# Not just another EQ-5D-5L value set for the UK: using the 'OPUF' approach to study health preferences on the societal-, group-, and individual person-level

Paul Schneider[1,*], Nancy Devlin[2], John Brazier[1]

[1]ScHARR, the University of Sheffield, Sheffield, UK
[2]School of Population and Global Health, University of Melbourne, Melbourne, Australia

## ABSTRACT

### Background

We recently reported on the development of a new method for valuing health states, called 'Online elicitation of Personal Utility Functions' (OPUF). In contrast to established methods, such as time trade-off or discrete choice experiments, the OPUF approach does not require hundreds or thousands of respondents, but allows estimating utility functions for small groups and even on the individual level. The objective of this study was to generate and compare EQ5-5D-5L value sets on the societal-, group-, subgroup-, and individual person-level.

### Methods

The OPUF tool is a new type of online survey – a demo is available at: https://eq5d5l.me**.** It broadly consists of three valuation steps: dimension weighting, level rating, and anchoring. Responses were combined on the individual level to construct personal utility functions, using an additive linear model. Every respondent also completed three conventional discrete choice experiments. We assessed the heterogeneity of preferences between observed and latent groups using PERMANOVA and k-means cluster analysis.

### Results

A representative sample (N = 1,000) of the UK population was recruited through the prolific online platform. On average, it took participants about nine minutes to complete the survey. Data of 874 respondents were included in the analysis. For each respondent, we constructed a personal EQ-5D-5L value set. The derived utility functions predicted respondents' choices in discrete choice experiments with an accuracy of 78%. On the societal level, the predicted values for the EQ-5D-5L health states ranged from -0.376 to 1. Health state preference varied greatly between individuals. This was largely due to differences in the anchoring (i.e. the range of the utility scale respondents used), while there was near consensus on the relative importance of the five EQ-5D dimensions between groups. Demographic characteristics explained only a small proportion of the variability.

### Conclusion

Using the OPUF approach, we were not only able to estimate a new EQ-5D-5L value set for the UK, but also to examine the underlying individual preferences in an unprecedented level of detail.

---

* Contact: p.schneider@sheffield.ac.uk

## 1 INTRODUCTION

Preference-based measures of health, such as the EQ-5D-5L, are an essential component of health economic evaluations. They map health states to a common currency, that is usually referred to as 'utility'. Utility values are needed to compute quality-adjusted life years (QALYs) and to assess and compare the health effects of different treatment options (Whitehead & Ali, 2010).

Preference-based measures of health have two components. Firstly, a descriptive system which defines a number of mutually exclusive health states. Secondly, a value set, which assigns each health state a utility value. These utility values are preference-based. They require the preferences of a target population, in most cases the general population, but occasionally also patients, as input (Brazier et al., 2017b).

Health state preferences can be elicited with various different methods. TTO, DCE, and SG, are probably the ones most commonly used in the context of the QALY framework (Brazier et al., 2017a). These methods, however, have a severe limitation: they only allow the elicitation of (large) group preferences. Since only little information is obtained from each individual, data from hundreds, if not thousands of individuals are usually required to estimate a statistical preference model. Work by Oppe & van Hout (2017) suggests, for example, that the minimum sample size required to derive a linear additive preference model with 20 coefficients for the EQ-5D-5L is about 1,000 participants. This means, health state preferences can only be captured on a (large) group level, and it is generally not feasible to draw inferences about the preferences of any given individual. As a consequence, little is known about the heterogeneity of preferences between individuals.

We recently developed a new approach, for eliciting personal utility functions online (OPUF) (Schneider et al., 2021). It allows estimating health preference models on the individual person-level. The approach is based on previous work by Devlin et al. (2019), but has thus far only been applied in small pilot studies.

In this paper, we report on the results of a large survey of the UK population, in which we used the OPUF approach to elicit preferences for EQ-5D-5L health states. We exploit the approach's ability to construct value sets on the social, group, and individual level, to study the heterogeneity of preferences in an unprecedented level of detail. More specifically, we investigate two research questions:

1. To what extent do health preferences differ between members of the UK general public?
2. How much of these differences can be explained by observed group characteristics or latent preference groups?

## 2 METHODS

### 2.1 Sample

We recruited 1,000 participants through the prolific online platform (Palan & Schitter, 2018) in August 2021. The sample was selected to be broadly representative of the UK general population in terms of age, sex, and ethnicity. All participants completed the EQ-5D-5L OPUF survey.

### 2.2 The EQ-5D-5L instrument

The EQ-5D-5L instrument is a generic preference-based measure of health. It consists of two components: a descriptive system, which defines a number of mutually exclusive health states and, secondly, a set of (social) values, that reflect their respective desirability.

The descriptive system defines health states along five dimensions: mobility (MO), self-care (SC), usual activities (UA), pain or discomfort (PD), and anxiety or depression (AD). Each dimension has five levels: no, slight, moderate, severe, and extreme problems. The instrument can describe a total of 3,125 health states. These states are usually referred to by a 5-digit code, representing the severity levels: '11111' denotes full health, for example; '21111' denotes slight mobility problems but no problems on any other dimension; and '55555' denotes the (objectively) worst health state (Herdman et al., 2011, Devlin et al., 2018).

The social value set maps each health state to a health-related quality of life or, so called, utility value. Utility values range from 1, assigned to perfect health ('11111') to 0, assigned to being dead. Health states that are considered worse than being dead have a negative utility value.

EQ-5D-5L health state preferences are most commonly represented by a linear additive model. It includes 20 coefficients, – four on each dimension – representing the disutility associated with the move from no problems to slight, moderate, severe, and extreme problems (Devlin et al., 2018).

### 2.3 The online elicitation of personal utility functions (OPUF) approach

The OPUF approach is an adaptation of the PUF method (Devlin et al., 2019) for use as a stand alone online survey. In contrast to traditional preference elicitation techniques (TTO, DCE, SG, etc), which are alternative-based (decompositional), the OPUF approach is attribute-based (compositional). The theoretical foundation for both, compositional and decompositional methods, lie in multi-attribute value theory. The difference between the two is the *direction* in which preferences are (de)constructed (Belton & Stewart. 2002, Thokala et al., 2016).

Decompositional methods start with valuing health states. In a second step, the responses are decomposed into their components, using statistical methods. This means, the 20 EQ-5D-5L preference model parameter coefficients are inferred from respondents' holistic evaluation of health states.

In a compositional approach, the partial values for the different components of health states are elicited directly. The components are 1) dimension weights, they determine the relative importance of each dimension; 2) level ratings, they determine the relative position of the five severity levels (no, slight, moderate, severe, extreme) within each dimension; and 3) anchoring, which maps the dimension weights and level ratings on to the QALY scale. These components are then combined to construct values for entire health states. This makes it possible to construct preference functions not only on the group level, but also on the individual person level.

### 2.4 The EQ-5D-5L OPUF survey

The EQ-5D-5L OPUF survey consists of nine steps, of which four are essential for the construction of PUFs. In the following, the steps will be briefly described. A more detailed description of the OPUF survey and its development is provided in Schneider et al., (2021). Much effort went into the design of an intuitive and easy-to-use interface. We thus recommend readers to consult the online demo version of the OPUF survey while reading through this section. It is available at: https://eq5d5l.me.

1) Warm-up (own EQ-5D-5L health state, EQ-VAS)

The first  The survey began with a question  asking the participants to report their own EQ-5D-5L health state and to rate their overall health status, using the EQ-VAS.

2) Level rating

Level ratings were elicited by asking participants to position 'slight', 'moderate', and 'severe health problems' on a visual analogue scale between 0% and 100%. The instructions stated that "a person with 100% health has no health problems", and"a person with 0% health has extreme health problems". It was then asked "[h]ow much health does a person with slight, moderate, and severe health problems have left?".

Ideally, level ratings should be obtained for each dimensions separately. However, the level descriptions of the EQ-5D-5L are very similar across dimensions. The second best level is referred to as 'slight' on all five dimensions for example ('I have slight problems walking about', 'I have slight pain or discomfort', etc). We thus decided to simplify the survey by eliciting the level ratings for health problems in general, i.e. without reference to any particular dimension, and then applied the level ratings to all five dimensions.

3) Dimension ranking

Participants were asked to rank the worst levels of the five EQ-5D dimensions (i.e. 'I am unable to walk about', 'I am unable to wash and dress myself', etc) from worst to less worse. Ties were not permitted. The selected rank order was used to tailor the presentation of the following task (4) to the individual participant.

4) Dimension swing weighting

The task showed five sliders, one for each EQ-5D-5L dimension, describing an improvement from the worst (extreme problems) to the best level (no problems) on the respective dimension. The sliders were presented in the same order as the participant had ranked them before. The first slider (the most important dimension) was set to 100. Participants were asked to use this as a yardstick to evaluate the importance of the the four other dimensions. The instructions for this task were personalised. If, for example, pain/discomfort was ranked first in the previous exercise, the instructions stated: "If an improvement from 'I have extreme pain or discomfort' to 'I have no pain or discomfort' is worth 100 'health points', how many points would you give to improvements in other areas?".

5) Validation DCE

The survey also included three DCEs. The choice sets were personalised, to cover a broad range in terms of severity (mild, moderate, severe health states) and utility differences between scenarios (easy, moderate, difficult). The choice sets always involved trade-offs, i.e. dominant or dominated states were excluded. The responses were not used to construct PUFs. The task was only included to assess the consistency between PUFs and participants' DCE choices.

6) Anchoring I: position-of-dead

Two different methods were used to anchor PUFs on the QALY scale: all participants were asked to consider a pairwise comparison between the worst health state '55555' (scenario A) and being dead (scenario B). If they preferred '55555' over 'being dead', they immediately moved on to task 7. If they preferred 'being dead' over '55555', a binary search algorithm was initiated, during which the health state shown in scenario A changed, adaptively, depending on the participant's choices, to find the health state that they considered to be equivalent to 'being dead' (Sullivan et al., 2020).

To enable the search algorithm, all 3,125 EQ-5D-5L health states were ranked from the best to the worse, based on the participant's responses to the level rating and dimension weighting. After the first comparison ('55555' vs 'being dead'), the algorithm selects the median state (which may be different for each participant). It then jumps up or down, narrowing down on the rank fo the health state that is equal to being dead. After six

iterations, the search ended. At this point, the rank of the equal-to-dead state is being identified with a maximum error of +/- 49 ranks (corresponding to 1.6% of the total number of EQ-5D-5L health states).

7) Anchoring II: dead-VAS

If participants prefer the worst health state, '55555', over 'being dead', the utility of '55555' could take any value between 1 and 0. We therefore asked those participants to locate the position of '55555' on a visual analogue scale between 'No health problems' (=100) and 'being dead' (=0). The selected value was then used as the anchor point for the PUF.

8) Demographic questionnaire

The OPUF survey included questions about personal characteristics, which were assumed or shown to be associated with EQ-5D-5L health preferences. These included: age, sex, having children, importance of religion or spirituality, the frequency of engaging in religious or spiritual activities, level of education, income, and experience with severe health problems – see table 1 for more details (Golicki et al., 2019, MVH. 1995; Feng et al., 2018, Peeters & Stiggelbout 2010).

9) Results page

As a thank-you to the participants, the last page of the survey showed a comparison between some of their own responses and aggregate results from English general population (obtained from Devlin et al. (2018)).

### 2.5 Constructing Personal Utility Functions (PUFs)

PUFs were constructed for all participants. In this section, we provide an overview of the preference construction procedure and illustrate the steps with an example.

Overview

1. The level ratings for no, slight, moderate, severe, and extreme health problems were rescaled between 0 (no problems) and 1 (extreme problems).
2. The five dimension weights were normalised to sum 1.
3. The outer product of the dimension weights and the level ratings was taken to generate a set of 20 (un-anchored) model coefficients (+5 zero coefficients).
4. Depending on wether the participants considered state '55555' better or worse than dead, we either used the response from the 'dead-VAS' or from the 'position-of-dead' task to anchor the model coefficients and map them on to the QALY scale.
5. Finally, the model coefficients were used to generate utility values for all 3,125 EQ-5D-5L health states – this vector of utility values represents the PUF

Example

To illustrate the procedure, suppose a participant gave the following level ratings l with $l_{no} = 100,\ l_{slight} = 90,\ l_{moderate} = 50,\ l_{severe} = 30,$ and $l_{extreme} = 0$; and the following dimension weights *w with* $w_{MO} = 100,\ w_{SC} = 60,\ w_{UA} = 45,\ w_{PD} = 80,$ and $w_{AD} = 70$. After rescaling the level ratings and the dimension weights, we derive the two vectors:

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

*Taking the outer product provides a matrix* $\widetilde{M}$*, containing 20 (1-0 scaled) coefficients (+ zero coefficients for 'no problems' on each dimension).*

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

*Suppose the respondent considered state '51255' (approximately) equivalent to being dead in the 'Position-of-Dead' task. To rescale and anchor* $\widetilde{M}$ *on the QALY scale, we first compute the scaled disutility for the state equal to being dead with* $\widetilde{u}(51255) =$ *0.29+0+0.02+0.23+0.2= 0.74. Subsequently, we set the utility of that state to zero and rescale the entire matrix accordingly, by simply dividing it by the value:*

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

*Note that the constructed preference model assigns state '51255' a value of 0 (= 1 - (0.39+0+ 0.02+0.31+0.27) ); '11111' is still equal to 1 (= 1 - (0+0+0+0+0)), and the worst health state ('55555') now has a value of –0.35 (= 1-(0.39+ 0.23+0.15+0.31+0.27) ). The model can be used to assign utility values to all EQ-5D-5L health states. The resulting vector of 3,125 utility values is taken to be a representation of the participant's PUF.*

### 2.6 Preference Heterogeneity

Investigating the heterogeneity of preferences between individuals, requires a measure of dis/similar to quantify how far apart two PUFs are. As stated above, a PUF was represented by a vector of 3,125 utility values (one for each EQ-5D-5L health state). It would obviously not be useful to compare the utility values of individual health states, nor would it provide much insight to compute means or medians in this case. Instead, we assessed the dissimilarity between PUFs using the euclidean distance (ED) measure.

Analogous to a line between two points on a two dimensional plane, the ED between two PUFs denotes the shortest path length in a 3,125 dimensional space. It is computed as the square root of the sum of the squared differences between the PUFs of individuals *i* and *j*:

$$d_{EUD}(i,j) = \sqrt{\sum \left(u_i(s_1) - u_j(s_1)\right)^2 + \ldots + \left(u_i(s_{3125}) - u_j(s_{3125})\right)^2}$$

with $s = \{11111, 21111, \ldots, 55555\}$

The ED has a lower bound of 0, which indicates that two PUFs are identical. Theoretically, it does not have an upper bound, but due to the design of the EQ-5D-5L OPUF survey (negative values were capped at -31), the maximum ED between two PUFs was 1,789.

### 2.7 Statistical analysis

After we constructed PUFs for all participants, we computed all pairwise ED. We then used two different approaches to partition the distance matrix. Firstly, we performed permutational multivariate analysis of variance (PERMANOVA) to investigate the heterogeneity of preferences between observed groups. Secondly, we used k-means cluster analysis to identify latent preference profiles.

PERMANOVA

PERMANOVA is a geometric partitioning of variation across a multivariate data cloud, defined in the space of any given dissimilarity measure, in response to one or more groups (Anderson. 2014; Anderson & Walsh. 2013). Originally developed to test for differences in dispersion in ecological data (e.g. Souza et al., 2013), in this study, we used it to investigate the variability in EQ-5D-5L health state preferences.

*Analogous to ANOVA,* PERMANOVA decomposes the total distances between observations ($SS_T$) into within-groups ($SS_W$) and between groups sum-of-squares ($SS_B$), with

$$SS_T = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d(i,j)^2 ;$$

and

$$SS_W = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d(i,j)^2 \epsilon_{ij}^{\ell} / n_{\ell}$$

where N is the total sample size (=874), $d(i,j)^2$ is the squared distance between the PUFs of participants *i* and *j*,$\epsilon_{ij}$ *is an indicator which is 1 if participants i and j belong to the same group, and 0 if they do not, and* $n_{\ell}$ *is the* size for group $\ell$. *Then,*$SS_B$ *can then be calculated as* $SS_B = SS_T - SS_W$, which allows calculating the pseudo *F statistic:*

$$F = \frac{\left(\frac{SS_B}{p-1}\right)}{\left(\frac{SS_W}{N-p}\right)}$$

where p is the number of groups.

Semiparametric inference is achieved by permutations. The data is resampled (without replacement) and each time the F statistic is recorded. The original F statistic is then compared to the F statistics of the permutations to derive a p-value. This allows robust statistical analysis in situations where more response variables than participants are observed or when the data is severely non-normal or zero−inflated.

The null hypothesis that is investigated is that the centroids and the dispersion (however defined by the distant measure) are equivalent for all groups. The null hypothesis can be rejected either because the centroids or the spread of the distances is different.

PERMANOVA was performed on the ED matrix. We first tested each of the group characteristics shown in table 1 individually, and then combined them all in one model. P-values were based on 10,000 permutations and a value below 0.05 was considered statistically significant.

Cluster analysis

In addition to the decomposition of the variability of health state preferences with respect to observed group characteristics, we also performed a k-means cluster analysis. The aim was to identify latents groups of participants with distinct health preference profiles. We ran k-means cluster analyses on the ED matrix with 2 to 10 clusters. To determine the optimal number of clusters, we evaluated the change in the explained variance and tried to identify the knee of the curve. This approach, referred to as elbow method, is a commonly used heuristic to identify the point at which including an additional cluster only provides small improvements in model fit.

## 3 RESULTS

### 3.1 Sample

We recruited 1,000 participants through the prolific online platform. Data from 126 participants, who skipped one or more valuation steps, had to be excluded, because no meaningful PUF could be constructed. Characteristics of the 874 participants included in the study are shown in table 1.

Although we sought to recruit a representative sample of the UK population, it was apparent that the included sample was younger (e.g. only 3% were aged 70+ versus 15% in the UK population), and more highly educated (e.g. 56% had a degree versus 40% in the population).

### 3.2 EQ-5D-5L OPUF survey results

On average, it took participants about nine minutes to complete the survey. The median was eight; the shortest duration was three; and the longest was 32 minutes.

**TABLE 1** Sample characteristics

| | n (%%) |
|---|---|
| **Sex** | |
| Female | 456 (52%) |
| Male | 413 (47%) |
| Other/prefer not to say | 5 ( 1%) |
| **Age** | |
| 18-29 | 189 (22%) |
| 30-39 | 188 (22%) |
| 40-49 | 162 (19%) |
| 50-59 | 147 (17%) |
| 60-69 | 164 (19%) |
| 70+ | 23 ( 3%) |
| Prefer not to say | 1 ( 0%) |
| **Children** | |
| No | 410 (47%) |
| Yes | 458 (52%) |
| Prefer not to say | 6 ( 1%) |
| **Education** | |
| without qualifications | 10 ( 1%) |
| GCSE/Standard grade | 93 (11%) |
| A-Level/Higher grade | 161 (18%) |
| Certificate/Diploma/NVQ | 118 (14%) |
| Degree | 305 (35%) |
| Post-graduate | 181 (21%) |
| Prefer not to say | 6 ( 1%) |
| **Income** | |
| £0 – £20,000 | 207 (24%) |
| £20,001 – £30,000 | 161 (18%) |
| £30,001 – £50,000 | 216 (25%) |
| £50,001 – £70,000 | 132 (15%) |
| £70,001+ | 99 (11%) |
| Prefer not to say | 59 ( 7%) |
| **Religious/spiritual practice** | |
| Never/practically never | 545 (62%) |
| A few times a year | 132 (15%) |
| A few times a month | 47 ( 5%) |
| Once a week | 32 ( 4%) |
| A few times a week | 48 ( 5%) |
| Every day | 60 ( 7%) |
| Prefer not to say | 10 ( 1%) |
| **Importance of religion/spirituality** | |
| Not important | 476 (54%) |
| Slightly important | 201 (23%) |
| Moderately important | 100 (11%) |
| Very important | 88 (10%) |
| Prefer not to say | 9 ( 1%) |
| **Experience with health problems*** | |
| Health care professional | 76 ( 9%) |
| Carer | 86 (10%) |
| Family member | 429 (49%) |
| Past own experience | 199 (23%) |
| Present own experience | 49 ( 6%) |
| No experience | 285 (33%) |
| Prefer not to say | 11 ( 1%) |

*non-exclusive categories

Warm-up (own EQ-5D-5L health state, EQ-VAS)

Most participants had no or only mild health problems: 216 (25%) were in full health and 404 (46%) reported slight problems on one or more dimensions. Overall, problems were most frequently reported for the AD (n=470; 53%) and the PD dimension (n=458, 52%).

Most participants also reported high EQ-VAS scores: the mean (SD) and median (IQR) was 77.56 (15.59) and 80 (70-90), with a range of 12 to 100.

Level ratings

The mean (SD) ratings assigned to the 'slight', 'moderate', and 'severe health problems' were 80.23 (11.23); 55.61 (11.55); and 23.47 (13.18), respectively. Participants tended to assign round values: for example, 182 (21%) participants assigned a rating of 80 to the 'slight' level, and another 112 (13%) assigned it a value of 90.

Dimension weights

The EQ-5D-5L dimension that was, on average, considered to be most important was pain/ discomfort with a mean (SD) weight of 90.05 (16.61), followed by mobility and self-care, which nearly identical weights of 82.88 (20.71) and 82.87 (20.47), and then anxiety/ depression with a mean weight of 75.80 and the highest standard deviation of 24.15. The least important dimension was usual activities, with a mean (SD) weight of 73.71 (22.15).

Anchoring (position-of-dead and dead-VAS)

For 342 (39%) participants, who indicated that they would prefer state '55555' over 'being dead', we took the anchor point from the dead-VAS task. For the remaining 532 (61%) participants, who considered '55555' worse than dead, we anchored the PUF using their responses to the position-of-dead task. Figure 1 below shows the resulting bi-modal distribution of utility values for state '55555'. The mean (SD) utility of state '55555' was -0.37 (0.83), and the lowest and highest values were -9.42 and 1.

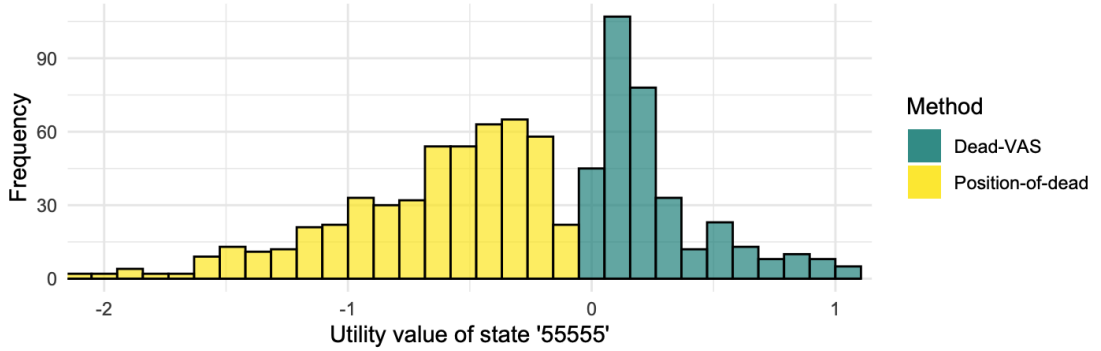


**FIGURE 1** distribution of utility values for state '55555', based on the responses from either the dead-VAS or the position-of-dead task. Values below -2 are not shown (n=24).

### 3.3 Personal utility functions and an alternative EQ-5D-5L social value set for the UK

Descriptive statistics for the constructed personal EQ-5D-5L preference models are provided in table 2. It may be interesting to note that for all model coefficients, the lowest observed value was zero. This was the case because 5 (0.6%) participants assigned a utility of 1 to state '55555', with the implication that all health states were set equivalent to full health, i.e. there was no disutility associated with problems on any dimension. The reported mean model coefficients may also be interpreted a social utility function; they could be used to generate an alternative EQ-5D-5L social value set for the UK.

### 3.4 Validation DCE

Overall, PUFs predicted participants' DCE responses with an accuracy of 78.5%. The responses of 453 (52%) participants were fully consistent, while 299 (34%) made one, 101 (12%) made two, and 21 (2%) made three 'mistakes'. Moreover, we found that the consistency varied depending on the difficulty of the DCE choice set. When the utility difference between the two presented health states was large (>0.3 measured on a personalised 1-0 utility scale) 82% (325 of 395) choices were consistent. Yet, even when the utility difference was small (<0.1) and the choice was difficult, a participant's PUF still predicted their choices with an accuracy of 68% (143 of 209 of choices).

**TABLE 2** Descriptive statistics of personal EQ-5D-5L model coefficients (n=874)

| | Mean (95% CI) | Min. | Q1 | Median | Q3 | Max. |
|---|---|---|---|---|---|---|
| **Mobility** | | | | | | |
| Level 2 | 0.055 (0.053; 0.059) | 0.000 | 0.024 | 0.044 | 0.071 | 1.021 |
| Level 3 | 0.123 (0.121; 0.130) | 0.000 | 0.071 | 0.109 | 0.156 | 1.271 |
| Level 4 | 0.213 (0.210; 0.223) | 0.000 | 0.128 | 0.193 | 0.267 | 1.794 |
| Level 5 | 0.283 (0.278; 0.297) | 0.000 | 0.168 | 0.252 | 0.346 | 2.270 |
| **Self-Care** | | | | | | |
| Level 2 | 0.055 (0.054; 0.058) | 0.000 | 0.026 | 0.045 | 0.071 | 0.613 |
| Level 3 | 0.124 (0.122; 0.130) | 0.000 | 0.072 | 0.110 | 0.158 | 0.813 |
| Level 4 | 0.213 (0.210; 0.222) | 0.000 | 0.133 | 0.192 | 0.267 | 1.250 |
| Level 5 | 0.282 (0.278; 0.294) | 0.000 | 0.174 | 0.256 | 0.350 | 2.083 |
| **Usual activities** | | | | | | |
| Level 2 | 0.048 (0.047; 0.051) | 0.000 | 0.022 | 0.038 | 0.062 | 0.623 |
| Level 3 | 0.108 (0.106; 0.113) | 0.000 | 0.062 | 0.096 | 0.138 | 0.813 |
| Level 4 | 0.186 (0.184; 0.194) | 0.000 | 0.110 | 0.168 | 0.236 | 1.250 |
| Level 5 | 0.248 (0.245; 0.260) | 0.000 | 0.150 | 0.220 | 0.317 | 2.083 |
| **Pain/Discomfort** | | | | | | |
| Level 2 | 0.060 (0.059; 0.063) | 0.000 | 0.029 | 0.050 | 0.080 | 0.534 |
| Level 3 | 0.136 (0.134; 0.141) | 0.000 | 0.082 | 0.122 | 0.171 | 0.813 |
| Level 4 | 0.234 (0.231; 0.243) | 0.000 | 0.147 | 0.214 | 0.293 | 1.273 |
| Level 5 | 0.309 (0.305; 0.322) | 0.000 | 0.190 | 0.275 | 0.387 | 2.083 |
| **Anxiety/Depression** | | | | | | |
| Level 2 | 0.049 (0.048; 0.052) | 0.000 | 0.020 | 0.040 | 0.065 | 0.652 |
| Level 3 | 0.111 (0.110; 0.117) | 0.000 | 0.061 | 0.099 | 0.145 | 0.813 |
| Level 4 | 0.192 (0.189; 0.200) | 0.000 | 0.114 | 0.173 | 0.246 | 1.250 |
| Level 5 | 0.254 (0.250; 0.266) | 0.000 | 0.153 | 0.227 | 0.322 | 2.083 |

*95% CI = 95% confidence intervals, based on 10,000 bootstrap iterations; Q1 = first quartile; Q3 = third quartile

### 3.5 Preference heterogeneity

The average utility values for the EQ-5D-5L health states ranged from 1 to -0.37. The variability of utility values increased with severity: the mean and standard deviation (SD) of states '22222', '33333', '44444', and '55555' were 0.73 (0.22), 0.40 (0.38), -0.04 (0.60), and -0.37 (0.83), respectively. (N.B.: by definition, '11111' has a value of 1).

Figure 2 illustrates the substantial variation in participants' health state preferences. It shows are the average utility values across all participants, i.e. the social value set, for a subset of 100 health states, ranked from the best to the worst (according to the social preference). The thin lines represent the 874 individual PUFs. The colour of the line indicates the ED from the average social value set.

We computed the ED between the PUFs of all participants, which yielded a 874 x 874 distance matrix with 381,501 unique pairwise comparisons. The mean (SD) and median (IQR) ED was 23.36 (23.02) and 17.95 (9.72; 29.37). The highest and lowest observed ED were 259.93 and 0.

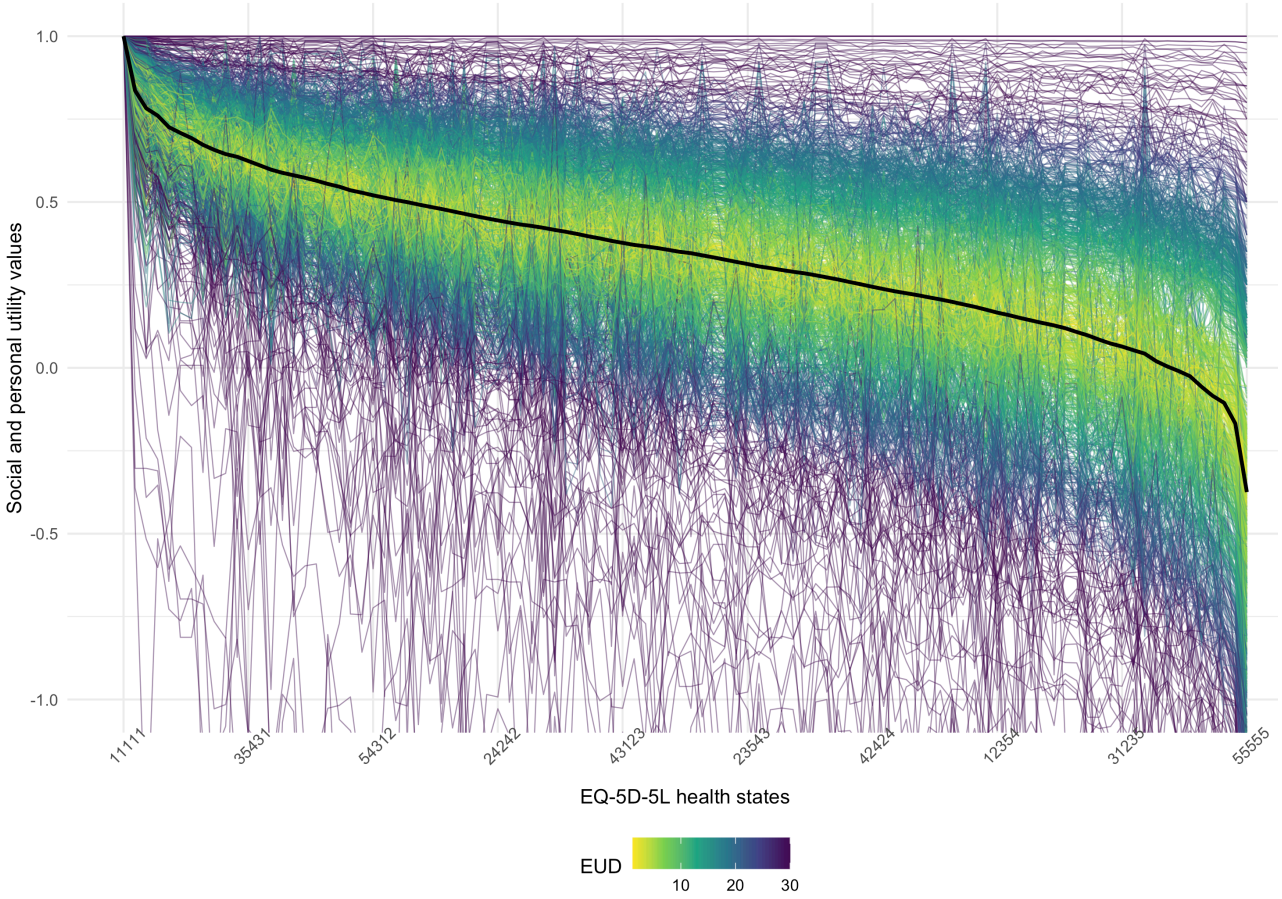


**FIGURE 2** simplified illustration of the aggregate group preference (thick black line) and the PUFs of all 874 participants. Shown are the utility values for a sample of 100 health states, ranked from the best on the left to the worst on the right (according to the aggregate group preference). The colours of the individual PUF lines indicate their euclidean distance from the average preference. Values below -1 are not shown.

### 3.6 PERMANOVA

Table 3 provides the results of the PERMANOVA. Shown are the within-group sum-of-squares ($SS_W$) for each group individually and for all groups combined, and the corresponding $R^2$, pseudo F, and p values. The between groups sum-of-squares ($SS_B$) can be computed by subtracting the $SS_W$ from the $SS_T$.

Significant differences between groups were observed for four group characteristics: age, having children, importance of religion/spirituality, and EQ-VAS quintiles. In addition, the effect of currently experiencing severe health problems ('present own experience') was borderline significant (p =0.0504). However, the proportions of the variance that were explained by these group characteristics individually were rather small: $R^2$ values ranged between 2.6% (for age) and 1.2% (for importance of religion/spirituality). It may be interesting to note that, contrary to our expectations, the effects of group characteristics that reflected experience with health problems (e.g. being a healthcare professional, carer) were not statistically significant. The model that included all group characteristics explained 8.5% of the differences between participants' PUFs.

**TABLE 3** Results of PERMANOVA – testing for differences in EQ-5D-5L health state preferences between groups characteristics

| Group variable | $SS_W$ | Df | $R^2$ | F | p |
|---|---|---|---|---|---|
| Sex | 473 | 2 | 0.1% | 0.44 | 0.630 |
| Age | 12180 | 6 | 2.6% | 3.85 | 0.008* |
| Having children | 7877 | 2 | 1.7% | 7.43 | 0.008* |
| Education | 4142 | 6 | 0.9% | 1.29 | 0.238 |
| Income | 4160 | 5 | 0.9% | 1.55 | 0.166 |
| Importance of religion/spirituality | 5708 | 4 | 1.2% | 2.67 | 0.034* |
| Religious/spiritual practice | 5698 | 6 | 1.2% | 1.78 | 0.098 |
| Experience w/ health problems | | | | | |
| Health care professional | 410 | 1 | 0.1% | 0.76 | 0.373 |
| Carer | 188 | 1 | 0.0% | 0.35 | 0.569 |
| Family member | 146 | 1 | 0.0% | 0.27 | 0.633 |
| Past own experience | 179 | 1 | 0.0% | 0.33 | 0.582 |
| Present own experience | 1977 | 1 | 0.4% | 3.69 | 0.050 |
| No experience | 180 | 1 | 0.0% | 0.33 | 0.586 |
| EQ-VAS (quintiles) | 5699 | 4 | 1.2% | 2.67 | 0.027* |
| All groups together | 39918 | 48 | 8.5% | 1.60 | 0.031* |
| Total ($SS_T$) | 469540 | 873 | | | |

$SS_T$ = total sum-of-squares; $SS_W$ = within-group sum-of-squares; df = degrees of freedom; F = pseudo F statistics; p values based on 10,000 permutations; * = p<0.05

To give some intuition for kind of differences that existed between groups, the (sub)group-specific value sets for different age groups are shown in figure 3 as an example. The colours of the plotted group-level (thick lines) and personal utility functions (thin lines) indicate group membership. For simplicity, the 'prefer not to say' group is not shown.

The age group specific value sets differ from each other in two ways. Firstly, there appears to be some differences in *scale*.The curve for the youngest group (age 18-29) is the lowest. The curve then seem to move upwards with increased age, and the curve for the oldest age group (70+) is the highest. This suggests that the older the participants are, the higher they set their anchor point. Secondly, the group-specific curves are not strictly decreasing, i.e. they move up and down and fluctuate, especially the curves for the oldest and for the youngest group. This indicates differences in the relative importance of health state attributes, i.e. groups assign different weights to the five EQ-5D-5L dimensions and/or differ in their level ratings. It should be noted that due to the simplified visualisation of EQ-5D-5L utility functions (we only show 100 of the 3125 utility scores) this effect may appear smaller in the figures than it actually is.

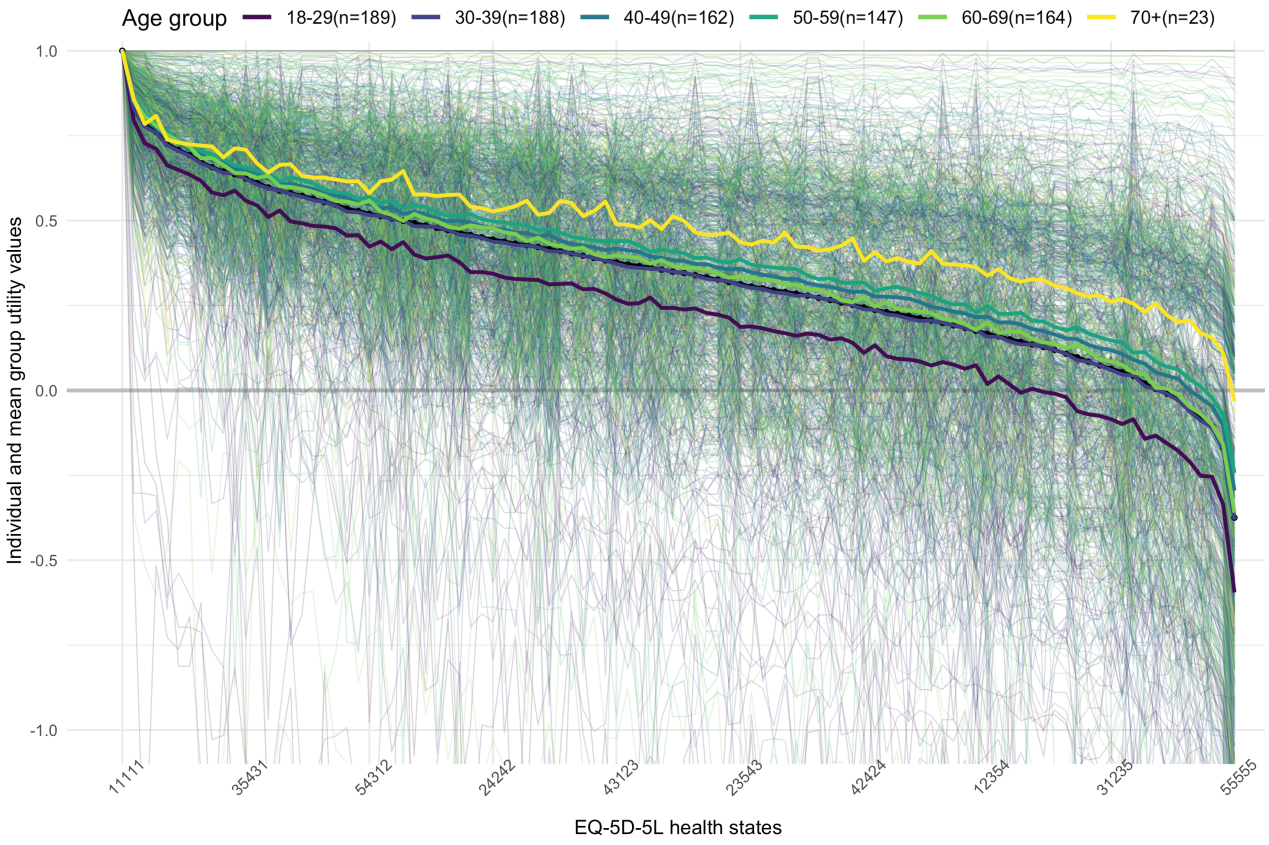


**FIGURE 3** Age-specific EQ-5D-5L health state preferences. Shown are the group level value sets (thick lines) and the underlying PUFs (thin lines), as well as the social value set (thick black line). Values below -1 and the 'prefer not to say' group are not shown.

### 3.7 K-means cluster analysis

Figure 4 shows the the proportion of the variance explained as a function of the number of k-means clusters. The $R^2$ values for 1 to 10 clusters were 0%, 44%, 66.4%, 78.0%, 81.6%, 89.1%, 92.1%, 93.7%, 94.4%, and 94.5%. The respective marginal changes of moving from 1 to 2 clusters, from 2 to 3 clusters, etc, were 44%, 22.4%, 11.6%, 3.6%, 7.5%, 3%, 1.6%, 0.7%, and 0.1%.

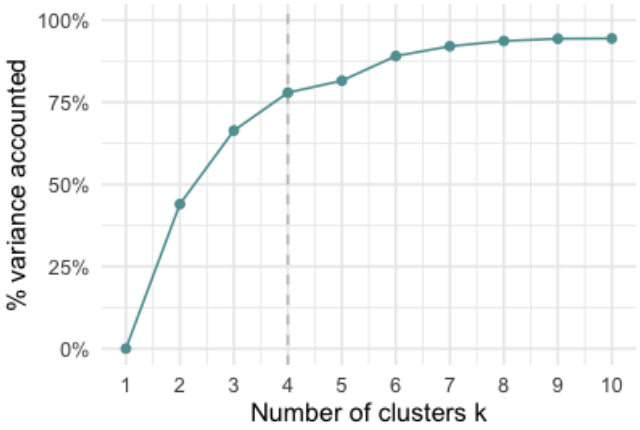


**FIGURE 4** Proportion of the variance explained as a function of the number of k-means clusters

Based on these results, we determined that the most plausible elbow point was at k=4. We thus partitioned the data into four clusters, and plotted the cluster-specific value sets alongside the underlying PUFs and the overall social value set of the full sample (black) – see figure 5. The clusters could be characterised as: 1) participants who considered 'being dead' the worst health state (blue, n=345), 2) participants average preferences (purple, n=386), 3) participants with low anchor points (yellow, n=131), 4) outliers, participants with very low anchor points (green, n=12).

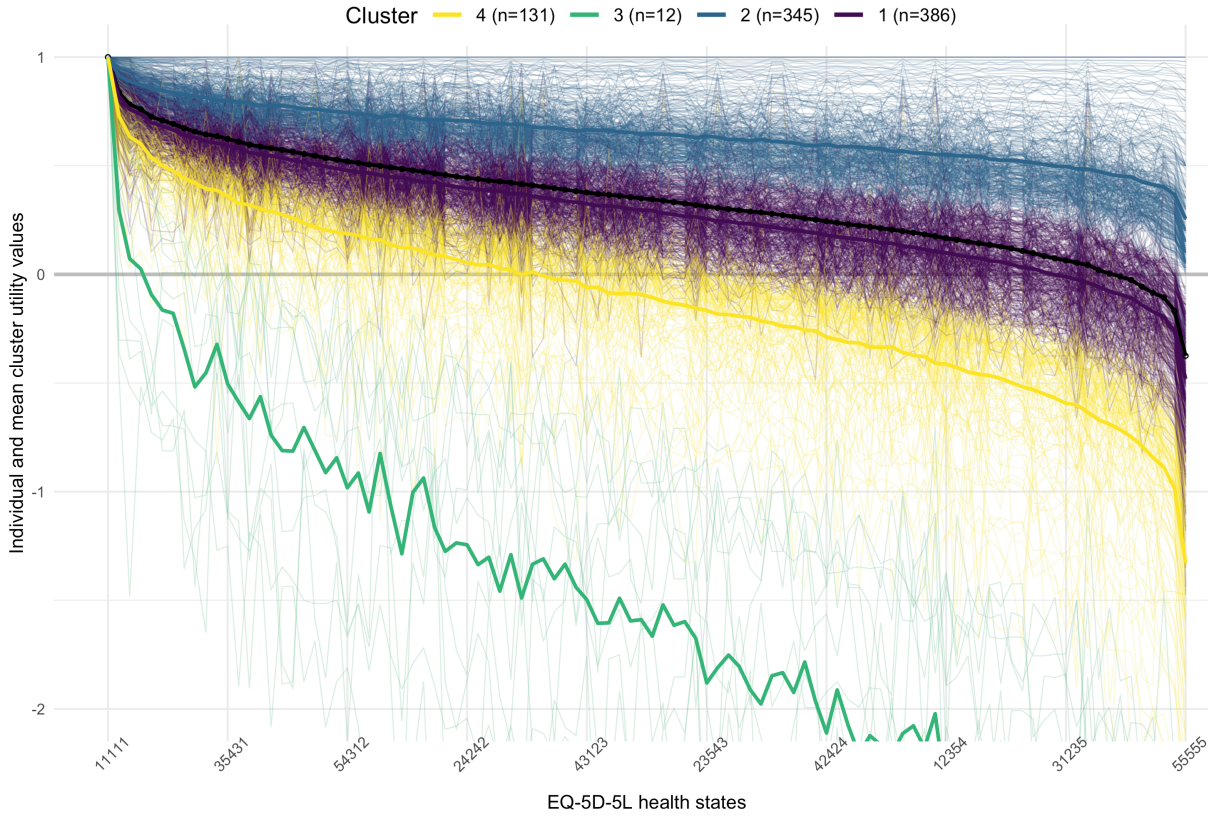


**FIGURE 5** Results of the k-means cluster analysis. Shown are the cluster-specific value sets (thick lines) and the underlying PUFs (thin lines), as well as the social value set (thick black line). Values below -2 are not shown.

## 4 DISCUSSION

This study is the first application of the newly developed OPUF approach for eliciting health state preferences in a large sample of the UK population. We constructed EQ-5D-5L value sets on the societal-, group-, and individual person level, to explore the, hitherto largely ignored, heterogeneity of health state preferences.

We found that health state preferences systematically differed between groups. Significant effects were observed in the PERMANOVA for age, having children, importance of religion/spirituality, and the EQ-VAS quintile. However, the variability of preferences within groups was substantial, and individual group characteristics explained only small proportions of the ED between PUFs. For other demographic factors (sex, education, income), we observed no systematic differences between groups. Contrary to our expectations, participants' experience with severe health problems (captured by 6 non mutually exclusive categories) were also not associated with the differences in PUFs. It should be noted though, that the participants in our sample were quite 'healthy' – a large majority reported no or only slight problems in any of the EQ-5D dimensions. If, and if so, how, patients' preferences differ from the preferences of members of the general public should be further investigated in future applications of the OPUF approach.

When all characteristics were taken into account together, group membership accounted for just 8% of the variance. This result should not be considered surprising. The formation of health preferences is a complex task, which is likely to be influenced by various emotional, cognitive, and social factors (Russo et al. 2019). There is no compelling reason why demographic factors, such as age, should be good predictors of people's health preferences. It may be a trivial point to make, but we would like to add that the results also illustrate that aggregate group-level value sets usually say little about the preferences of any given individual – in our study, preferences differed greatly between individuals within all the groups that we considered.

The findings from the cluster analysis provide additional insight into the heterogeneity of EQ-5D-5L health state preferences. The shapes of the identified cluster-specific value sets appear to confirm that most of variability between PUFs is determined by their scale, i.e. their anchor points. In comparison, the differences in the relative importance of health state attributes (i.e. dimension weights and level ratings) seem to be rather insignificant.

The main reason for this is probably the inherent structure of the EQ-5D-5L instrument: any given health state is dominating all states which are worse on at least one dimension, and not better on another. State '21111', for example, is dominated by '11111' and is itself dominating 2,499 other states ('31111', '21211', '31211', etc.). A simulation study by Ombler et al. (2018) found that the lowest correlation between any two (randomly created)

EQ-5D-5L value sets was 0.45. This covariance structure limits the extent to which preferences can differ and may explain why the *scale* seems to be by far the most important driver of the differences between individuals.

Our study has some limitations that should be considered when interpreting the findings.

Firstly, the participants that were included in the analysis were younger and more highly educated than the general UK population. The reported mean EQ-5D-5L model coefficients may not yield a representative social value set.

Secondly, preference heterogeneity can be investigated in many different ways. Designing this study thus required making several, somewhat contingent methodological choices. Instead of computing the ED between health state utility vectors, we could have assessed the differences in participants' model coefficients, or we could have computed a different distance measure – the Kendall correlation distance, for example, could be used to compare preference orderings (i.e. ordinal instead of cardinal preferences). Results may not be robust to these kinds of methodological choices.

Thirdly, we explored the variability of EQ-5D-5L health state preferences in a general sense. This means, we neither specified any hypotheses about the type or the direction of differences, nor did we test differences between subgroups. Even though the OPUF approach would have allowed us to study the health state preferences of small subgroups, in the absence of predefined hypotheses about subgroup differences, it did also not seem useful to consider the (up to 240) interaction effects between groups. To answer more specific research questions, such as, '*do older people with strong religious beliefs people assign higher utility values to health states than the general public?*' a different analytical approach may be required.

Finally, a key consideration for the interpretation of our findings is the validity of the OPUF approach. It is a new method, based on a different paradigm (compositional approach) than other, established preference elicitation methods, such as TTO, DCE, or SG (decompositional). Even though we observed a high consistency between constructed PUFs and participants' choices in DCEs 78%, more research is needed to better understand how the OPUF approach compares to other methods, and to determine how the online survey design affects participants' preference formation. Further refinement of the survey may also be help to prevent people from skipping essential valuation tasks, and thereby reduce the number of participants who have to be excluded from the analysis.

The OPUF approach provides a flexible, conceptually attractive, alternative approach for eliciting health state preferences. The ability to construct utility functions on the individual person level opens up new and, we think, exciting avenues for research. As demonstrated

in this study, the OPUF approach makes it possible to investigate the heterogeneity of health states preferences in an unprecedented level of detail. It may also enable researchers to derive value sets for small groups of participants (e.g. patients with rare diseases), for which this would otherwise be practically infeasible. Even though the OPUF approach has, thus far, only been implemented for the EQ-5D-5L, in principle, it could be applied to any descriptive system or patient-reported outcome measure.

**Funding and acknowledgements**

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

We are very grateful to Siobhan Daley, Jack Dowie, Barry Dewitt, Irene Ebyarimpa, Paul Kind, Johanna Kokot, Simon McNamara, Clara Mukuria, Monica Oliveira, Krystallia Pantiri, Donna Rowan, Erik Schokkaert, Koonal Shah, Robert Smith, Praveen Thokala, Ally Tolhurst, David Tordrup, Evangelos Zormpas, and the participants of the 2021 Summer HESG virtual meeting for helpful comments, discussions of the ideas expressed in this paper, and/or for providing feedback on earlier versions of the EQ-5D-5L OPUF survey. We would also like to thank all participants who took part in this study. The usual disclaimer applies.

**Ethical approval**

The study was approved by the Research Ethics Committee of the School of Health and Related Research at the University of Sheffield (ID: 030724).

**REFERENCES**

Anderson MJ. Permutational multivariate analysis of variance (PERMANOVA). Wiley statsref: statistics reference online. 2014 Apr 14:1-5.

Anderson MJ, Walsh DC. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing?. Ecological monographs. 2013 Nov;83(4):557-74.

Belton V, Stewart T. Multiple criteria decision analysis: an integrated approach. Springer Science & Business Media; 2002.

Brazier J, Ratcliffe J, Saloman J, Tsuchiya A. Measuring and valuing health benefits for economic evaluation. OXFORD university press; 2017a.

Brazier J, Ara R, Rowen D, Chevrou-Severac H. A review of generic preference-based measures for use in cost-effectiveness models. Pharmacoeconomics. 2017b Dec;35(1): 21-31.

Devlin NJ, Shah KK, Mulhern BJ, Pantiri K, van Hout B. A new method for valuing health: directly eliciting personal utility functions. The European Journal of Health Economics. 2019 Mar;20(2):257-70.

Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health−related quality of life: An EQ−5 D−5 L value set for E ngland. Health economics. 2018 Jan;27(1):7-22.

Feng Y, Devlin NJ, Shah KK, Mulhern B, van Hout B. New methods for modelling EQ−5D−5L value sets: An application to English data. Health Economics. 2018 Jan;27(1):23-38.

Golicki D, Jakubczyk M, Graczyk K, Niewada M. Valuation of EQ-5D-5L health states in Poland: the first EQ-VT-based study in Central and Eastern Europe. Pharmacoeconomics. 2019 Sep;37(9):1165-76.

Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Quality of life research. 2011 Dec;20(10):1727-36.

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

Ombler F, Albert M, Hansen P. How significant are "high" correlations between EQ-5D value sets?. Medical Decision Making. 2018 Aug;38(6):635-45.

Oppe M, Van Hout B. The "power" of eliciting EQ-5D-5L values: the experimental design of the EQ-VT. EuroQolWorking Paper Series. 2017 Oct;17003.

Palan S, Schitter C. Prolific. ac—A subject pool for online experiments. Journal of Behavioral and Experimental Finance. 2018 Mar 1;17:22-7.

Peeters Y, Stiggelbout AM. Health state valuations of patients and the general public analytically compared: a meta-analytical comparison of patient and population health state utilities. Value in Health. 2010 Mar 1;13(2):306-9.

Russo S, Jongerius C, Faccio F, et al. Understanding patients' preferences: a systematic review of psychological instruments used in patients' preference and decision studies. Value in Health. 2019 Apr 1;22(4):491-501.

Schneider P, van Hout B, Heisen M, Brazier J, Devlin N. A new online tool for valuing health states: eliciting personal utility functions for the EQ-5D-5L. Available online: https://github.com/bitowaqr/bitowaqr.github.io/raw/master/files/opuf2.pdf

Souza AT, Dias E, Nogueira A, Campos J, Marques JC, Martins I. Population ecology and habitat preferences of juvenile flounder Platichthys flesus (Actinopterygii: Pleuronectidae) in a temperate estuary. Journal of Sea Research. 2013 May 1;79:60-9.

Sullivan T, Hansen P, Ombler F, Derrett S, Devlin N. A new tool for creating personal and social EQ-5D-5L value sets, including valuing 'dead'. Social Science & Medicine. 2020 Feb 1;246:112707.

Thokala P, Devlin N, Marsh K, et al. Multiple criteria decision analysis for health care decision making—an introduction: report 1 of the ISPOR MCDA Emerging Good Practices Task Force. Value in health. 2016 Jan 1;19(1):1-3.

Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. British medical bulletin. 2010 Dec 1;96(1):5-21.