Article:
Marten, O., Mulhern, B., Bansback, N. et al. (1 more author) (2020) Implausible states: prevalence of EQ-5D-5L states in the general population and its effect on health state valuation. Medical Decision Making, 40 (6). pp. 735-745. ISSN 0272-989X

https://doi.org/10.1177/0272989X20940673

---

revised version for MDM May 2020

Implausible states: prevalence of EQ-5D-5L states in the general population and its effect on health state valuation

Running head: Implausible EQ-5D-5L health states

Author Information:

Ole Marten, MSc (School of Public Health, Bielefeld University, Germany),
contact: Universitaetsstrasse 25, Bielefeld, Germany, 33615; e-mail: <ole.marten@uni-bielefeld.de>

Brendan Mulhern, MRes (Centre for Health Economics Research and Evaluation, University of Technology Sydney, Australia),
contact: 1–59 Quay St, Haymarket, Sydney, Australia, NSW 2000; e-mail: <Brendan.Mulhern@uts.edu.au>

Nick Bansback, PhD (School of Population and Public Health, University of British Columbia, Canada),
contact: 2206 East Mall, Vancouver, BC Canada, V6T 1Z3; e-mail: <nick.bansback@ubc.ca>

Aki Tsuchiya, PhD (School of Health and Related Research, University of Sheffield, UK),
contact: Regent Court, 30 Regent Street, Sheffield, UK, S1 4DA; e-mail: <a.tsuchiya@sheffield.ac.uk>

Address correspondence to Ole Marten, School of Public Health, Bielefeld University, Universitaetsstrasse 25, Bielefeld, Germany; e-mail: ole.marten@uni-bielefeld.de

Very early versions of this paper using different datasets were presented at the UK Health Economists' Study Group (HESG) meeting in Aberdeen (June 2017), and as a poster at the EuroQol Group meeting in Barcelona (September 2017), but there are no working paper versions of these. And, the current paper is not in the public domain.

[Word-count 4978]

This work has not received any external funding. Our time has been supported by our respective institutions. The research team has independently designed the study, interpreted the data, and wrote the manuscript. All analyses are based on secondary data and do not require research ethics approval. Individual level data from the General Practitioner Patient Survey were accessed with permission by NHS England. Nick Bansback, Brendan Mulhern, and Aki Tsuchiya are members of the EuroQol Group.

Abstract

The EQ-5D is made up of health state dimensions and levels, where some combinations seem less "plausible" than others. If "implausible" states are used in health state valuation exercises, then respondents may have difficulty imagining them, causing measurement error. There is currently no standard solution: some valuation studies exclude such states, whilst others leave them in. This study aims to address two gaps in the literature: (1) to propose an evidence-based set of the least prevalent two-way combinations of EQ-5D-5L dimension-levels; and (2) to quantify the impact of removing perceived implausible states from valuation designs. For the first aim, we use data from two waves of the English General Practitioner Patient Survey (n=1,639,453). For the second aim, we re-model a secondary dataset of a Discrete Choice Experiment (DCE) with duration that valued EQ-5D-5L and compare across models that drop observations involving different health states: (i) implausible states as defined in the literature; (ii) the least prevalent states identified in stage (1); and (iii) randomly select states; alongside (iv) a model that does not drop any observations. The results indicate that two-way combinations previously thought to be implausible actually exist amongst the general population; there are other combinations that are rarer; and that removing implausible states from an experimental design of a DCE with duration leads to value sets with potentially different characteristics depending on the criterion of implausible states. We advise against the routine removal of implausible states from health state valuation studies.

[243wds]

1. Introduction

Preference-based generic health instruments are used to operationalise the Quality Adjusted Life Year (QALY) for the economic evaluation of health care interventions. Typically, they take the form of a *descriptive* or *classification system*, made up of dimensions of health with differing levels of severity that describe different health states; and an accompanying *value set*, or *tariff*, which specifies the preference weights for each of the health states that the classification system describes on an interval scale with 1 for full health and 0 for being dead. Examples of such instruments include the EQ-5D-3L[1] and EQ-5D-5L,[2] HUI3,[3] SF-6D,[4] and AQoL.[5] For example, EQ-5D-3L has five dimensions: mobility (MO), self-care (SC), usual activities (UA), pain or discomfort (PD), and anxiety or depression (AD); with three severity levels across each (1 for no problems and 3 for the most severe), thus distinguishing 243 different health states.[6] EQ-5D-5L is a later variant with five levels instead of three, distinguishing 3,125 different health states.[7]

The value sets are typically estimated by health state valuation studies that survey members of the public for their preferences for stylised health state descriptions (often referred to as "hypothetical health states") selected from the descriptive system. These valuation data are then modelled econometrically, to predict average preferences for all health states described by the classification system. One potential consideration in the selection of the stylised health states is their plausibility. The idea of "implausible states" can be broken down into two. One is where some health states appear unlikely and more difficult for a typical general public respondent to imagine (difficult-to-imagine states). The other is where some states are, as a matter of fact, much less prevalent than others (rare states). The extreme case of the latter are states that will never be observed (impossible states), but since self-reporting of states involves error, it is futile to distinguish empirically between rare states and impossible states. A given health state may be both difficult-to-imagine and rare. However, for the purpose of selecting the stylised health states for valuation, what matters are the difficult-to-imagine states, as perceived by survey respondents.

The Measurement and Valuation of Health (MVH) project is one of the earliest examples of a health state valuation study, and used the time trade-off (TTO) method to value EQ-5D-3L states.[8] When selecting the 42 EQ-5D states that were valued, the researchers "wanted to exclude states which seemed prima facie implausible to respondents, so as to sustain motivation and credibility".[8] Thus, they excluded states that combine no problems in one dimension and an extreme problem in another (hereafter, "two-way contrasting combinations"), specifically, between no problems in usual activities and: extreme problems in mobility, described as being 'confined to bed' (UA1xMO3); or extreme problems in self-care (UA1xSC3).[1] However, there was no empirical basis to select these combinations.

The methodological literature on health state valuation since then has largely left this topic untouched, and studies have either excluded (or "backed off" from) health states considered to be implausible albeit without empirical support (e.g. Viney et al.[9]), or have included all possible combinations in the design development process (e.g. Mulhern et al.[10]). Where health states are selected through any experimental design procedure, simply dropping certain states may result in inefficiency (error) and/or inaccuracy (bias). However, if those states are difficult for respondents to imagine, then keeping them in the design may also result in measurement error introducing error and/or bias. The exception is the Australian EQ-5D-3L valuation study using a Discrete Choice Experiment (DCE), which included a simulation exercise where the effect of removing certain states was examined.[11] A design with all possible EQ-5D-3L states was compared to a design that dropped 45 states judged to be implausible - this time defined as two-way contrasting combinations of the worst severity level in mobility with:

either no problems in usual activity (MO3xUA1), or with no problems in self-care (MO3xSC1)[9]. The findings suggested the variability in data was greater in the all-states design, but the statistical efficiency was lower in the design where these perceived implausible states were dropped. However, as the authors state, the exercise only examined the efficiency of designs, and not the measurement error attributable to the difficulty of imagining these states. In a separate study, Jakubczyk and Golicki[12] applied an experimental TTO design to examine the amount of imprecision around health preferences from a Polish student sample. The authors found that values of states with two-way contrasting combinations have higher imprecision and that imprecision is mostly driven by UA and AD.

The aims of this study were to first analyse the prevalence of self-reported EQ-5D-5L states in a large sample of the general population, in order to propose an evidence-based set of the least prevalent EQ-5D-5L states. Second, to re-model a rich secondary DCE-with-duration health state valuation dataset to quantify the impact of excluding choice sets that involve health states that we find the least prevalent. This is compared against models that remove health states considered implausible by existing literature and a model that does not drop any observations. In order to distinguish between the effect of removing specific states and the effect of removing any states, we also have a series of models that remove choice sets at random.

2. Methods

2.1 Stage 1: Prevalence of EQ-5D-5L states in the general population

For the first stage, self-reported EQ-5D-5L data from the General Practitioner (GP) Patient Survey were used. This is a large cross-sectional survey conducted by NHS England, "designed to give patients the opportunity to feed back about their experiences of their GP practice".[13] The sample consists of individuals aged 18 or above with a valid National Health Service (NHS) number, who have been registered with a GP practice in England for 6 months or more. The sample, in effect recruited from the general population, are asked a wide range of questions including self-reported EQ-5D-5L. We used data from two years: 2013 (the first time EQ-5D-5L was included in the survey), and 2015 (the last year with data collection in the summer and winter).*

The analysis was conducted at the level of EQ-5D-5L states. First, the states were ranked from the most frequently observed to those with no observations. Second, the ranks at which different levels of a given dimension first appeared were analysed. Third, the frequencies of all two-way contrasting and non-contrasting combinations were cross-tabulated. Thus, we identify two kinds of two-way combinations that characterise prevalence-based implausibility: those that are the least frequently reported; and those that are the most common amongst states that are never observed. These need not be contrasting combinations.

2.2 Stage 2: Gauging the effect of different exclusion strategies in DCE_{TTO}

In Stage 2, in order to explore the effect of removing rare state from a design, we re-modelled EQ-5D-5L valuation data that used DCE-with-duration (DCE_{TTO}), by systematically excluding observations from

*EQ-5D-5L is coded as five different questions, and the publicly available version of the dataset does not allow building the EQ-5D-5L profile. Therefore, for this analysis, access to the individual data was applied for to NHS England.

choice tasks that included health states of varying plausibility. Secondary data from the 'Further Exploration of DCE with duration to value EQ-5D-5L' (FEDEV)[14] project were used, which had included substantially more choice tasks than were required to estimate main effects, and enabled us to drop some tasks without compromising the integrity of the experimental design.

The FEDEV study was conducted online with a representative sample of the UK general population. The valuation data in this study is a subset of the FEDEV dataset, based on DCE_{TTO} tasks that were designed using non-informative (zero) priors, and administered to 800 respondents. The design included 120 health profile pairs, where each profile comprises an EQ-5D-5L state and one of six duration levels ranging from 0.5 to 10 years. The pairs were selected using a swapping algorithm implemented in the DCE design software Ngene,[15] based on minimising the D-error. Each respondent completed 10 choice tasks from the underlying design.[14]

The analysis re-estimated the DCE_{TTO} models based on conditional logit regressions, which estimate utility decrements for each severity level of each dimension.[16] The coefficients on the latent scale were then anchored onto a scale with 1 for full health and 0 for being dead, producing unanchored and anchored sets of coefficients. This was repeated by excluding data generated from certain choice tasks.

The original FEDEV experimental design had 120 choice tasks, to enable exploration of interactions between duration and dimension severity level, and is substantially larger than the minimum number of choice tasks required for estimating the main effects required for this study, with 21 parameters. We were consequently confident that, while excluding some tasks would impact the precision of estimates (d-efficiency), it would not impact the statistical identification of unbiased parameter estimates, particularly using a design with zero (non-informative) priors where two-way combinations should appear a similar number of times. However, since there is a limit to the number or combination of tasks that could be removed, we checked the experimental design properties of each selection of tasks for important correlations and statistical properties using Ngene.[15]

Building on the original experimental design, we developed four *treatment* datasets dropping observations from the source data depending on the definition of implausible states to be excluded. It would not be appropriate to compare these treatments directly to the full model, since that would confound the effect of removing specific observations with the effect of removing any observations. Therefore, we created *comparator* datasets that excluded similar combinations of levels at random over multiple draws to reduce bias in comparisons. Table 1 summarises the models used in Stage 2, consisting of four different treatments and three comparators.

[PLEASE INSERT TABLE 1 HERE]

The first two treatment models (T1 and T2) were generated based on criteria of perceived implausibility proposed for EQ-5D-3L by Dolan[1] and Viney et al.[9,11] We adapt their two-way contrasting combinations to the EQ-5D-5L and re-define them as one dimension with severe or extreme problems (level 4 or 5) combined with another dimension with no problems (level 1). For comparability, the third and fourth treatment models (T3 and T4) were based on data that excluded observations arising from the same number of two-way combinations as the previous models (i.e. four combinations), but based on combinations that were *empirically identified* as the least prevalent in Stage 1. T3 removed the four least prevalent two-way combinations amongst self-reported states, while T4 removed the four most common two-way combinations amongst never-reported states.

Comparator model 1 (C1avg) removed observations from the full dataset based on four *randomly selected two-way combinations* from the set of all possible two-way combinations. Given the random nature of the exercise, any result will be subject to random error. Therefore, 100 models were estimated based on random draws, and their parameter values averaged. Since treatment models 1 and 2 removed observations involving *two-way contrasting* combinations, this may influence the experimental design in a systematically different way from comparator model 1. Comparator models 2 (C2avg) excluded observations from the full dataset if they concerned at least one of four two-way combinations chosen randomly from the set of 40 two-way contrasting combinations. This is repeated 100 times and the anchored parameter values are averaged. Comparator model 3 (C3) does not exclude any choice tasks, and produces a single comparator model using the full dataset.

A total of four treatment models and 201 comparator models were estimated. Predicted values for the three most frequently self-reported states with problems in at least two dimensions were estimated from anchored parameters to each model, along with values for more plausible states across the range of severity (22222, 33333, 44444, and 55555). A series of further descriptive analyses compare across the designs and the estimated parameters in order to distinguish between the effect of removing specific observations and that of removing random observations.

3. Results

3.1 Stage 1: Descriptive analysis

The GP Patient Survey dataset included 1,639,453 individuals, distributed across 2,707 unique EQ-5D-5L health states. This covers approximately 87% of all possible health states, leaving 418 health states that were not reported in our sample. About one third of the respondents reported being in the best health state (11111). The three most prevalent states (11111, 11121, 11112) cover just over 50%, and the 33 most prevalent states cover just over 80% of the observations (see Table 2). Among the 33 most prevalent states only one includes SC2, while 27 include PD 2 or worse. None of the 33 most prevalent states includes levels 4 or 5. About half of the observed states are reported by less than 0.1% of the sample (<1,640).

[PLEASE INSERT TABLE 2 HERE]

Table 3 tabulates the overall frequencies of the different levels by dimension, with varying patterns across the dimensions. Problems in SC are the rarest, while problems in PD are the most prevalent. Furthermore, the prevalence of levels 4 or 5 in MO, UA and PD (ranging from 5.7-7.1%) is relatively higher than the prevalence of levels 4 or 5 in SC (2%) and AD (3%).

[PLEASE INSERT TABLE 3 HERE]

Table 4 shows select results of cross-tabulations of the frequencies of all two-way combinations (full results are in Appendix Section A). When one dimension is fixed at level 1 the prevalence of a particular two-way combination rapidly decreases with increasing severity of the other dimension. Hence, contrasting combinations of level 1 with levels 4 or 5 are less prevalent than non-contrasting combinations between levels 2 and 3. It also appears that non-contrasting combinations of closer levels are relatively more prevalent. Therefore, these results support the approach, to operationalise implausible states as two-way contrasting combinations between level 1 with levels 4 or 5. This insight

holds when we look at the distribution of two-way combinations amongst the 418 states that are never self-reported (Appendix Section B).

[PLEASE INSERT TABLE 4 HERE]

In terms of the implausibility of the states, all 250 possible two-way combinations are observed in the dataset. Nevertheless, some of these are extremely rare: e.g. only 122 individuals self-report having SC4 alongside UA1 (viz. less than 0.01% of the sample). The two-way combinations excluded in Viney et al.[9] (MO4/5xUA1; MO4/5xSC1) are not the least frequent combinations. However, a two-way combination proposed by Dolan[1] and adapted for the EQ-5D-5L (SC4/5xUA1) is one of the 10 least frequently self-reported two-way combinations. Looking at the two-way combinations that make up the 418 *unobserved* states, the same is true: SC4/5 with UA1 are among the 11 most common combinations of unobserved states. Furthermore, AD4/5xUA1 and PD5xUA1 are more common amongst the unobserved EQ-5D-5L profiles than the four combinations proposed by Viney et al.[9]

These distributions informed the selection criteria for treatment models T3 and T4 in Stage 2. Treatment model T3 removes the least prevalent of the two-way combinations among all observed health states. These are: SC4xUA1, SC4xUA2, SC5xUA2 and SC3xUA1. Treatment model T4 removes the two-way combinations that appeared most often amongst those 418 health states that were never reported in the dataset. These are: SC4xUA1, PD5xUA1, AD5xUA1 and AD4xUA1. Across the two approaches, only SC4xUA1 is included in both treatments (and T1).

3.2 Stage 2: Re-modelling the DCE_{TTO} dataset

Excluding choice sets based on the Dolan[1] criteria (T1) resulted in the removal of responses from 30 choice sets (leaving responses from 90). The design properties were analysed, and the highest correlation between two-way combinations was below 0.3. The corresponding number of choice sets included for T2 (exclusion based on the Viney et al.[9] criteria), T3 (exclusion based on observed states from Stage 1) and T4 (exclusion based on unobserved states from Stage 1) are 86, 84 and 90, and for all correlations remained low (below 0.3) and no other statistical issues in the designs are found. This indicates that the exclusion of the choice sets does not impact the robustness of consequent designs for the statistical identification of parameter estimates. For further descriptive analyses of the designs and parameters, see Appendix Sections C to E.

Figure 1 presents the anchored coefficients of the DCE_{TTO} models. All the plots are expected to be non-positive and downward sloping. This is not always the case, however, and reversed orderings (with the implication that utility improves as severity increases) are observed, between levels 2 and 3: for MO (T1, T2, T4, C2avg), SC (T1, T4), PD (T3, C1avg, C2avg, C3), and AD (all models). The coefficient decrements for the UA dimension are monotonic.

[PLEASE INSERT FIGURE 1 HERE]

Visual inspection of the plots suggest comparator models C1avg and C3 are consistently very close to each other - since C3 uses the full dataset while C1avg is based on a random subset of those observations, the gap between C1avg and C3 can be interpreted as the magnitude of the effect of *randomly* excluding tasks associated with four two-way combinations from the study design. Similarly, the gap between C2avg and C3 represents the effect of randomly excluding tasks associated with four two-way *contrasting* combinations. These three plots suggest that PD and AD are more robust than the

other dimensions. The other plots are consistently further away from C3, indicating that there is an excess effect beyond the gaps observed for C1avg and C2avg. The plots for T1 and T4 are close to each other and tend to be shallower than the others. Across the models, it appears that UA is the most sensitive to the removal of different states, followed by AD. For example, while the models have different anchored coefficients for MO5, SC5 and PD5, all models agree on their relative ranking – however, the models disagree on which of UA5 and AD5 is worse. Across the dimensions and levels, there is no particular curve that is consistently above or below C1avg and C3. For example, the curves for C1avg and C3 mostly below the others for MO, mostly above the others for UA, and mixed with the rest for SC, PD and AD. This means that there will be no consistent patterns across the predicted values of the different models.

Figure 2 plots the predicted values of seven select states across the different models. Three of them are the most prevalent self-reported states with problems in at least two dimensions, identified in Step 1 (11122 is the 4th, 21221 is the 5th, and 21231 is the 10th most prevalent state) and jointly cover 7.5% of all observations. The remaining four are the states with the same level for each dimension (22222, 33333, 44444 and 55555), which jointly cover 0.52% of all observations. All anchored coefficients are taken at face value, including where they are disordered. The predicted values from all models are very similar to each other for state 11122, but less so for the other states. As would be expected from Figure 1, the plots from the comparator models (C1avg, C2avg and C3) for the seven states are very similar to each other. Conversely, the treatment models deviate from these comparator models, but there seems to be no consistent pattern across the treatment models. Predicted values from T3 and T4, both of which use evidence-based criteria of implausibility, are not similar to each other, and in fact, the predicted values from T4 are much more similar to those from T1, which uses a judgement-based criterion of implausibility.

[PLEASE INSERT FIGURE 2 HERE]

4. Discussion

This paper has attempted to introduce some systematic evidence for the treatment of so-called implausible states into an area of health state valuation, where practice has been dominated by judgement (and precedents) with little empirical basis. The analysis raises a number of questions that require further consideration.

- What do we mean by the term “implausible”?

In this paper, we have introduced a distinction between difficult-to-imagine states (a matter of perception or judgement of survey respondents) and rare states (a factual matter). We then explored the latter. At a conceptual level, rare states encompass two further possibilities. One is that some states are extremely rare but if the sample size were large enough, all 3,125 EQ-5D-5L states would eventually be self-reported. This possibility cannot be ruled out since we observe all two-way combinations in the self-report data. The second is that some extremely rare states involve combinations that are inherently not possible and no sample, however large, would include these states - unless self-reported in error. However, this scope for error, where people misreport their own health, makes these two possibilities practically non-distinguishable. For example, in our dataset of 1.6 million respondents, there are 274 states with just one respondent reporting that state, and some of these may be due to measurement error; but we cannot say for certain. We do, however, suggest that future research in this area does not conflate these concepts and propose to distinguish between: states that are *difficult to imagine* (independently of their actual prevalence); states that are very *rare* (which may or may not be observed

in a sample of finite observations); and states that are physiologically *impossible* (but which may still be self-reported in error). Using this terminology, this study reports on an empirically identified set of rare EQ-5D-5L states, and contrasts this with the *perceived* implausible states established in the established literature.

- Why not ask survey respondents which states are difficult to imagine?

The ideal prospective study to decide which states are difficult to imagine would ask general public respondents to score all 3,125 states on how difficult it is for them to imagine the health state. In the absence of this data, researchers have traditionally avoided using certain states in valuation studies based on their own judgement. The two main disadvantages of this approach are: first, different people may find different states difficult to imagine, and there is no guarantee that the remaining states are accepted as imaginable by all respondents; and second, the effect of avoiding those states from the valuation is not known.

Recently, Yang et al.[17] asked medical students to indicate the perceived plausibility of each of the 3,125 possible EQ-5D-5L states but did not provide an explicit definition of implausibility. One of their findings was that, while variation in levels across dimensions of a state increased perceived implausibility, there was little agreement over which states were perceived as the least plausible.

However, a different possibility that this paper contributes to might be to examine how prevalent different states actually are, remove those from the design, and explain to respondents that while some states in the design may seem unlikely, all those states have been self-reported by a general public sample. Furthermore, since the effect of removing various states from valuation designs are not known, we compare the effects of two of the most established sets of implausible states in the literature, and our evidence-based sets of rare states.

- Why use general population datasets for Stage 1?

This analysis used a dataset with self-reported health of the general population. One may question whether general population datasets cover a narrower range of EQ-5D-5L states than some patient datasets arising from trials or observational studies would. For example, Devlin et al (2010)[18] has analysed a heterogeneous patient and general public dataset and found that 161 of the 243 EQ-5D-3L states were not self-reported. The justification for carrying out this analysis on a general population dataset is that when members of the public are surveyed in health state valuation studies, the range of health states that they can *typically* imagine is more likely to come from their community (which includes people who are ill). While general population datasets can claim to represent health states in the general community, patient datasets cannot. Moreover, since the dataset used in this analysis is very large, and it is likely it includes individuals who would be eligible in various patient surveys and trial studies.

Furthermore, the analysis using general population data has another use for designing "experience-utility based value sets"[19–21] - studies where members of the public are asked to report and value their own current health state. The data are then econometrically modelled, as in conventional health state valuation studies, to generate a population value set. One technical challenge associated with such an exercise is the inefficiency associated with the highly skewed distribution of health states observed amongst the general population. Across different countries, about a third of the general population self-report full health using EQ-5D-5L[22–24] and around a half self-report full health in EQ-5D-3L.[25,26] The prevalence of health states in the general population reported in this study can inform the design of such studies, to predict the number of people necessary either to be screened or recruited in order to

observe a sufficient range of health states to make these models valid.

- Why operationalise implausibility in terms of two-way contrasting combinations?

The health state valuation literature since MVH[8] and Dolan[1] to Viney et al.[9,11] has operationalised *perceived* implausible states in terms of two-way contrasting combinations (also see Jakubczyk and Golicki,[12] Lim et al.[27] or Bagust[28]). However, there are three things to note. First, the dataset used in Stage 1 includes all 250 two-way combinations, so arguably none of these are *impossible*. Second, notwithstanding this, around 13% (423) of all possible EQ-5D-5L *states* remain unobserved and thus potentially *impossible* after 1.6 million individuals being surveyed. Third, the least prevalent two-way combinations are not necessarily contrasting combinations. These suggest that if the aim is to remove rare states from health state valuation, dropping a handful of two-way contrasting combinations is unlikely to be the most efficient way. On the other hand, we have operationalised rare/impossible states in terms of two-way combinations in two approaches based on evidence: the four least prevalent two-way combinations amongst self-reported states (T3) and the four most common two-way combinations amongst never-reported states (T4). Unlike the combinations used in the previous literature, some of these are non-contrasting combinations. We show that these two approaches result in different combinations, and they have different effects on predicted health state values.

- How to select the choice tasks to remove?

Using Dolan[1] and Viney et al.[9] as templates, we used selection criteria based on four two-way combinations in EQ-5D-5L. However, this does not control for the number of choice tasks for which observations are removed from the re-estimation process. For example, while T2 excludes 34 of the 120 choice sets and the conditional logit regression model has 5,874 observations, T3 excludes 24 choice sets and the model has 6,396 observations. Furthermore, C3 uses the full sample, with 120 choice sets and 8,020 observations. Two alternatives might have been either (a) to control for the number of *choice tasks* to remove from the re-estimation; or (b) to control for the number of *states* to be removed. While these would allow a more like for like comparison of the re-estimation results, the generation of the comparator models would be substantially more complex. Furthermore, since C1avg and C3 have very similar results, the estimates appear to be driven largely by what is dropped from the models rather than the volume of what is dropped.

- What should health state valuation studies do about difficult-to-imagine states?

Health state valuation studies have either identified perceived implausible states based on judgement and removed or avoided them from the study design, or ignored the matter altogether. Our findings suggest that removing different states from a DCE_{TTO} health state valuation design because of judgement-based or evidence-based implausibility will result in different tariff values. An unexpected finding is that the two evidence-based criteria for rare states (T3 and T4) have different effects. Results from the further analyses reported in the Appendix suggests that some of the change in parameter estimates may be due to the exclusions from the initial designs itself. However, there appears to be an excess effect attributable to the exclusion of health states that are rare and/or perceived implausible. Given this, and the findings from Yang et al.,[17] that there is little agreement across medical students on the imagined plausibility of different EQ-5D states, we believe that it is premature to exclude two-way combinations from choice designs to address the concern that respondents may feel some states are difficult to imagine. Instead, research might focus on (1) the main factors that make stylised states difficult for members of the public to imagine and (2) how best to inform respondents that all states used in the survey have been self-reported by the general population, however unlikely they may appear.

- What are strengths and limitations of this study?

To our knowledge, this is the first study to examine the prevalence of different EQ-5D-5L health states in the general population, putting forward an evidence-based set of rare states. In order to achieve the large self-report dataset necessary for this exercise, we pooled two structurally equivalent waves of the GPPS comprising 1.6 million observations. However, the great majority of respondents (80%) concentrated in only 33 health states significantly limits the sample size available for the analysis conducted in stage 1.

To explore the effects of excluding implausible states of varying definitions on the modelling of DCE_{TTO} tariffs we exploited a secondary health state valuation dataset. Ideally, a future study may address this topic prospectively, focussing on comparative experimental designs, which include and exclude (by varying definitions) implausible states, while preserving important design features. However, for the time being, re-analysing existing data is an efficient use of resources to first explore this topic.

5. Conclusion

This paper challenges the existing literature on health states that are considered "implausible" in the context of health state valuation. The literature has operationalised implausible states in terms of two-way contrasting combinations for EQ-5D. However, we find that all 250 two-way combinations of EQ-5D-5L are actually observed in a large self-report dataset of the general population, and therefore there appears to be no physiologically impossible combinations. We also identify health states that have not previously been discussed as implausible but are rarer than those that have, some of which are non-contrasting combinations. Importantly, we find that removing different health states from a valuation study may lead to value sets with potentially different characteristics, but that this will depend on the criterion of implausible states. Without empirical evidence of difficult-to-imagine states, we recommend that valuation studies of EQ-5D do not routinely remove health states from DCE_{TTO} designs. Research should also focus on how best to communicate with study participants about valuing difficult-to-imagine states.

Acknowledgements

All analyses are based on secondary data. Individual level data from the General Practitioner Patient Survey were accessed with permission by NHS England. The FEDEV dataset was funded by the EuroQol Foundation. We are grateful to Arne Risa Hole, Kim Huong Nguyen, Stephen Pudney, and two anonymous referees for Medical Decision Making for helpful suggestions. Nick Bansback, Brendan Mulhern, and Aki Tsuchiya are members of the EuroQol Group. The usual disclaimers apply.

References

1. Dolan P. Modeling valuations for EuroQol health states. *Med Care*. 1997; 35: 1095–1108.
2. Devlin NJ, Shah KK, Feng Y, Mulhern B and van Hout B. Valuing health-related quality of life: An EQ-5D-5L value set for England. *Health Econ*. 2018; 27: 7–22.
3. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Med Care*. 2002; 40: 113–128.
4. Brazier J, Roberts J and Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ*. 2002; 21: 271–292.
5. Richardson J, Sinha K, Iezzi A and Khan MA. Modelling utility weights for the Assessment of Quality of Life (AQoL)-8D. *Qual Life Res*. 2014; 23: 2395–2404.
6. Brooks R. EuroQol: the current state of play. *Health Policy*. 1996; 37: 53–72.
7. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011; 20: 1727–1736.
8. MVH Group. *The Measurement and Valuation of Health: First Report on the Main Survey*.
9. Viney R, Norman R, King MT, et al. Time trade-off derived EQ-5D weights for Australia. *Value Health*. 2011; 14: 928–936.
10. Mulhern B, Bansback N, Brazier J, et al. Preparatory study for the revaluation of the EQ-5D tariff: methodology report. *Health Technol Assess*. 2014; 18: vii-xxvi, 1-191.
11. Viney R, Norman R, Brazier J, et al. An Australian discrete choice experiment to value eq-5d health states. *Health Econ*. 2014; 23: 729–742.
12. Jakubczyk M and Golicki D. Elicitation and modelling of imprecise utility of health states. *Theory Decis*. 2019; 17: 5.
13. GP Patient Survey, https://gp-patient.co.uk/faq.
14. Mulhern B, Bansback N, Hole AR and Tsuchiya A. Using Discrete Choice Experiments with Duration to Model EQ-5D-5L Health State Preferences. *Med Decis Making*. 2017; 37: 285–297.
15. Choice Metrics. Ngene: software for experimental design (2012).
16. Bansback N, Brazier J, Tsuchiya A and Anis A. Using a discrete choice experiment to estimate health state utility values. *J Health Econ*. 2012; 31: 306–318.
17. Yang Z, Feng Z, Busschbach J, Stolk E and Luo N. How Prevalent Are Implausible EQ-5D-5L Health States and How Do They Affect Valuation? A Study Combining Quantitative and Qualitative Evidence. *Value Health*. 2019; 22: 829–836.
18. Devlin NJ, Parkin D and Browne J. Patient-reported outcome measures in the NHS: new methods for analysing and reporting EQ-5D data. *Health Econ*. 2010; 19: 886–905.
19. Burström K, Sun S, Gerdtham U-G, et al. Swedish experience-based value sets for EQ-5D health states. *Qual Life Res*. 2014; 23: 431–442.

20. Sun S, Chen J, Kind P, Xu L, Zhang Y and Burström K. Experience-based VAS values for EQ-5D-3L health states in a national general population health survey in China. *Qual Life Res*. 2015; 24: 693–703.

21. Versteegh MM and Brouwer WBF. Patient and general public preferences for health states: A call to reconsider current guidelines. *Soc Sci Med*. 2016; 165: 66–74.

22. Golicki D and Niewada M. EQ-5D-5L Polish population norms. *Arch Med Sci*. 2017; 13: 191–200.

23. Hinz A, Kohlmann T, Stöbel-Richter Y, Zenger M and Brähler E. The quality of life questionnaire EQ-5D-5L: psychometric properties and normative values for the general German population. *Qual Life Res*. 2014; 23: 443–447.

24. McCaffrey N, Kaambwa B, Currow DC and Ratcliffe J. Health-related quality of life measured using the EQ-5D-5L: South Australian population norms. *Health Qual Life Outcomes*. 2016; 14: 133.

25. Kularatna S, Whitty JA, Johnson NW, Jayasinghe R and Scuffham PA. EQ-5D-3L derived population norms for health related quality of life in Sri Lanka. *PLoS ONE*. 2014; 9: e108434.

26. Sørensen J, Davidsen M, Gudex C, Pedersen KM and Brønnum-Hansen H. Danish EQ-5D population norms. *Scand J Public Health*. 2009; 37: 467–474.

27. Lim S, Jonker MF, Oppe M, Donkers B and Stolk E. Severity-Stratified Discrete Choice Experiment Designs for Health State Evaluations. *Pharmacoeconomics*. 2018; 36: 1377–1389.

28. Bagust A. Improving valuation sampling of EQ-5D health states. *Health Qual Life Outcomes*. 2013; 11: 14.

TABLES

Table 1: Summary of models

	Criterion	Which four 2-way interactions to drop	Dropped interactions	Number of choice tasks used in Stage 2
T1	**Dolan**	As specified in Dolan (1997)	MO4/5 x UA1; SC4/5 x UA1	90
T2	**Viney**	As specified in Viney et al (2011)	MO4/5 x UA1; MO4/5 x SC1	86
T3	**Empirically reported**	Four least prevalent in reported health states in Stage 1	SC4 x UA1; SC4 x UA2; SC5 x UA2; SC3 x UA1	84
T4	**Empirically unreported**	Four most common in unreported health states in Stage 1	SC4 x UA1; PD5 x UA1; AD5 x UA1; AD4 x UA1	90
C1	**Random**	Random four from all possible interactions	Average of 100 draws	Mean approx. 91
C2	**Random/40**	Random four from the 40 contrasting interactions	Average of 100 draws	Mean approx. 92
C3	**Do nothing**	Do not drop anything	n/a	120

Note: M - Mobility; SC - Self-Care; UA - Usual Activities; PD - Pain/Discomfort; AD - Anxiety/Depression; The attached number per dimension refers to the corresponding severity level in that dimension; e.g. MO4 x UA1 refers to a health state with severe mobility problems and no problems in usual activities.

Table 2: Frequencies of the 33 most prevalent EQ-5D-5L states

EQ State	Rank	Observed prevalence	Share	Cumulative share
11111	1	593,664	36.21%	36.21%
11121	2	201,238	12.27%	48.49%
11112	3	85,526	5.22%	53.70%
11122	4	66,014	4.03%	57.73%
21221	5	38,339	2.34%	60.07%
21121	6	37,814	2.31%	62.37%
11131	7	31,711	1.93%	64.31%
11221	8	27,031	1.65%	65.96%
11113	9	25,692	1.57%	67.52%
21231	10	18,345	1.12%	68.64%
21222	11	17,006	1.04%	69.68%
11123	12	15,078	0.92%	70.60%
11222	13	13,940	0.85%	71.45%
31331	14	13,347	0.81%	72.26%
11132	15	11,228	0.68%	72.95%
21122	16	10,577	0.65%	73.59%
31231	17	10,334	0.63%	74.23%
21232	18	10,093	0.62%	74.84%
21131	19	9,468	0.58%	75.42%
21111	20	8,776	0.54%	75.95%
11231	21	8,765	0.53%	76.49%
31221	22	6,877	0.42%	76.91%
31332	23	6,819	0.42%	77.32%
11211	24	6,415	0.39%	77.71%
11133	25	5,785	0.35%	78.07%
11232	26	5,077	0.31%	78.38%

31232	27	4,880	0.30%	78.67%
11223	28	4,493	0.27%	78.95%
21211	29	4,317	0.26%	79.21%
21223	30	4,296	0.26%	79.47%
31131	31	4,267	0.26%	79.73%
31333	32	4,209	0.26%	79.99%
32332	33	4,180	0.25%	80.25%

Table 3: Overall frequencies of levels by dimensions

Dimensions and levels		Freq.	Percent	Cum.
Mobility	**1**	1,153,397	70.35	70.35
	2	227,757	13.89	84.24
	3	150,365	9.17	93.42
	4	95,655	5.83	99.25
	5	12,279	0.75	100
Self-care	**1**	1,454,836	88.74	88.74
	2	84,281	5.14	93.88
	3	66,839	4.08	97.96
	4	23,118	1.41	99.37
	5	10,379	0.63	100
Usual Activities	**1**	1,144,815	69.83	69.83
	2	254,141	15.5	85.33
	3	146,500	8.94	94.27
	4	64,210	3.92	98.18
	5	29,787	1.82	100
Pain/Discomfort	**1**	760,823	46.41	46.41
	2	503,078	30.69	77.09
	3	259,870	15.85	92.94
	4	94,738	5.78	98.72
	5	20,944	1.28	100
Anxiety/Depression	**1**	1,124,379	68.58	68.58
	2	323,255	19.72	88.3
	3	142,460	8.69	96.99
	4	33,631	2.05	99.04
	5	15,728	0.96	100

Table 4(a): Cross-tab of frequencies of respondents at different levels of mobility and self-care.

Self-Care / Mobility	1	2	3	4	5	Total
1	1,143,228	6,911	2,244	505	509	1,153,397
2	199,003	23,619	3,937	629	569	227,757
3	88,432	33,159	25,962	2,048	764	150,365
4	22,108	19,600	33,063	17,934	2,950	95,655
5	2,065	992	1,633	2,002	5,587	12,279
Total	1,454,836	84,281	66,839	23,118	10,379	1,639,453

Table 4(b): Cross-tab of frequencies of respondents at different levels of mobility and usual activities.

Usual Activities Mobility	1	2	3	4	5	Total
1	1,049,134	83,060	16,438	3,210	1,555	1,153,397
2	78,436	120,365	23,754	3,302	1,900	227,757
3	13,932	43,595	76,110	12,317	4,411	150,365
4	1,924	6,565	29,154	43,260	14,752	95,655
5	1,389	556	1,044	2,121	7,169	12,279
Total	1,144,815	254,141	146,500	64,210	29,787	1,639,453

Table 4(c): Cross-tab of frequencies of respondents at different levels of self-care and usual activities.

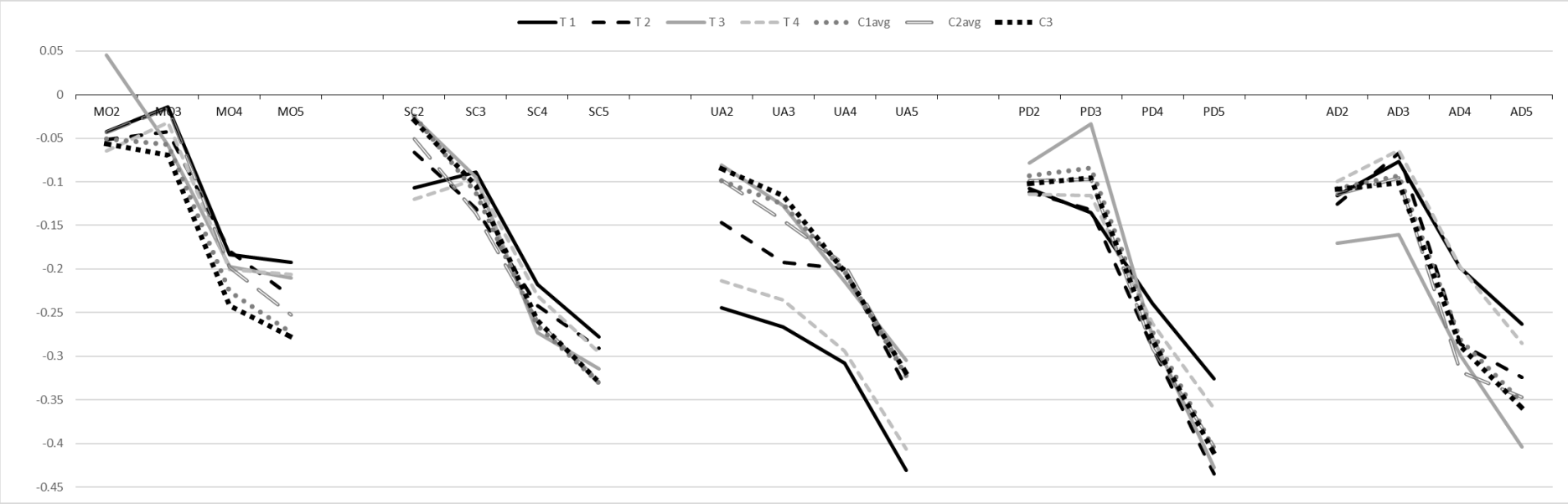
Usual Activities Self-Care	1	2	3	4	5	Total
1	1,141,250	219,602	76,849	12,395	4,740	1,454,836
2	2,452	30,931	34,628	12,296	3,974	84,281
3	433	3,001	32,689	23,853	6,863	66,839
4	122	226	1,827	14,354	6,589	23,118
5	558	381	507	1,312	7,621	10,379
Total	1,144,815	254,141	146,500	64,210	29,787	1,639,453

Table 4(d): Cross-tab of frequencies of respondents at different levels of usual activities and pain/discomfort.

Pain/Discomfort Usual Activities	1	2	3	4	5	Total
1	723,413	342,443	71,433	6,710	816	1,144,815
2	24,167	131,627	83,765	13,470	1,112	254,141
3	8,428	21,204	79,700	34,045	3,123	146,500
4	2,414	4,001	16,308	31,927	9,560	64,210
5	2,401	3,803	8,664	8,586	6,333	29,787
Total	760,823	503,078	259,870	94,738	20,944	1,639,453

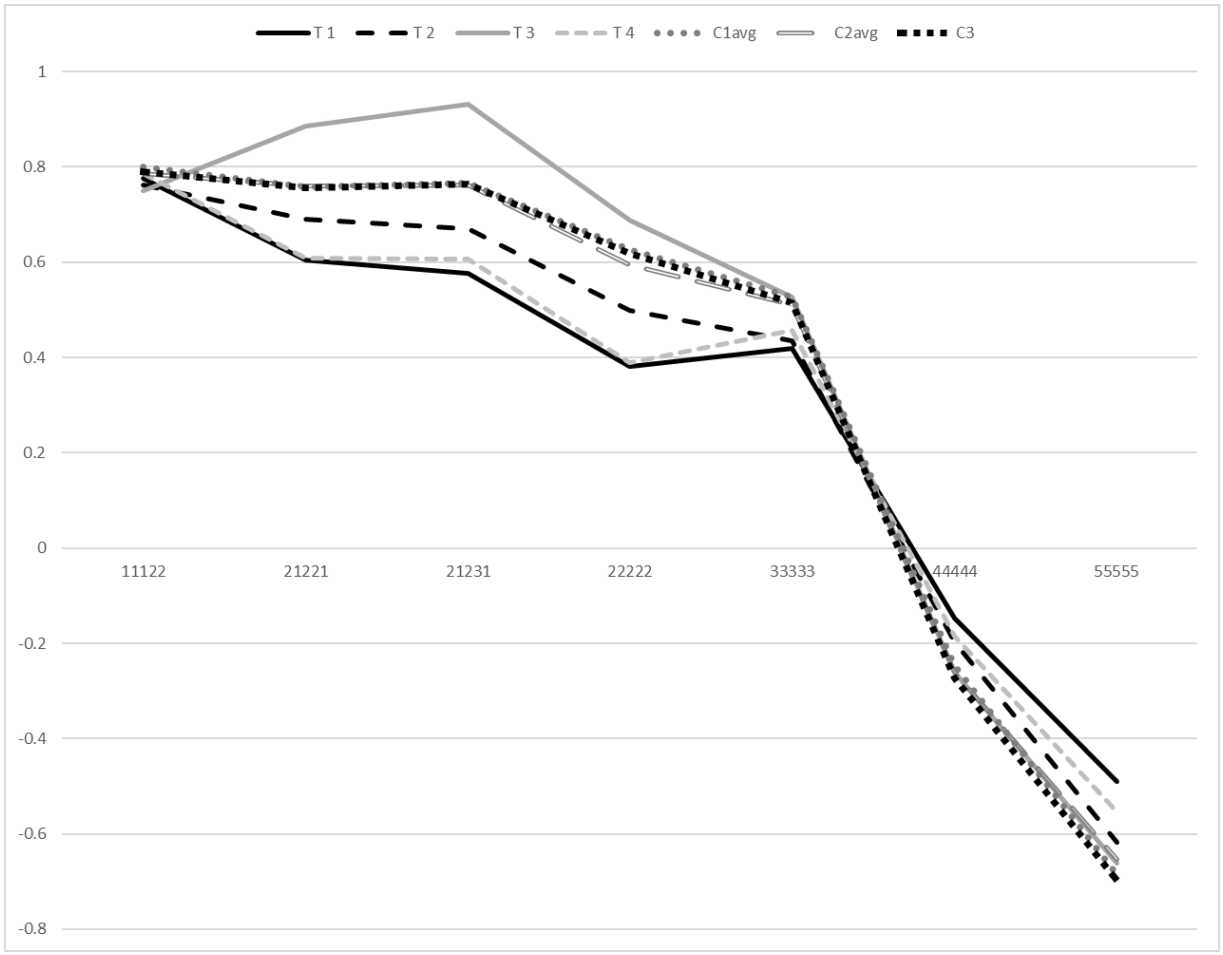
Figures

Figure 1: Comparison of (mean) anchored coefficients.



- T1: Treatment model 1 (exclusion based on the Dolan criteria)
- T2: Treatment model 2 (exclusion based on the Viney et al criteria)
- T3: Treatment model 3 (exclusion based on observed states from Stage 1)
- T4: Treatment model 4 (exclusion based on unobserved states from Stage 1)
- C1avg: average anchored coefficients from the 100 comparator models (random exclusion)
- C2avg: average anchored coefficients from the 100 comparator models (random exclusion of contrasting interactions), and
- C3: Comparator model 3 (the full model).

Figure 2: Comparison of (mean) predicted values of select states.



- T1: Treatment model 1 (exclusion based on the Dolan criteria)
- T2: Treatment model 2 (exclusion based on the Viney et al criteria)
- T3: Treatment model 3 (exclusion based on observed states from Stage 1)
- T4: Treatment model 4 (exclusion based on unobserved states from Stage 1)
- C1avg: average anchored coefficients from the 100 comparator models (random exclusion)
- C2avg: average anchored coefficients from the 100 comparator models (random exclusion of contrasting interactions), and
- C3: Comparator model 3 (the full model).