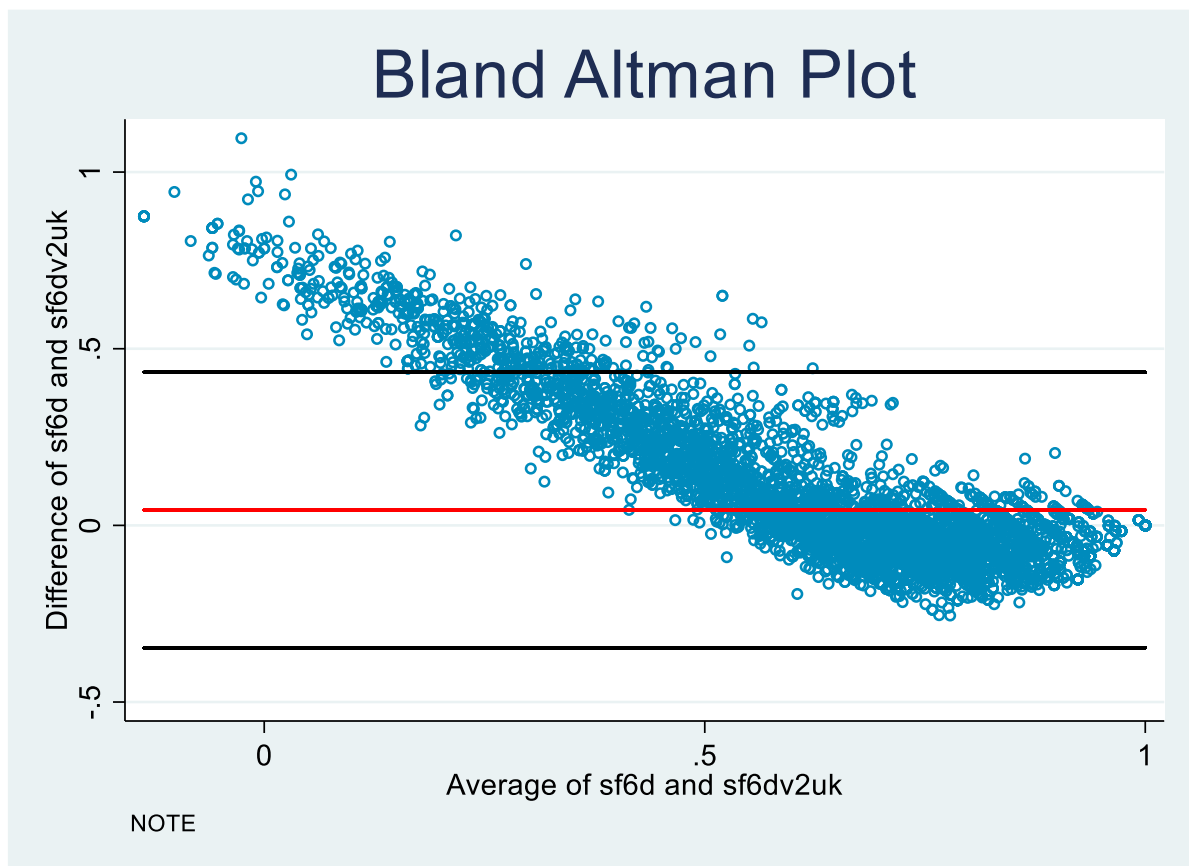


Figure 2 SF-6Dv2 and SF-6D Bland Altman Plot



Notes: The Bland Altman plot shows agreement for higher values of SF-6Dv2 and SF-6D but poor agreement for people with low index values of approximately 0.4 or less in both instruments. The red line indicates the mean difference and the black lines represent the difference (+/-) 1.96 multiplied by the standard deviation

References

- [1] Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics*. 2002; 2:271-292
- [2] Brazier J, Mulhern B, Bjorner J, Gandek B, Rowen D, Alonso J, Vilagut G, Ware J. Developing a New Version of the SF-6D Health State Classification System From the SF-36v2: SF-6Dv2 *Medical Care*. 2020;58(6): 557-565
- [3] Ferreira P, Ferreira L, Pereira, L. How consistent are health utility values? *Quality of Life Research*. 2008;17:1031–1042.
- [4] Mulhern B, Bansback N, Norman R, Brazier J. Valuing the SF-6Dv2 Classification System in the United Kingdom Using a Discrete-choice Experiment With Duration. *Medical care*. 2020;58(6):566-573
- [5] Whitehurst D, Brazier J, Viney R, Mulhern M. The SF-6Dv2: How Does the New Classification System Impact the Distribution of Responses Compared with the Original SF-6D? *PharmacoEconomics*. 2020 doi: 10.1007/s40273-020-00957-9.
- [6] Rowen D, Azzabi Zouraq I, Chevrou-Severac H, Van Hoult B. International Regulations and Recommendations for Utility Data for Health Technology Assessment. *PharmacoEconomics*. 2017; 35, 11-19
- [7] Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, Bonnel G, Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L) *Quality of Life Research*. 2011;20:1727-1736
- [8] Richardson J, Sinha K, Lezzi A, Khan M. Modelling utility weights for the Assessment of Quality of Life (AQoL)-8D. *Quality of Life Research*. 2014;23(8):2395–2404.
- [9] van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, Lloyd A, Scalone L, Kind P, Pickard AS. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health*. 2012 Jul-Aug;15(5):708-15.
- [10] Lovibond S, Lovibond P. *Manual for the Depression Anxiety & Stress Scales*. (2 Ed.) Sydney: Psychology Foundation; 1995
- [11] Marks GB, Dunn SM, Woolcock AJ. A scale for the measurement of quality of life in adults with asthma. *Journal of Clinical Epidemiology*. 1992;45: 461–472.
- [12] Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Filiberti A, Flechtner H, Fleishman SB, De Haes JC, Kaasa S. The European Organization for Research and Treatment of Cancer QLQ-C30: a

quality-of-life instrument for use in international clinical trials in oncology. *JNCI: Journal of the National Cancer Institute*. 1993;85(5):365-376.

[13] The European Organisation for Research and Treatment of Cancer. *EORTC QLQ-C30 Scoring Manual*. 1995. <https://www.eortc.org/app/uploads/sites/2/2018/02/SCmanual.pdf>

[14] Boyer J, Earp J. The development of an instrument for assessing the quality of life of people with diabetes. *Diabetes-39. Medical care*. 1997;35(5):440-453

[15] Cox R, Alexander G. The Abbreviated Profile of Hearing Aid Benefit (APHAB). *Ear and Hearing*. 1995;16:176-186

[16] Meenan R, Gertman P, Mason J. Measuring health status in Arthritis: the Arthritis Impact Measurement Scales. *Arthritis Rheumatism*. 1980;23:146-152.

[17] Lim LY, Valenti LA., Knapp JC, Dobson AJ, Plotnikoff R, Higginbotham N, Heller RF. A self-administered quality-of-life questionnaire after acute myocardial infarction. *Journal of clinical epidemiology*. 1993;46(11):1249-1256.

[18] Bland J, Altman D. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*. 1999;8(2):135-160.

[19] Janssen M, Pickard A, Golicki D, Gudex C, Niewada M, Scalone L, Swinburn P, Busschbach J. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Quality of Life Research*. 2013;22(7):1717–1727.

[20] Cohen J. A power primer. *Psychological bulletin*. 1992;112(1):155

[21] Janssen M, Bonsel G, Luo N. Is EQ-5D-5L better than EQ-5D-3L? A head-to-head comparison of descriptive systems and value sets from seven countries. *PharmacoEconomics*. 2018;36:675-697

[22] Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Economics*. 2004;13(9):873-84

[23] Richardson J, Khan MA, Iezzi A, Maxwell A. Comparing and Explaining Differences in the Magnitude, Content, and Sensitivity of Utilities Predicted by the EQ-5D, SF-6D, HUI 3, 15D, QWB, and AQL-8D Multiattribute Utility Instruments. *Medical Decision Making*. 2015;35(3):276-291.

[24] Mulhern B, Feng Y, Shah K, Janssen M, Herdman M, van Hoult B, Devlin N. Comparing the UK EQ-5D-3L and English EQ-5D-5L Value Sets. *PharmacoEconomics*. 2018;36:699–713

[25] Tsuchiya A, Brazier J, Roberts J. Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets. *Journal of Health Economics* 2006;25:334-346

^[26] Hernandez Alava M, Wailoo A, Grimm S, et al. EQ-5D-5L versus EQ-5D3L: the impact on cost effectiveness in the United Kingdom. *Value in Health* 2018;21:49-56.