PharmacoEconomics (2020) 38:575–591
https://doi.org/10.1007/s40273-020-00893-8

ORIGINAL RESEARCH ARTICLE

A Comparison of the EQ-5D-3L and EQ-5D-5L

Alexander James Thompson[1] · Alex James Turner[2]

Published online: 17 March 2020

Abstract

Introduction The EQ-5D-3L (3L) and EQ-5D-5L (5L) are both frequently used measures of health status. Previous studies have found the EQ-5D-5L to have superior measurement properties but no study has compared the two measures in a large general population survey using matched respondents.

Methods Using data from the GP Patient Survey, coarsened exact matching was used to match individuals completing the 3L in 2011 with those completing the 5L in 2012. Measurement properties were assessed for a general population and multimorbid population (chronic conditions ≥ 2), with ceiling effects, informativity and distribution of response compared. Changes in the direction of response, as well as the impact on utility distributions, were quantified.

Results Matching resulted in a cohort of 1,023,218 respondents (2011: 511,609; 2012: 511,609) for analysis. Ceiling effects for the 5L were lower than the 3L (43.8% vs. 54.4%). The 5L had improved informativity and broader spread of responses than the 3L (5L top 50 profiles: 77.4% vs. 3L: 98.8%). Overall, there was an upwards shift in utility values for the 5L versus the 3L as respondents using the 5L reported ill health more frequently but with less severity. Measurement improvements and effects on utility values were more pronounced for the multimorbid population.

Conclusion The 5L had superior measurement properties than the 3L and should be preferred in general population surveys and for use in individuals with multimorbidity. At increasing levels of morbidity, the 5L is currently associated with higher utility values than the 3L.

Key Points

The EQ-5D-5L has superior measurement properties to the EQ-5D-3L and should be preferred in general population and multimorbid populations.

The EQ-5D-5L leads respondents to report more ill-health. However, it is less severe than when using the EQ-5D-3L.

For higher-levels of multimorbidity, use of the EQ-5D-5L in England will currently lead to higher utility values than the EQ-5D-3L, irrespective of the tariff used.

1 Introduction

The EQ-5D is a frequently used measure of health status that is instrumental in calculating quality-adjusted life-years (QALYs) for cost-effectiveness analysis. The first version of the EQ-5D covers five domains (mobility, self-care, usual activities, pain or discomfort, and anxiety or depression), with each having three levels (no problems, moderate problems, or extreme problems) [1]. This gives a total of 243 unique health states from which, using a preference-based tariff, it is possible to generate utility values to calculate QALYs.

Widespread use of the three-level version of the EQ-5D (EQ-5D-3L) highlighted persistent problems with its measurement properties. Most notable are so-called 'ceiling effects' whereby a large proportion of individuals within a sample state they have 'no problems' in all domains, which may indicate a lack of discriminatory power to detect ill health in individuals with mild conditions. In response, the EQ-5D-5L was developed that kept the same number of domains in an almost identical format but which incorporated five-levels for each domain (no problems, some problems, moderate problems, severe problems or extreme problems) [2].

A recent systematic review of studies comparing the performance of EQ-5D-5L and EQ-5D-3L found the EQ-5D-5L

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s40273-020-00893-8) contains supplementary material, which is available to authorized users.

✉ Alexander James Thompson
alexander.thompson@manchester.ac.uk

[1] Manchester Centre for Health Economics, University of Manchester, Manchester, UK

[2] Health Organisation, Policy, and Economics (HOPE) Group, Centre for Primary Care and Health Services Research, University of Manchester, Manchester, UK

to have either similar or better measurement properties than the EQ-5D-3L [3]. Importantly, there was evidence of reduced ceiling effects with the EQ-5D-5L and also a better spread of respondents using all levels within each domain. By reducing the clustering around full health and improving the distribution of respondents across levels, the EQ-5D-5L has potentially greater scope to capture smaller changes in health status. The majority of previous studies have been 'head-to-head' (H2H) comparisons using responses for both measures completed by the same individuals [3]. There is also a subset of 'indirect' studies [4, 5] that compare responses using different individuals, potentially allowing for a much larger sample size; however, a key challenge for previous indirect studies are that respondents have been different on observable characteristics, potentially biasing any findings.

The primary aim of this study was to compare the measurement properties of the EQ-5D-3L and EQ-5D-5L, drawing on data from a large survey of the general population. We examined how the descriptive system impacts ceiling effects, discriminatory power and the distributional characteristics of both EQ-5D responses and EQ-5D utilities. We used matching methods to ensure comparability in 3L and 5L respondents using characteristics that could be predictive of the level, or reporting, of health. We also exploited this large dataset to explore heterogeneity in measurement properties in individuals with multimorbidity and for increasing numbers of health conditions.

2 Methods

2.1 Data

Data were drawn from the General Practice Patient Survey (GPPS), a survey designed to provide information on individuals' experience of services provided by general practices in England [6]. Approximately 2.6 million surveys are sent annually to randomly selected adults (18+ years) who must have been registered with a practice for at least 6 months [7]. As almost the entirety of the English population is registered with a GP practice [8], the GPPS represents a general population sample. The annual response rate is approximately 35%, resulting in approximately 1 million individuals per year. The GPPS is a repeated cross-section, meaning that although it is possible for individuals to be randomly selected in different years, the same individual cannot be tracked over time.

The GPPS collects information on individual sociodemographic characteristics and multiple measures of individual's health, including the EQ-5D. Our study used data from the 2011 and 2012 surveys years. Importantly, between these years the EQ-5D version changed from the EQ-5D-3L to the EQ-5D-5L.

We restricted our sample to individuals with complete data on EQ-5D and relevant health and sociodemographic characteristics. Missing data were not imputed. The total eligible sample size was 1,411,680 individuals (2011: 718,239; 2012: 693,441).

2.2 Matching

We used coarsened exact matching (CEM) to ensure individuals in the 2011 (EQ-5D-3L) and 2012 (EQ-5D-5L) samples were identical on characteristics that are either predictive of health or predictive of how health is reported [9]. CEM first coarsens continuous variables into bins, with binary or categorical variables left uncoarsened. Strata are then created for each unique combination of the coarsened and binary/categorical variables and exact matching is applied to these strata. Strata containing individuals in the 2011 sample only or 2012 sample only were pruned from the study sample. *k-to-k* matching was used to ensure the same number of individuals from the 2011 and 2012 years were included in each strata.

CEM improves on exact matching by reducing the number of unmatched observations when matching on continuous variables [10], and outperforms other matching methods on a variety of other criteria [11]. Information on characteristics used in the matching process and the chosen degree of coarsening are provided in Table 1, including individual-level sociodemographic characteristics and measures of health and health behaviours. Area-level deprivation was measured using the overall score of the 2010 Index of Multiple Deprivation (IMD) [12]. The sole continuous variable, the IMD score, was coarsened into quintiles.

To measure balance in the distributions of characteristics following matching, we computed differences in the mean, median, range and 25th and 75th quantile between the 2011 and 2012 samples. We additionally use the L_1 statistic to examine balance on the joint distribution of all characteristics [9]. Further details on CEM and the L_1 statistic are provided in Electronic Supplementary Appendix 1.

2.3 Analysis

We conducted all our analysis using both the whole sample (i.e. general population analysis) and by considering individuals with multimorbidity. We followed the most commonly used definition of multimorbidity, i.e. an individual with two or more long-term health conditions [13]. All analyses were conducted in Stata version 14 (StataCorp LLC, College Station, TX, USA).

2.3.1 Distributional Properties of EQ-5D Responses

We first examined how the change to the EQ-5D-5L altered average response patterns. This was done by comparing

Table 1 Matching variables and the degree of coarsening

Variable	Categories	Coarsening
Sex	Male, female	None
Age group (years)	18–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75–84, 85+	None
Ethnic group	White, mixed, Asian, Black, other	None
Long-term health conditions		
Alzheimers disease		None
Angina or long-term heart problem		None
Arthritis or long-term joint problem		None
Asthma or long-term chest problem		None
Blindness or severe visual impairment		None
Cancer in the last 5 years		None
Deafness or severe hearing impairment		None
Diabetes		None
Epilepsy		None
High blood pressure		None
Kidney or liver disease		None
Learning difficulty		None
Long-term back problem		None
Long-term mental health problem		None
Long-term neurological problem		None
Another long-term condition		None
IMD score[a]		Quintiles
Economic activity	Employed (full- or part-time), permanently sick or disabled, unemployed, retired, other (student, looking after home, ‘something else’)	None
Health limits daily activity	A lot, a little, no	None
Smoking habits	Never smoker, former smoker, current smoker	None
Sexual orientation	Heterosexual/straight, gay/lesbian/bisexual/other/prefer not to say	None
Wave	January–March, July–September	None

IMD Index of Multiple Deprivation, *LSOA* lower super output area

[a] The IMD measures deprivation across multiple domains at an LSOA level. LSOAs are census groupings that contain approximately 1500 people

the percentage of respondents reporting each EQ-5D level in both descriptive systems. We also explored whether the introduction of additional levels reduced ceiling effects present in the EQ-5D-3L, by comparing the proportion of individuals reporting ‘no problems’ across all dimensions.

2.3.2 Inconsistency

We compared 3L and 5L responses within pairs of randomly matched individuals within an identical stratum and used matched responses to examine inconsistency. We recoded 3L responses to the equivalent for the 5L, e.g. 1 = 1, 2 = 3, and 3 = 5, and assumed that in the absence of the version change, the 5L and recoded 3L responses within matched pairs would have been identical. Thus, any differences in responses can be attributed to the version change.

Following Janssen et al. [14], we defined an inconsistent response for a domain as a response to the EQ-5D-3L that is at least two levels away from the EQ-5D-5L, e.g. an individual reporting ‘no pain or discomfort’ on the EQ-5D-3L but their matched respondent reporting ‘moderate pain or discomfort’ on the EQ-5D-5L. For total inconsistency, we defined an inconsistent profile as one where any individual domain is different by more than two levels. For example, the EQ-5D-3L (coded to EQ-5D-5L) profile 12221 would be consistent with the EQ-5D-5L profile 12223 but would be inconsistent with the EQ-5D-5L profile 12224.

For calculations of inconsistency, we repeated ($n = 10$) the process of randomly matching individuals within stratum. All results are based on mean values across all randomisations. We examined variation in inconsistency across different randomisations as an indirect test of the matching assumption. Similarity in changes when 3L respondents are matched to different 5L respondents (with potentially different unobserved characteristics) provides evidence supporting the matching assumption.

2.3.3 Informativity

We used Shannon's indices to examine the relative discriminatory power of the 3L and 5L [3]. These indices are derived from informational theory and assume that a measure provides the most information if responses are distributed equally across all response categories. They also provide a quantitative measure of the degree of response redistribution due to the additional categories. We first calculated the Shannon–Weaver index (H'), which is calculated as (Eq. 1):

$$H' = -\sum_{i=1}^{L} p_i \log_2 p_i, \tag{1}$$

where L denotes the total number of response categories, and p_i is the proportion of individuals reporting health in response category i. H' ranges from zero (least informative) to $\log_2 L$ (most informative), the latter indicating proportions in each response category are identical. However, it is difficult to compare H' across measures as its maximum level is determined by the number of levels ($\log_2 3 = 1.58$ for the 3L and $\log_2 3 = 2.32$ for the 5L). We therefore additionally computed the Shannon's Evenness index (J'), which scales H' by the maximum H' for a measure with the same number of response categories (Eq. 2):

$$J' = \frac{H'}{H'_{\max}}. \tag{2}$$

We computed $H'$and J' for response categories for each individual domain.

We also explored how the full descriptive systems are used by calculating the proportion of health states selected by individuals for the EQ-5D-3L and EQ-5D-5L. We rank the health states in terms of frequency selected and plot the cumulative frequency.

2.3.4 Response Change

We first measured the impact on responses by assessing the sensitivity of the two measures for picking up small movements away from full health for both the general population and for those reporting having no chronic health conditions. We calculated the proportion of individuals reporting 'no problems' for each domain and stratified the results by the response to the question "Have your activities been limited today because you have recently become unwell or been injured? ('no limitations'/'some limitations')". We calculated the change in the proportion reporting 'no problems' when moving from no limitations to some limitations for each domain and then calculated the difference in the changes between the 3L and the 5L to calculate a difference-in-difference. A positive difference-in-difference is indicative of greater sensitivity of the 5L compared with the 3L.

To provide a consideration of response change across all levels, we also used matched pairs from the inconsistency analysis to examine whether the version change caused individuals overall to select levels indicative of poorer or better health. We did this first at a domain-level by comparing mean level responses between the recoded EQ-5D-3L (1 = 1; 2 = 3; 3 = 5) and the EQ-5D-5L. Mean level response differences are calculated for matched pairs by subtracting EQ-5D-3L from EQ-5D-5L responses, with positive scores implying lower levels of health are reported with the EQ-5D-5L and negative scores implying lower levels of health are reported on the EQ-5D-3L. We then compared the completion of the total profile by comparing EQ-5D-3L and EQ-5D-5L respondents on a 'misery index' that is equivalent to the sum of the levels across all domains. Calculations of response change are based on randomly matched samples of individuals within strata.

For a selection of matched strata where total profile inconsistency was below 2.5%, we also depicted flows in responses between the EQ-5D-3L and EQ-5D-5L. To do this, we randomly matched individuals across the EQ-5D-3L and EQ-5D-5L datasets who were in the same strata, i.e. those who have identical matched characteristics apart from being given the two different EQ-5D versions across 2 different years. We then calculate cross-tabulations, showing for each level and domain of the EQ-5D-3L the corresponding level and domain for the EQ-5D-5L. We then depicted these cross-tabulations graphically using 'Sankey' diagrams [15, 16].

2.3.5 Impact on Utility Indices

We examined the consequences of the EQ-5D version change on the distribution of utility scores by generating utility values from the EQ-5D-3L using the value set derived by Dolan [17], and values for the EQ-5D-5L using the mapping algorithm developed by van Hout et al. [18]. We also calculated utility values for 5L respondents using the value set derived by Devlin et al. [19]. We compared the utility distributions by examining differences in kernal density and exploring the impact on mean utility for condition counts of up to five concurrent conditions.

3 Results

Implementing CEM resulted in 280,842 strata, of which 66,398 contained both EQ-5D-3L and EQ-5D-5L respondents. The sample size following matching was 1,023,218 (2011: 511,609; 2012: 511,609). Differences in the characteristics of 3L and 5L respondents at the mean prior to and following matching are shown in Table 2, alongside summary statistics for the matched cohort. Differences in characteristics at the minimum, maximum and 25th, 50th and 75th percentiles are shown in Tables A1 and A2 (Electronic

Supplementary Appendix 2). There was a reduction in imbalance following matching. The mean difference in the IMD score dropped considerably and imbalance at all other points of the distribution were reduced. Given exact matching is conducted for binary and categorical variables, the distributional characteristics of all other covariates were now identical for EQ-5D-3L and EQ-5D-5L respondents. The L_1 statistic also fell from 0.496 to 0.394, indicating a reduction in imbalance in the joint distribution of all characteristics.

3.1 Distributional Properties of EQ-5D Responses

In the general population, there was a reduction in those selecting 'no problems' for the EQ-5D-5L versus the EQ-5D-3L, with the biggest fall in the usual activities domain (Fig. 1). Using the EQ-5D-5L, the new 'slight problems' level was always more than twice as frequently selected as the 'moderate problems' level, apart from in self-care where it was still more frequent. The proportion of those reporting 'some problems' despite also reporting having 'no conditions' fell when using the EQ-5D-5L, with far more reporting 'slight problems' instead (Fig. 2). Overall, the proportion selecting 'no problems' for all domains was 54.4% for the EQ-5D-3L and 43.8% for the EQ-5D-5L, suggesting a mean absolute fall in 'ceiling effects' for the EQ-5D-5L of 10.6 percentage points.

For the multimorbid population, the number of individuals selecting 'moderate problems' more than halved for the EQ-5D-5L compared with those selecting 'some problems' for the EQ-5D-3L for most domains (Fig. 1). Instead, more individuals made use of the two new levels (slight problems and severe problems), particularly as multimorbidity increased (Fig. 2). For those with at least two conditions, the new 'severe problems' level was always more frequently selected than extreme problems. Overall, the number of multimorbid individuals selecting no problems for all domains was 20.4% for the EQ-5D-3L and 14.1% for the EQ-5D-5L, suggesting a mean absolute fall in 'ceiling effects' for the EQ-5D-5L of 6.3 percentage points.

3.2 Inconsistency

For the general population, inconsistency varied markedly by domain, with the lowest inconsistency being for the self-care (3.8%) domain and the highest inconsistency being in the pain and discomfort (18.1%) domain (Table 3). Overall profile inconsistency was 4.4% between the EQ-5D-3L and EQ-5D-5L (differed by more than two levels in any domain). Inconsistency was higher in the multimorbid population versus the general population for all domains and was lowest in the self-care (12.8%) domain and highest in the usual activities domain (26.2%). Overall profile inconsistency was 9.8% between the EQ-5D-3L and EQ-5D-5L for the multimorbid population. Inconsistency results were robust to the random pairing used within stratum providing support for the matching assumption.

3.3 Informativity

Using Shannon's absolute measure of informativity (H'), the EQ-5D-5L produced higher values, indicating it was consistently more informative than the EQ-5D-3L in each domain for both the general population and the multimorbid population (Table 4). However, in the general population, the relative measure of informativity (J'), which adjusts for the different number of levels between the measures, revealed the EQ-5D-5L and EQ-5D-3L to be approximately equal for the self-care, pain and discomfort, and anxiety and depression domains. There was a large increase in informativity using the EQ-5D-5L for the usual activities domain and a small increase for the mobility domain. For individuals with multimorbidity, relative informativity was higher for the EQ-5D-5L for all domains apart from anxiety and depression.

In both the general population and multimorbid populations, there was a much wider spread of profiles used for the EQ-5D-5L versus the EQ-5D-3L (Fig. 3). This was much more apparent for the multimorbid population, where the top 50 profiles for the EQ-5D-3L represented 98.8% of all the profiles selected, whereas for the EQ-5D-5L, the top 50 profiles represented only 77.4% of profiles selected.

3.4 Response Change

In individuals who responded 'no' to the question "Have your activities been limited today because you have recently become unwell or been injured?", the EQ-5D-5L was more sensitive in capturing ill health versus the EQ-5D-3L, with fewer individuals reporting 'no problems' across each domain (Fig. 4). In individuals who responded as 'limited a little' to the same question, the EQ-5D-5L had a greater reduction in those reporting 'no problems' versus the EQ-5D-3L, suggesting an increased sensitivity to detect small deviations away from health states close to full health versus the EQ-5D-3L. The observed improvements in sensitivity for capturing ill health was the same for both the general population and those who reported having no chronic health conditions.

When exploring response changes, considering the severity of levels selected, respondents using the EQ-5D-3L tended to select lower levels within the usual activities, pain and discomfort, and anxiety and depression domains versus the 5L, suggesting better reported health when using the EQ-5D-5L versus the EQ-5D-3L in the general population (Table 5). For the multimorbid population, respondents using the EQ-5D-3L tended to select lower levels on the usual activities and pain and discomfort domains. For both the general population and the multimorbid population, the misery index was lower for the EQ-5D-5L population versus

Table 2 Imbalance in the raw data and summary characteristics for the general population and those with multimorbidity

Characteristic [*N* (%)]	Raw data	Matched sample		
	Difference in mean	Difference in mean	General population [*n* (%)]	Multimorbid
Sex (male = 1)	< 0.00001	0	449,778 (43.96)	242,170 (47.31)
Age (years)				
18–24	– 0.00448	0	49,666 (4.85)	8226 (1.61)
25–34	– 0.00638	0	119,058 (11.64)	23,404 (4.57)
35–44	– 0.00863	0	155,950 (15.24)	44,130 (8.62)
45–54	– 0.00193	0	200,532 (19.60)	83,596 (16.33)
55–64	– 0.00115	0	214,378 (20.95)	127,294 (24.87)
65–74	0.01632	0	183,790 (17.96)	139,514 (27.26)
75–84	0.00578	0	83,526 (8.16)	71,332 (13.94)
85+	0.00047	0	16,318 (1.59)	14,386 (2.81)
Ethnic group				
White	– 0.00086	0	940,820 (91.95)	491,968 (96.11)
Mixed	< 0.00001	0	4040 (0.39)	554 (0.11)
Asian	– 0.00144	0	47,886 (4.68)	11,562 (2.26)
Black	0.00037	0	18,676 (1.83)	5498 (1.07)
Other	0.00202	0	11,796 (1.15)	2300 (0.45)
Long-term health conditions				
Alzheimer's disease	– 0.00027	0	1326 (0.13)	1326 (0.26)
Angina or long-term heart problem	0.00095	0	37068 (3.62)	37,068 (7.24)
Arthritis or long-term joint problem	0.00481	0	111160 (10.86)	111,160 (21.72)
Asthma or long-term chest problem	– 0.00035	0	75602 (7.39)	75,602 (14.77)
Blindness or severe visual impairment	– 0.00035	0	1934 (0.19)	1934 (0.38)
Cancer in the last 5 years	0.00248	0	23,644 (2.31)	23,644 (4.62)
Deafness or severe hearing impairment	0.00114	0	21,616 (2.11)	21,616 (4.22)
Diabetes	0.0054	0	53,040 (5.18)	53,040 (10.36)
Epilepsy	– 0.00062	0	3052 (0.30)	3052 (0.60)
High blood pressure	0.00966	0	201,960 (19.74)	201,960 (39.45)
Kidney or liver disease	0.00097	0	4010 (0.39)	4010 (0.78)
Learning difficulty	< 0.00001	0	1578 (0.15)	1578 (0.31)
Long-term back problem	0.00119	0	64,418 (6.30)	64,418 (12.58)
Long-term mental health problem	0.00207	0	16,868 (1.65)	16,868 (3.30)
Long-term neurological problem	0.00113	0	5802 (0.57)	5802 (1.13)
Another long-term condition	0.00799	0	103,770 (10.14)	103,770 (20.27)
Index of Multiple Deprivation	0.18207	0.0024		
Quintile 1 (least deprived)			229,685 (22.45)	115,498 (22.56)
Quintile 2			219,737 (21.48)	112,846 (22.05)
Quintile 3			208,836 (20.41)	106,326 (20.77)
Quintile 4			193,522 (18.91)	94,370 (18.44)
Quintile 5 (most deprived)			171,438 (16.75)	82,842 (16.18)
Economic activity				
Employed (full- or part-time)	– 0.01177	0	603,036 (58.94)	225,092 (43.97)
Permanently sick or disabled	– 0.00133	0	14,140 (1.38)	13,620 (2.66)
Unemployed	– 0.00016	0	30,860 (3.02)	10,856 (2.12)
Retired	0.0142	0	303,794 (29.69)	235,834 (46.07)
Other (student, looking after home, 'something else')	– 0.00094	0	71,388 (6.98)	26,480 (5.17)
Health limits daily activity				
A lot	0.00579	0	15,216 (1.49)	10,238 (2.00)
A little	< 0.00001	0	104,230 (10.19)	60,732 (11.86)

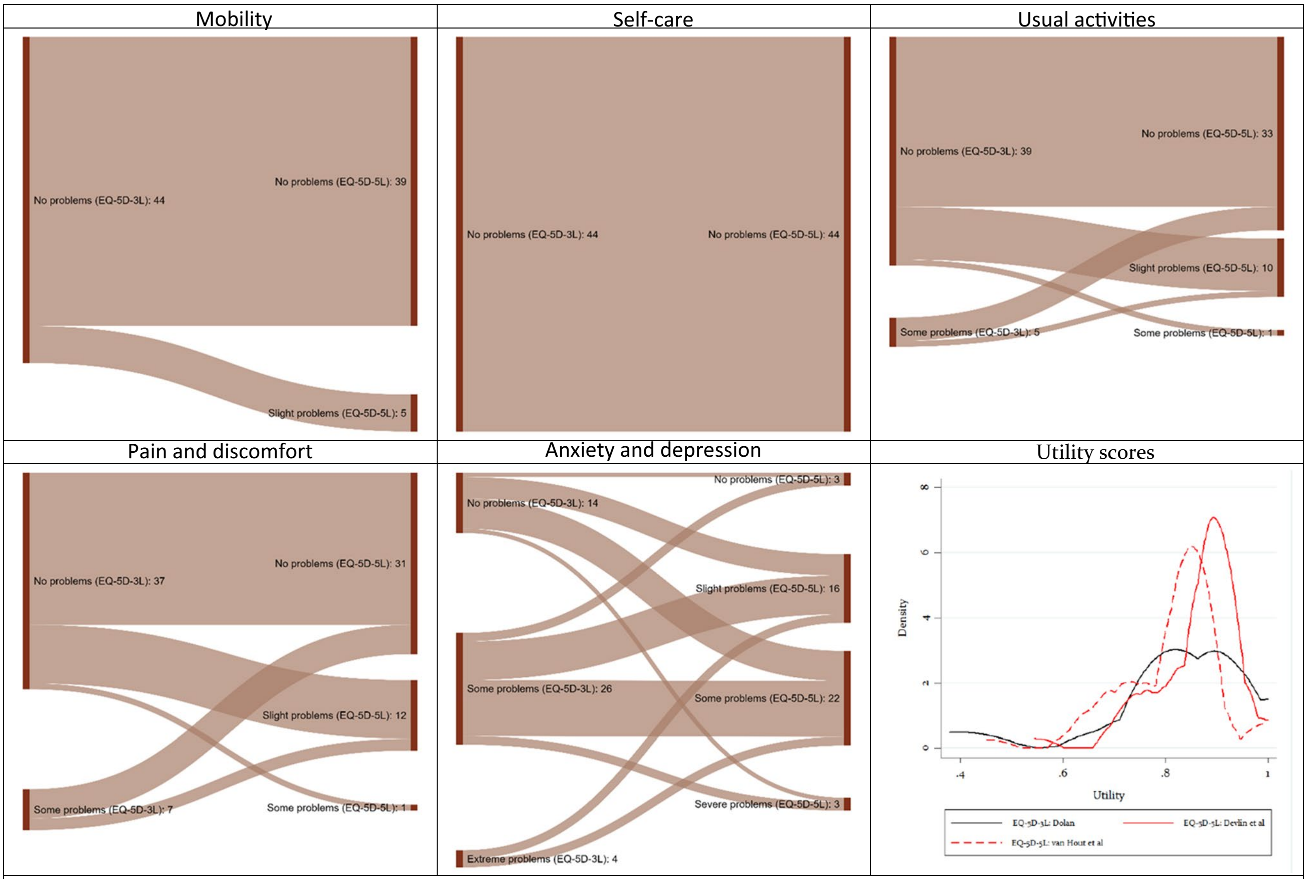
Table 2 (continued)

Characteristic [*N* (%)]	Raw data	Matched sample		
	Difference in mean	Difference in mean	General population [*n* (%)]	Multimorbid
No	− 0.00576	0	903,772 (88.33)	440,912 (86.14)
Sexual orientation (gay/lesbian/bisexual/prefer not to say = 1)			38,710 (3.78)	12,588 (2.46)
Smoking habits				
Never smoked	0.00711	0	562,222 (54.95)	254,370 (49.69)
Former smoker	− 0.0107	0	313,212 (30.61)	192,976 (37.70)
Current smoker	0.00359	0	147,784 (14.44)	64,536 (12.61)
Wave (July–September = 1)	0.02744	0	512,282 (50.07)	255,828 (49.98)

the EQ-5D-3L population, indicating that overall the EQ-5D-5L version led to respondents choosing levels indicative of better health. Response changes were robust to the random pairing used within stratum, providing further support for the matching assumption.

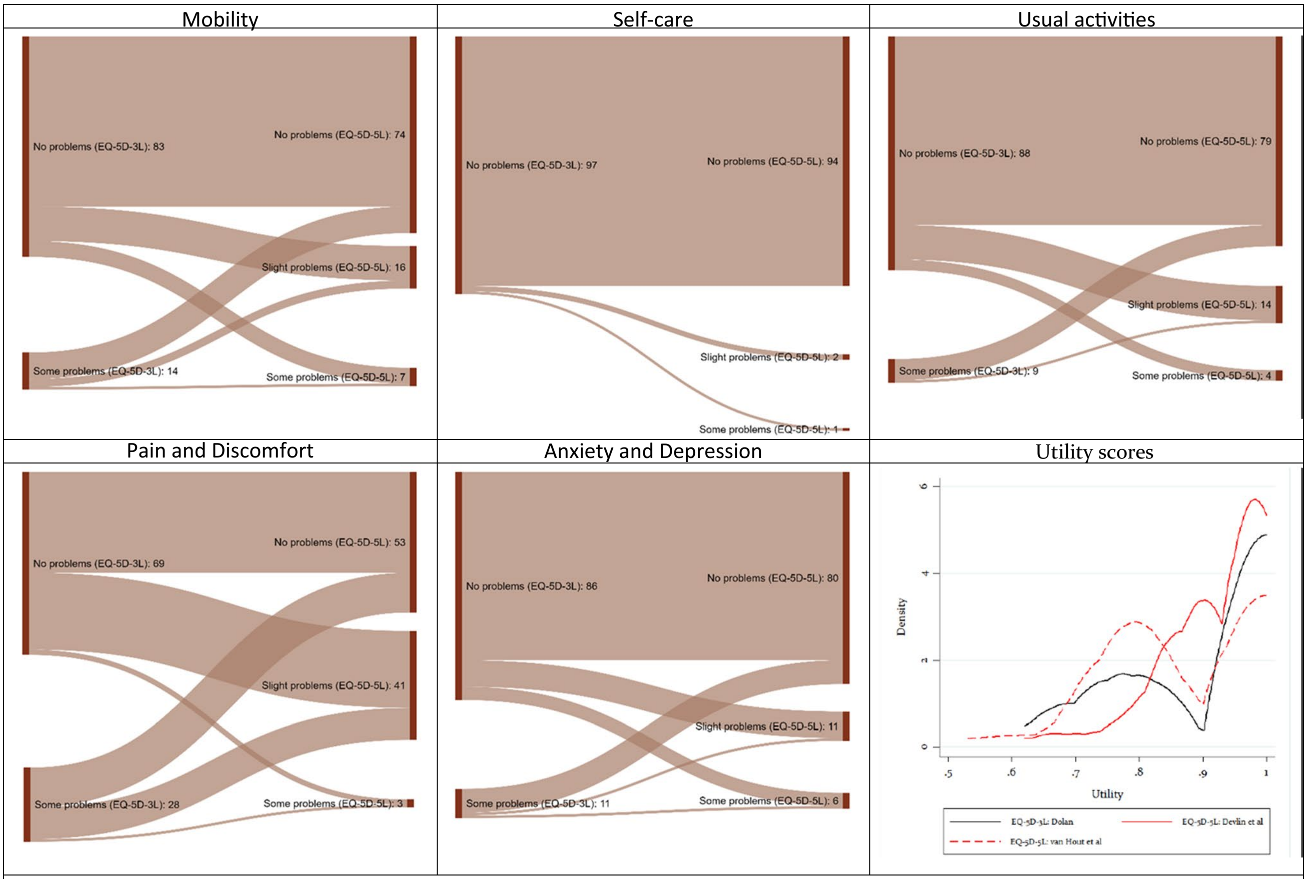
Box 1 and Box 2 present vignettes for matched stratum. For the 'working with depression' stratum (Box 1), there was a general shift towards levels suggestive of better health when moving from the EQ-5D-3L to the EQ-5D-5L. However, for the anxiety and depression domain, there was a general movement towards the 'middling' levels from the outer levels. For the 'retired with diabetes' stratum, across all domains, there was a movement towards the use of 'slight problems' in the EQ-5D-5L from both the 'no problems' and 'some problems' levels on the EQ-5D-3L. There was no comparative movement towards the 'extreme problems' levels on the EQ-5D-5L. This indicates heterogenous response changes across individuals with different characteristics.

Box 1: Sankey diagrams depicting response shifts and utility distributions for individuals "Working with depression[a]"



[a] A 'Sankey' diagram for a stratum where individuals completing the EQ-5D-3L and EQ-5D-5L in separate years have identical matched characteristics. Flows are calculated by a cross-tabulation of levels and domains between the 3L and 5L.Characteristics of stratum: Cohort n=88. Age: 45-64, Sex: Male, Condition: Only depression, Ethnicity: white, Working status: In work or education, IMD: 2nd quintile deprived, Smoking status: Never smoked, Sexuality: Heterosexual, Activities limited today: No. Mean EQ-5D-3L utility Dolan: 0.914 (sd: 0.122); Mean EQ-5D-5L utility van Hout et al: 0.881 (sd: 0.124); Mean EQ-5D-5L utility Devlin et al: 0.929 (sd: 0.087)

Box 2: Sankey diagrams depicting response shifts and utility distributions for individuals "Retired with diabetes[a]"



Mobility

Self-care

Usual activities

Pain and Discomfort

Anxiety and Depression

Utility scores

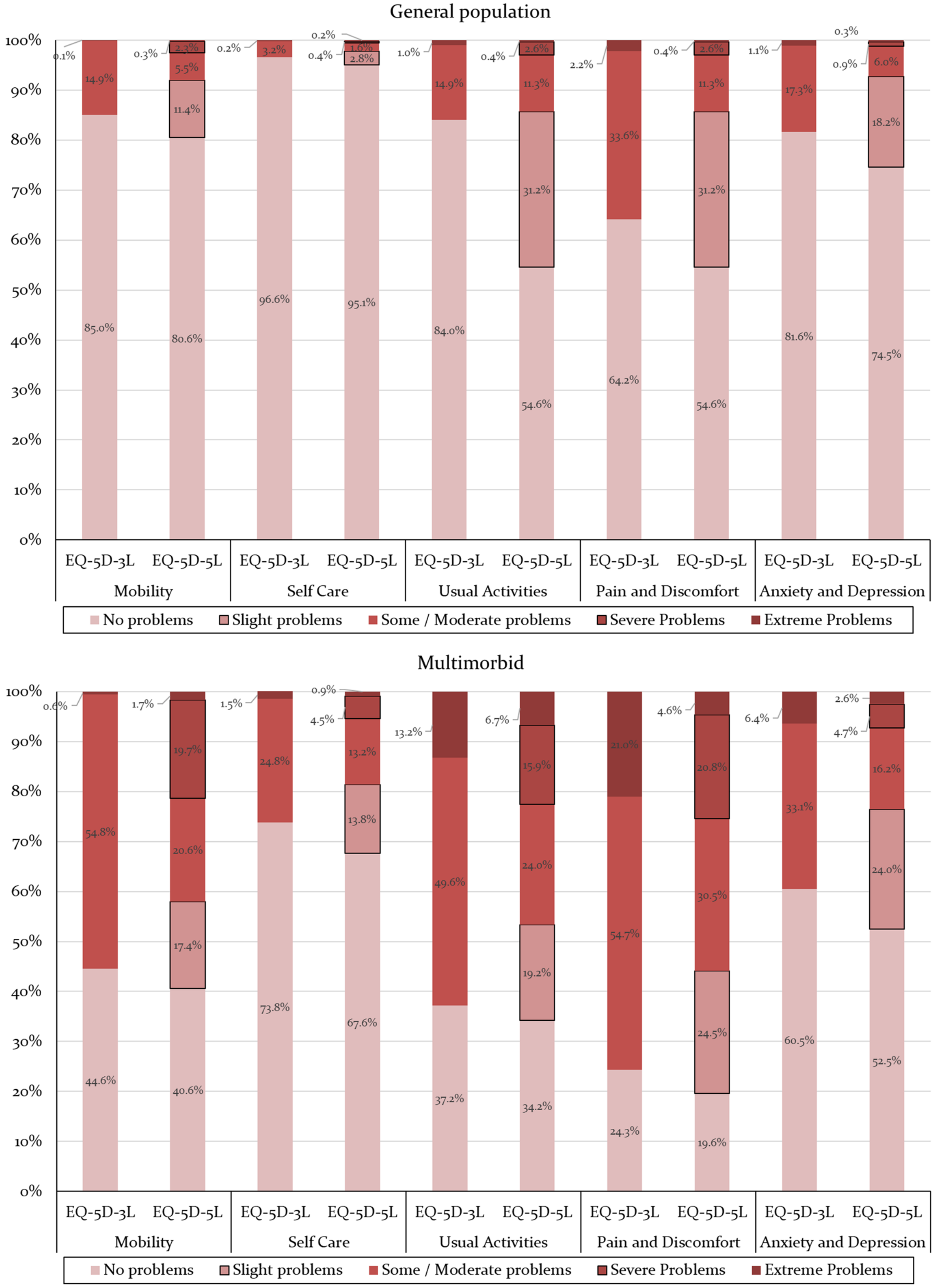
[a] A 'Sankey' diagram for a stratum where individuals completing the EQ-5D-3L and EQ-5D-5L in separate years have identical matched characteristics. Flows are calculated by a cross-tabulation of levels and domains between the 3L and 5L.Characteristics of stratum: cohort n=174. Age: 65-74, Sex: Male, Condition: Only diabetes, Ethnicity: white, Working status: Retired, IMD: 2nd quintile deprived, Smoking status: Former smoker, Sexuality: Heterosexual, Activities limited today: No. Mean EQ-5D-3L utility Dolan: 0.830 (sd: 0.165); Mean EQ-5D-5L utility van Hout et al: 0.810 (sd: 0.102); Mean EQ-5D-5L utility Devlin et al: 0.864 (sd: 0.085)

3.5 Impact on Utility Indices

Mean utility values for individuals completing the EQ-5D-3L, EQ-5D-5L van Hout et al. and EQ-5D-5L Devlin et al. in the general population were 0.867 (standard deviation [SD] 0.194), 0.856 (SD 0.173) and 0.905 (SD 0.142), respectively. Mean utility values for individuals completing the EQ-5D-3L, EQ-5D-5L van Hout et al. and EQ-5D-5L Devlin et al. in a multimorbid population were 0.710 (SD 0.269), 0.713 (SD 0.219) and 0.789 (SD 0.206) respectively. Consistent with the response changes, the EQ-5D-3L Dolan utility distributions (using EQ-5D-3L respondent data) and EQ-5D-5L van Hout et al. utility distribution (using EQ-5D-5L respondent data) mirrored each other but with a noticeable upwards shift for EQ-5D-5L van Hout et al. utilities (Fig. 5). There was poor overlap between the EQ-5D-3L Dolan and the EQ-5D-5L Devlin et al. (measured using EQ-5D-5L respondents) utility distributions. As populations became increasingly multimorbid, the mean utility measured by EQ-5D-5L van Hout et al. was higher than the equivalent for the EQ-5D-3L Dolan (Fig. 6). The Devlin et al. EQ-5D-5L generated significantly higher mean utility values than both the EQ-5D-3L Dolan and EQ-5D-5L van Hout et al. for all levels of morbidity.

4 Discussion

This is the first study to compare the measurement properties, response patterns and impact on population utility values of using the EQ-5D-3L and EQ-5D-5L within a large general population survey. Overall, we found the EQ-5D-5L to have better measurement performance than the EQ-5D-3L. This improvement occurs both for those close to full health (i.e. those in the general population) as well as those in poorer health (those with multimorbidity). Like some previous studies, we found the use of the EQ-5D-5L leads to ill health being more commonly reported compared



General population

Multimorbid

Fig. 1 Proportion of individuals selecting levels within each domain of the EQ-5D-3L and EQ-5D-5L

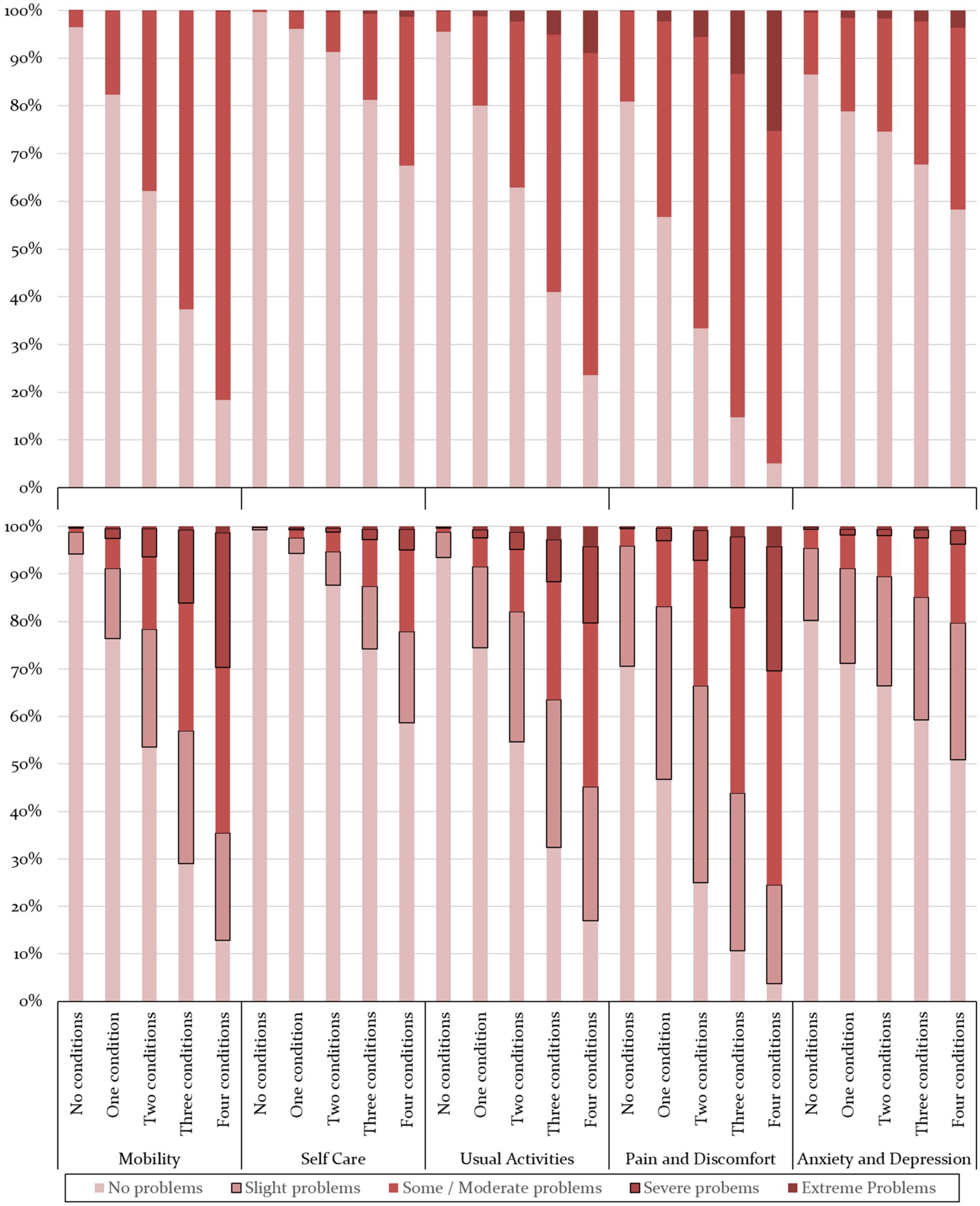


Fig. 2 Proportion of individuals selecting levels for the EQ-5D-3L (top) and EQ-5D-5L (bottom) as multimorbidity increases

with the EQ-5D-3L, but that levels of ill health reported are also less severe. As a consequence, we found the EQ-5D-5L generates higher mean utility values, at higher levels of multimorbidity, irrespective of the utility tariff used.

A key problem for the EQ-5D-3L is its tendency to generate significant ‘ceiling effects’ whereby a high proportion of respondents report having ‘no problems’ on all or one domain [3]. In line with previous studies using general population samples, we found a 10.6 percentage point fall in full health (no problems on all domains) reported for the EQ-5D-5L versus the EQ-5D-3L [5, 20, 21]. A similar phenomenon for the EQ-5D-3L, although not noted in the current

Table 3 Mean inconsistency between the EQ-5D-3L and EQ-5D-5L

	Mobility	Self-care	Usual activities	Pain and discomfort	Anxiety and depression
General population					
Domain inconsistency	10.8% (SD 0.00)	3.8% (SD 0.00)	12.0% (SD 0.00)	18.1% (SD 0.00)	17.3% (SD 0.00)
Total profile inconsistency	4.4% (SD 0.00)				
Multimorbid					
Domain inconsistency	23.9% (SD 0.00)	12.8% (SD 0.00)	26.2% (SD 0.00)	20.8% (SD 0.00)	23.1% (SD 0.00)
Total profile inconsistency	9.8% (SD 0.00)				

SD standard deviation across random pairings (n = 10) within strata

literature, is the tendency for respondents to report ‘some problems’ in any or all health domains, despite also reporting having no ongoing health conditions. This mismatch could be a consequence of the number of levels within the EQ-5D-3L forcing, for those close to full health with minor ailments, a choice between some problems or no problems [22]. In our sample of EQ-5D-5L respondents reporting no conditions, we observed a sizeable fall in those selecting ‘some problems’ versus the EQ-5D-3L, with a large proportion instead selecting ‘slight problems’. Taken together, these findings suggest that the richer descriptive system of the EQ-5D-5L has improved sensitivity to capture deviations from full health, where they exist, but has also improved specificity to appropriately characterise small reductions in health close to full health. Consequently, the EQ-5D-5L should be preferred in surveys of individuals who are generally healthy, such as population surveys.

We also examined the informational characteristics of the competing descriptive systems by using measures frequently reported in the literature, such as Shannon’s absolute and relative measures of informativity. Our findings were similar to the summaries reported by Buchholz et al. [3], with zero or modest improvements in relative informativity in the general population but gains in absolute informativity across all domains. However, for a multimorbid population, there was an improvement in both absolute and relative measures of informativity across most domains. Moreover, for a multimorbid population, it was clear that a high proportion of individuals were making use of the new ‘severe problems’ level, reducing in turn the proportion selecting ‘extreme problems’. Like Agborsangaya et al. [4], we found that the EQ-5D-5L had a much richer set of responses being more frequently used than the EQ-5D-3L, particularly in individuals reporting multimorbidity. These findings suggest that the EQ-5D-5L should also be preferred in populations who have multimorbidity, which is the norm for those suffering from chronic disease [23].

We explored whether there was any general upwards or downwards trend in the way that respondents completed the two instruments. For the usual activities and pain and discomfort domains, we found those completing the EQ-5D-5L were more likely to select levels indicating better health than those completing the EQ-5D-3L. Overall, when combining all the domains, we found the EQ-5D-5L was associated with respondents reporting better health. This was the case despite matching ensuring EQ-5D-5L and EQ-5D-3L respondents were identical on all observed characteristics.

Finally, we explored the impact of using the two instruments on utility values by scoring EQ-5D-3L responses using the EQ-5D-3L Dolan tariff and EQ-5D-5L responses using both the EQ-5D-5L van Hout et al. mapping algorithm and the EQ-5D-5L Devlin et al. tariff. The National Institute for Health and Care Excellence (NICE) have recently reaffirmed their position statement to use the EQ-5D-3L tariff, with EQ-5D-5L responses to be mapped to the original EQ-5D-3L utility space through the use of the van Hout et al. mapping algorithm [24]. Overall, we found that the utility distribution for EQ-5D-3L respondents closely matched the EQ-5D-5L van Hout et al. mapped distribution, but that there was a small upwards shift towards full health in the utility scale for the EQ-5D-5L van Hout et al. algorithm, particularly in those who were highly multimorbid.

Table 4 Shannon–Weaver index (H') and Shannon’s Evenness Index (J') for the EQ-5D-3L and EQ-5D-5L in the general population and in individuals who were multimorbid (condition count > 1)

	H		*J*	
	3L	5L	3L	5L
General population				
Mobility	0.62	0.99	0.39	0.42
Self-care	0.23	0.36	0.14	0.15
Usual activities	0.69	1.00	0.44	0.43
Pain and discomfort	1.06	1.52	0.67	0.66
Anxiety and depression	0.75	1.10	0.47	0.47
Multimorbid				
Mobility	1.04	2.00	0.66	0.86
Self-care	0.91	1.42	0.58	0.61
Usual activities	1.42	2.16	0.89	0.93
Pain and discomfort	1.44	2.16	0.91	0.93
Anxiety and depression	1.75	1.75	0.77	0.75

We suggest that the upwards shift in the utility scale, when using the mapping algorithm, was a function of an upwards response shift with individuals reporting better health for the 5L rather than due to a change in the valuation system. By contrast, the Devlin et al. distribution was unlike the EQ-5D-3L distribution or mapped EQ-5D-5L distribution, displaying a large shift upwards in utility values combined with a narrowing of the overall distribution. We suggest that the Devlin et al. distribution is a function of both a response shift and a change in the valuation system, and therefore is similar to the findings reported by Hernandez Alava et al. [26] when exploring the impact of switching from the EQ-5D-3L to the EQ-5D-5L Devlin et al. tariff on cost-effectiveness results. Overall, the distributional findings in this study validate those previously reported by Mulhern et al. [25] on a smaller sample for England and those reported by Janssen et al. [22] for EQ-5D-3L and EQ-5D-5L values across seven countries.

Given the recent NICE position statement recommending the use of the van Hout et al. mapping algorithm for studies that have collected 5L data, we suggest the implications of our results for cost-effectiveness analysis conducted in the near future could be similar (but smaller in magnitude) to those outlined by Hernandez Alava et al. [26]. Due to the upwards shift in utility values that we observed for the 5L respondents using the van Hout et al. algorithm, primarily driven by a response shift, there is less utility space for potential gains to occur versus identical EQ-5D-3L respondents scored using the Dolan EQ-5D-3L tariff. Consequently, despite the superior measurement properties of the EQ-5D-5L, our results suggest that, in general, interventions that promote health-related quality of life in individuals who are highly multimorbid are likely to be less cost effective, having a higher incremental cost-effectiveness ratio (ICER), versus similar interventions assessed using the EQ-5D-3L. Alternatively, interventions that improve life expectancy in individuals with multimorbidity are more likely to be cost effective and have a lower ICER, using the EQ-5D-5L rather than the EQ-5D-3L. Further research could focus on identifying the conditions or groups of conditions that lead to the largest response shift between the EQ-5D versions as it is these conditions where cost effectiveness is likely to be most influenced by the instrument used.

A potential limitation of this study is that we followed previous studies [4, 5] in comparing EQ-5D-3L and EQ-5D-5L responses using different individuals. Such indirect comparisons can induce bias if EQ-5D-3L and EQ-5D-5L respondents are dissimilar, as differences in measurement properties could be driven by differences in health (or characteristics that affect the reporting of health) rather than any causal effects of the version change [27, 28]. Previous indirect studies comparing the EQ-5D-3L and EQ-5D-5L had significant differences in sample characteristics. In our study, we mitigated the risk of bias by using matching to ensure respondents are close to identical on observed characteristics. Despite this, we cannot rule out bias from unobserved differences in characteristics across EQ-5D-3L and EQ-5D-5L respondents. However, given

Fig. 4 Proportion of individuals selecting ‘no problems’ for the EQ-5D-3L and EQ-5D-5L in the general population (top) and in those reporting no chronic health conditions (bottom) stratified by response to the question ‘‘Have your activities been limited today because you have recently become unwell or been injured?”

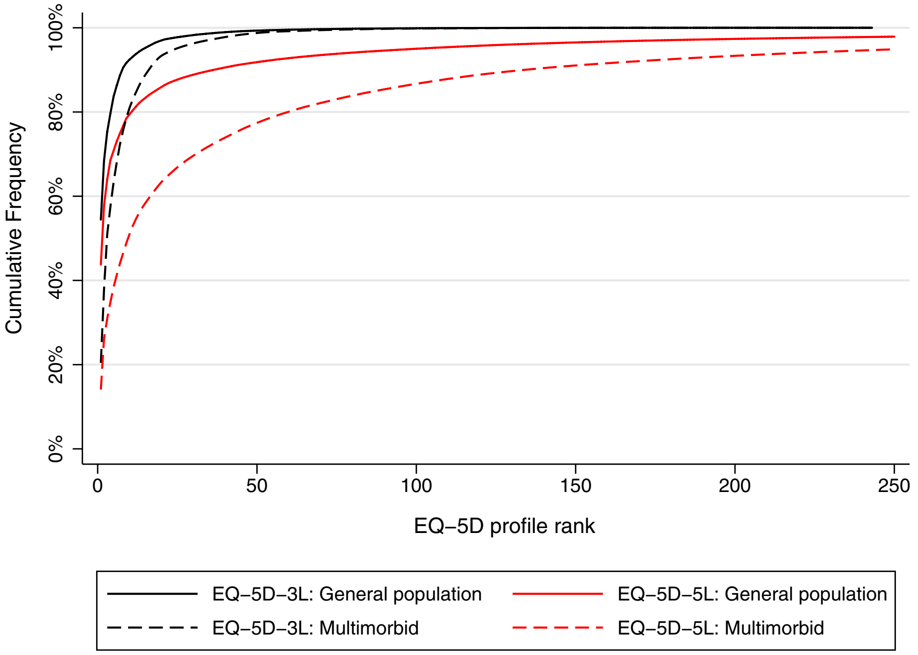


Fig. 3 Proportion of EQ-5D profile states completed by individuals[1]. [1]The EQ-5D profile rank refers to the particular combination of levels and domains available across the descriptive systems (‘profiles’) ordered by those profiles that are most common

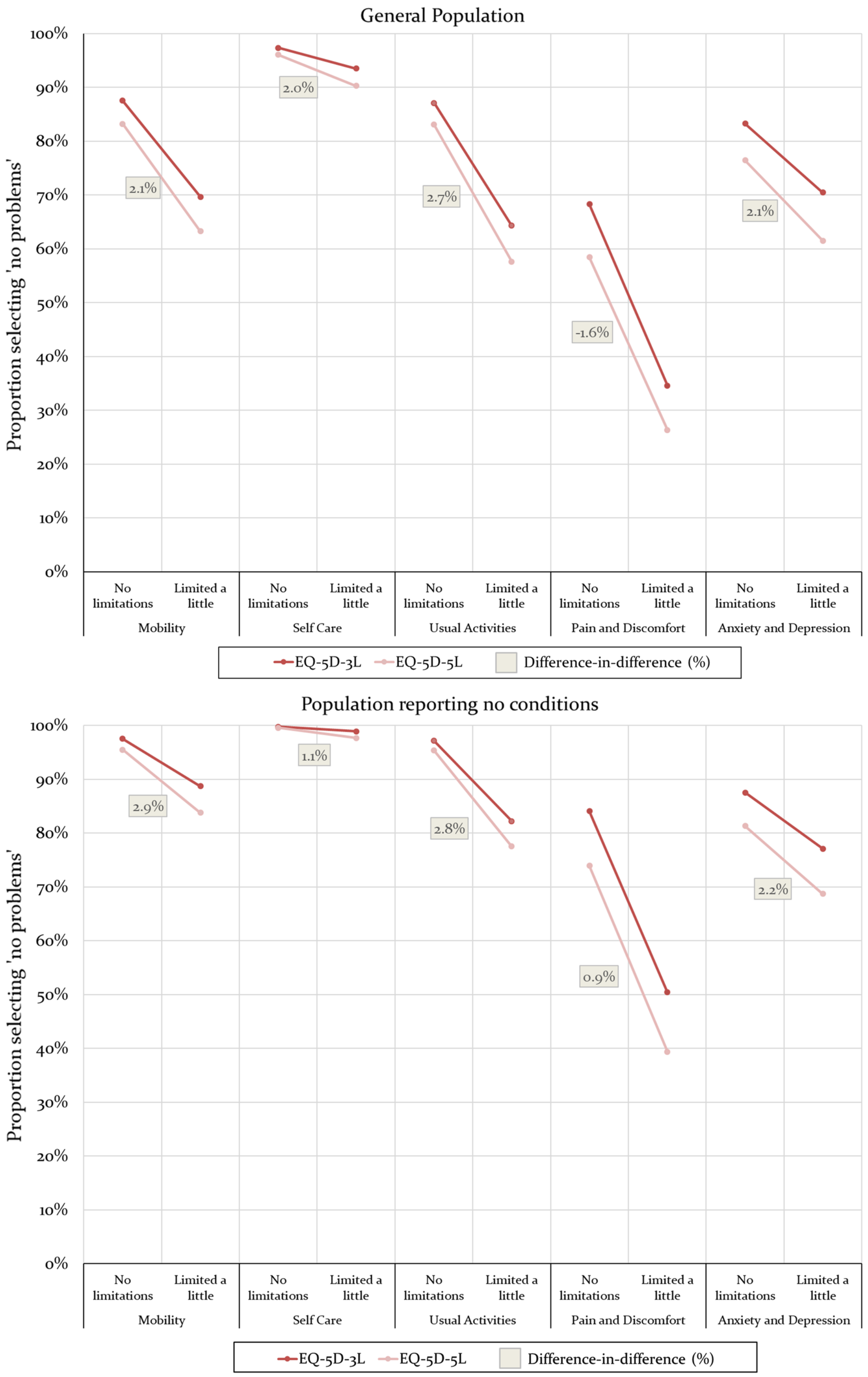
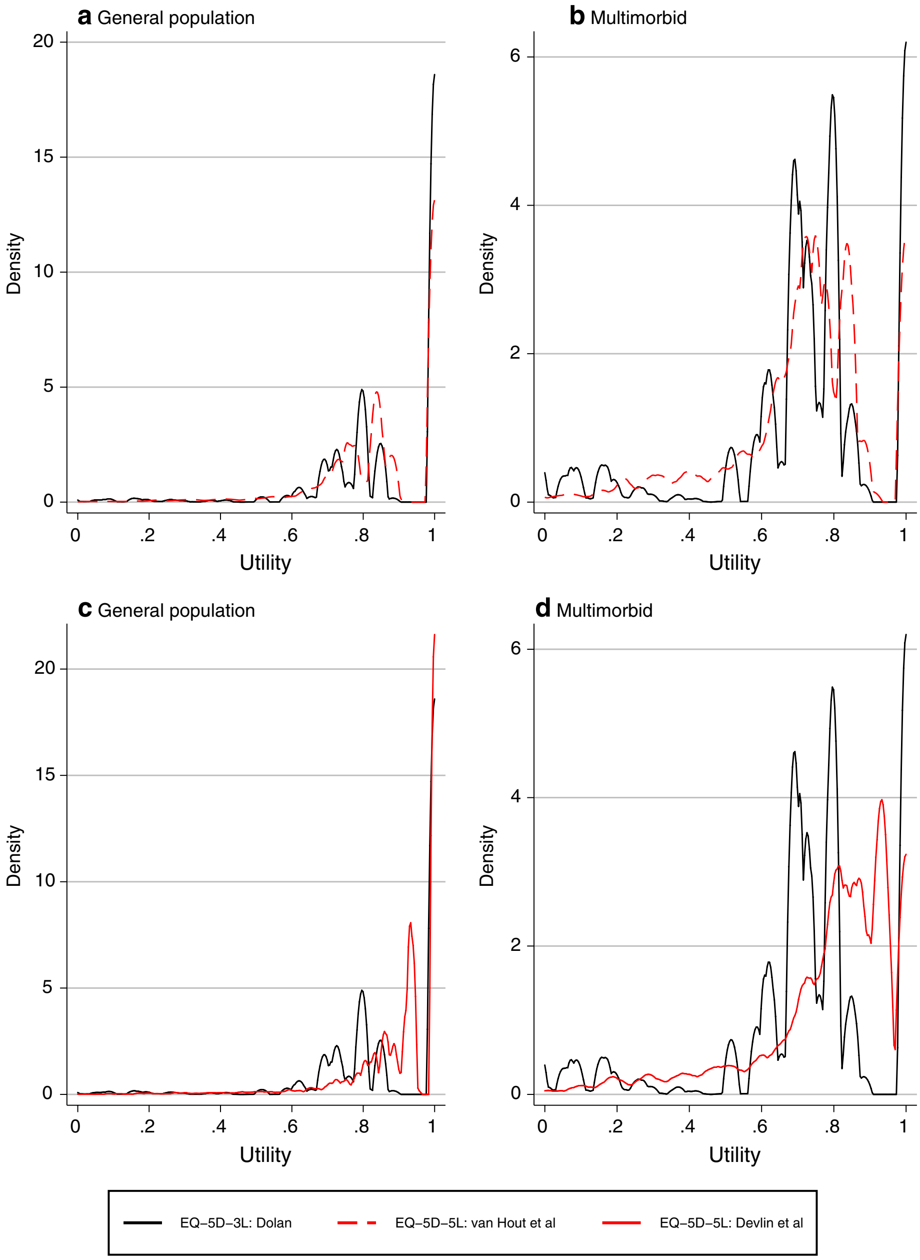

General Population
Proportion selecting 'no problems'
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2.1%
2.0%
2.7%
-1.6%
2.1%
No limitations
Limited a little
Mobility
Self Care
Usual Activities
Pain and Discomfort
Anxiety and Depression
EQ-5D-3L
EQ-5D-5L
Difference-in-difference (%)
Population reporting no conditions
Proportion selecting 'no problems'
2.9%
1.1%
2.8%
0.9%
2.2%
No limitations
Limited a little
Mobility
Self Care
Usual Activities
Pain and Discomfort
Anxiety and Depression
EQ-5D-3L
EQ-5D-5L
Difference-in-difference (%)

Table 5 Change in mean response between the EQ-5D-3L and EQ-5D-5L

	Mobility	Self-care	Usual activities	Pain and discomfort	Anxiety and depression
General population					
Response level change, mean	0.00 (SD 0.00)	0.01 (SD 0.00)	– 0.04 (SD 0.00)	– 0.13 (SD 0.00)	– 0.05 (SD 0.00)
Misery index change, mean	– 0.21 (SD 0.00)				
Multimorbid					
Response level change, mean	0.01 (SD 0.00)	0.01 (SD 0.00)	– 0.09 (SD 0.00)	– 0.45 (SD 0.00)	0.00
Misery index change, mean	– 0.45 (SD 0.00)				

SD standard deviation across random pairings (n = 10) within strata

Fig. 5 Utility distributions for the EQ-5D-3L and EQ-5D-5L. **a** General population using EQ-5D-3L Dolan [17] and EQ-5D-5L van Hout et al. [18]. **c** General population using EQ-5D-3L Dolan [17] and EQ-5D-5L Devlin et al. [19]. **b** Multimorbid population using EQ-5D-3L Dolan [17] and EQ-5D-5L van Hout et al. [18]. **d** Multimorbid population using EQ-5D-3L Dolan [17] and EQ-5D-5L Devlin et al. [19]



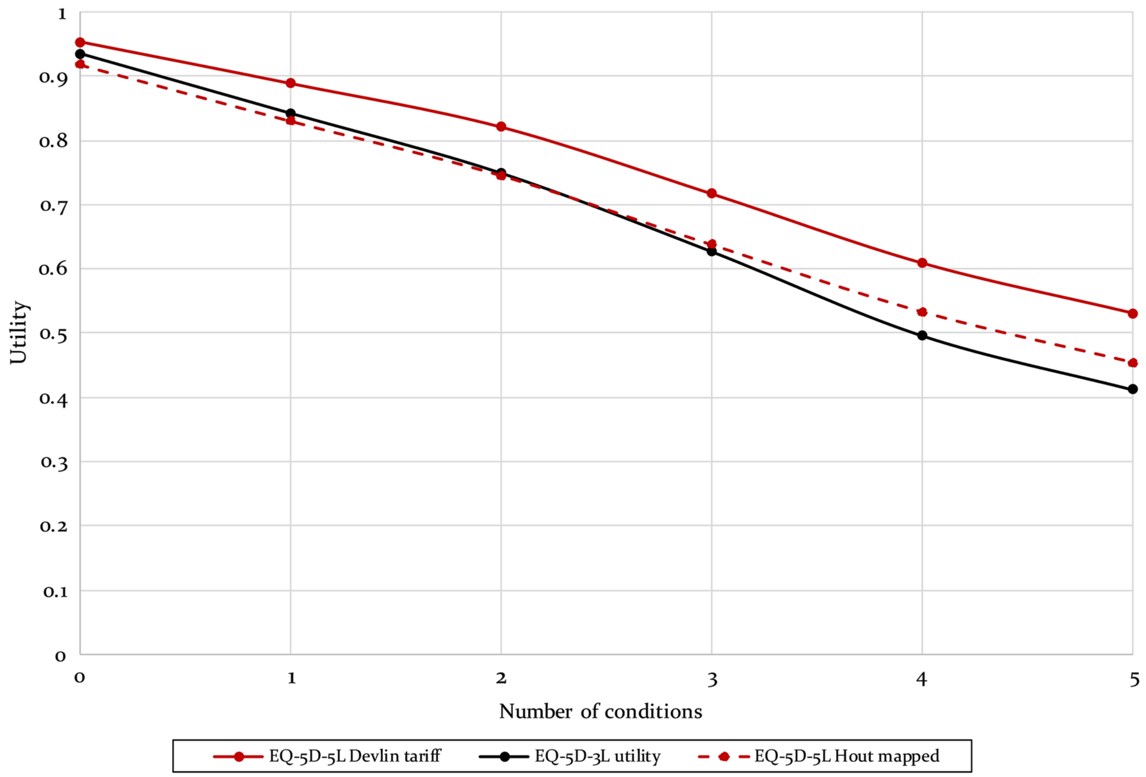


Fig. 6 Population mean utility values stratified by condition count

only small differences in observed characteristics prior to matching, and that results were unchanged when EQ-5D-3L respondents were matched to different observably identical EQ-5D-5L respondents (with potentially different unobserved characteristics), we believe this risk is minimal.

Indirect comparisons may also limit some concerns with H2H comparisons. Many H2H studies ask respondents to complete the EQ-5D-3L and EQ-5D-5L in the same survey [3]. Respondents may therefore anchor responses to the second version presented to those chosen for the first version, underestimating true variation in responses across versions. Although some studies attempt to mitigate this by asking respondents to complete the different versions 2 weeks apart [29], indirect comparison rules out the possibility of anchoring effects. The majority of H2H studies also present the EQ-5D-5L before the EQ-5D-3L version, which could induce an ordering effect [18], an effect discovered for other self-reported health measures [30]. Again, such ordering effects are ruled out by indirect comparisons.

Finally, previous studies have used levels of missing data in EQ-5D responses to compare the feasibility of the 5L and the 3L [31]. This was not conducted here due to evidence of improved response rates for all variables between the 2011 and 2012 GPPS survey years. This would have led to reduced missing data on the EQ-5D irrespective of the version change. Given the similarity in health and individual characteristics across survey years, we believe differences in response rates across years are unlikely to have biased other comparisons of the 5L and 3L made in this study.

5 Conclusions

Using matched responses from over 1 million individuals, this study found the EQ-5D-5L to have superior measurement properties to the EQ-5D-3L, including improvements in discriminatory power and reduced ceiling effects. Consequently, the EQ-5D-5L should be preferred for measuring health status in populations close to full health (i.e. general population) as well as those in poorer health (i.e. those with multimorbidity). However, the EQ-5D-5L also results in a re-distribution of responses such that individuals tend to select responses indicative of better health. This leads to increases in utility values for the EQ-5D-5L versus identical individuals using the EQ-5D-3L at higher levels of multimorbidity, irrespective of the choice of utility tariff or mapping function.

Author Contributions Alex Thompson came up with the original idea for the paper. Both authors designed the analysis. Alex Turner wrote the code to match the dataset across cohorts, and Alex Thompson analysed

the data. Both authors interpreted the results and wrote the manuscript together.

Data Availability Statement The data that support the findings of this study are available from NHS England but restrictions apply to the availability of these data, which were used under license for the current study and therefore are not publicly available.

Compliance with Ethical Standards

Funding No funding was received for this project.

Conflict of interest Alex Thompson and Alex Turner have no conflicts of interest to declare.

References

1. EuroQoL Group. EuroQol—a new facility for the measurement of health-related quality of life. Health Policy. 1990;16:199–208.
2. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011;20:1727–36.
3. Buchholz I, Janssen MF, Kohlmann T, Feng Y-S. A systematic review of studies comparing the measurement properties of the three-level and five-level versions of the EQ-5D. Pharmacoeconomics. 2018;36(6):645–61.
4. Agborsangaya CB, Lahtinen M, Cooke T, Johnson JA. Comparing the EQ-5D 3L and 5L: measurement properties and association with chronic conditions and multimorbidity in the general population. Health Qual Life Outcomes. 2014;12:74.
5. Feng Y, Devlin N, Herdman M. Assessing the health of the general population in England: how do the three- and five-level versions of EQ-5D compare? Health Qual Life Outcomes. 2015;13:171.
6. NHS England. GP Patient Survey 2020 [cited 17 Jan 2020]. https://www.england.nhs.uk/statistics/statistical-work-areas/gp-patient-survey/.
7. Ipsos MORI. Technical Annex for the GP Patient Survey: 2011–2012 Annual Report 2012 [cited 17 Jan 2020]. https://www.gp-patient.co.uk/downloads/archive/2012/June/June/2012/Technical/Annex.pdf.
8. NHS Digital. Patients registered at a GP practice: April 2018; Special Topic—number of patients registered compared to the projected resident population in England. 2018 [6 Feb cited 2020]. https://files.digital.nhs.uk/A7/EF50EA/gp-reg-pat-prac-topic-int.pdf.
9. Iacus SM, King G, Porro G. Causal inference without balance checking: coarsened exact matching. Polit Anal. 2011;20:1–24.
10. Blackwell M, Iacus SM, King G, Porro G. CEM: Coarsened exact matching in Stata. Stata J. 2009;9:524–46.
11. Iacus SM, King G, Porro G. Multivariate matching methods that are monotonic imbalance bounding. J Am Stat Assoc. 2011;104:345–61.
12. Department of Communities and Local Government. The English Indices of Deprivation 2010. London; 2011.
13. The Academy of Medical Sciences. Multimorbidity: a priority for global health research. 2018. https://acmedsci.ac.uk/file-download/82222577.
14. Janssen MF, Birnie E, Haagsma JA, Bonsel GJ. Comparing the standard EQ-5D three-level system with a five-level version. Value Health. 2008;11:275–84.
15. Schmidt M. The Sankey diagram in energy and material flow management. J Ind Ecol. 2008;12:82–94.
16. Lupton RC, Allwood JM. Hybrid Sankey diagrams: visual analysis of multidimensional data for understanding resource use. Resour Conserv Recycl. 2017;124:141–51.
17. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35:1095–108.
18. van Hout B, Janssen M, Feng Y-S, Kohlmann T, Busschbach J, Golicki D, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health. 2012;15(5):708–15. https://www.valueinhealthjournal.com/article/S1098-3015(12)00058-7/pdf.
19. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: an EQ-5D-5L value set for England. Health Econ. 2018;27:7–22.
20. Shiroiwa T, Fukuda T, Ikeda S, Igarashi A, Noto S, Saito S, et al. Japanese population norms for preference-based measures: EQ-5D-3L, EQ-5D-5L, and SF-6D. Qual Life Res. 2016;25(3):707–19.
21. Craig BM, Pickard AS, Lubetkin EI. Health problems are more common, but less severe when measured using newer EQ-5D versions. J Clin Epidemiol. 2014;67:93–9.
22. Janssen MF, Gouke JB, Bonsel J, 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(6):675–97.
23. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet. 2012;380:37–43.
24. National Institute for Health and Care Excellence. Position statement on use of the EQ-5D-5L value set for England. 2019. https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/technology-appraisal-guidance/eq-5d-5l.
25. Mulhern B, Feng Y, Shah K, Janssen MF, Herdman M, van Hout B, et al. Comparing the UK EQ-5D-3L and English EQ-5D-5L value sets. Pharmacoeconomics. 2018;36:699–713.
26. Hernandez Alava M, Wailoo A, Grimm S, Pudney S, Gomes M, Sadique Z, et al. EQ-5D-5L versus EQ-5D-3L: the impact on cost effectiveness in the United Kingdom. Value Health. 2018;21:49–56.
27. Bago T, O'Donnell O, Van doorslaer E. Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans. Int J Epidemiol. 2008;37:1375–83.
28. Lindeboom M, Van Doorslaer E. Cut-point shift and index shift in self-reported health. J Health Econ. 2004;23:1083–99.
29. Khan I, Morris S, Pashayan N, Matata B, Bashir Z, Maguirre J. Comparing the mapping between EQ-5D-5L, EQ-5D-3L and the EORTC-QLQ-C30 in non-small cell lung cancer patients. Health Qual Life Outcomes. 2016;14:60.

30. Garbarski D, Schaeffer NC, Dykema J. The effects of response option order and question order on self-rated health. Qual Life Res. 2015;24:1443–53.

31. Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. Qual Life Res. 2013;22:1717–27.