An EQ-5D-5L value set for Belgium - Electronic supplementary material

PharmacoEconomics Open

Nicolas Bouckaert[1,*]**, Irina Cleemput**[1], **Stephan Devriese**[1], **Sophie Gerkens**[1]

[1] Belgian Health Care Knowledge Centre (KCE), Boulevard du Jardin Botanique 55, 1000 Bruxelles, Belgium

***** corresponding author: nicolas.bouckaert@kce.fgov.be

Statements and Declarations

Funding. This research was co-funded by the Belgian Health Care Knowledge Centre (KCE) and the EuroQol Research Foundation (EQ Project 20170490 and 20170491). The KCE is a federal institution which is financed by the National Institute for Health and Disability Insurance (NIHDI, RIZIV – INAMI), the Federal Public Service of health, food chain safety and environment, and the Federal Public Service of social security. The development of health services research studies is part of the legal mission of the KCE. Although the development of the studies is paid by the KCE budget, the sole mission of the KCE is providing scientifically valid information.

Conflicts of interest. Nicolas Bouckaert, Irina Cleemput, Stephan Devriese, and Sophie Gerkens declare that they have no conflicts of interest directly relevant to the content of this article.

ESM1 – Sampling procedure and post-stratification weights

Sampling procedure

The sample was constructed using a multistage, stratified, cluster sampling with unequal probability design. Although it is a more complex sampling method than simple random sampling, it comes with several advantages. A clustered and stratified design allows for a larger sample to be interviewed in the same time frame and with similar resources than a simple random sampling would. Moreover, stratification brings the advantage of ensuring sufficient representation of the population on certain variables like age and sex. At the same time, stratification can increase efficiency of sample estimates by reducing variability (and thus standard errors) compared to random sampling without stratification [1]. In contrast, clustering and weighting can increase variability compared to simple random sampling. However, the advantages of feasibility and stratification were considered to outweigh the disadvantages of the clustering and weighing aspects of the design. The sampling procedure was similar in nature to the procedure used for the Belgian Health Interview Survey [2].

To assure a sufficient coverage of the Belgian territory, municipalities were sampled as clusters within provinces taking into account unequal probability of inclusion based on population size within the provinces and municipalities. This to find a balance between sufficient representativeness and feasibility. Additionally, sampling was further stratified on age categories and sex to obtain a sufficient balance in respondent characteristics. Age and sex were chosen as strata because of the availability at the National Register for sampling and because these variables were deemed important to have weighed proportionally on the construction the EQ-5D-5L value set. Age was categorised in eight groups to avoid too small strata ([18,30), [30,40), [40,50), [50,60), [60,70), [70,80), [80,90), [90,100)).

About 1000 successful interviews were targeted to produce the EQ-5D-5L value set with sufficient precision. We assumed a conservative estimated response rate of 10% in deciding how many potential participants to sample from the National Reguister. For each interview, 10 potential participants were sampled at random in the same stratum. Hence, a total of 10000 potential participants were then sampled.[a]

Post-stratification weights

Post-stratification were used to adjust the estimation to correct for differences between the planned and realised interviews and subsequently to obtain preference values representative for the Belgian population. Separate weights were calculated for the final sample of 892 respondents as well as for the larger sample used in the sensitivity analysis (913 respondents in the sample without exclusions based on implausible response patterns). They were calculated as follows:

1. We checked for 'empty strata', combinations of province, age category and sex where interviews were planned but where, in the end, no interviews were conducted. We found 7 out of 151 original strata to be empty. For the calculation of post-stratification weights, the empty strata need to be merged with other non-

[a] Note that not all 10 000 potential participants were contacted. For each required interview in a specific stratum, only one or a few potential participants needed to be contacted in order to find someone willing to participate in the survey. Exceptionally, all 10 potential candidates were contacted without successfully recruiting a respondent.

empty strata to create post-survey strata. We defined post-survey strata as close as possible to the original sampling strata and merge an empty stratum with the stratum in the same province and of the same sex, but of the age category just below, so that the interviews in that stratum represented the interviews in the empty strata as well.

2. For each post-survey stratum (province, age category, and sex), the number of people that each interview represents was calculated as the population size in a given post-survey stratum divided by the number of interviews realized in that stratum.
3. The post-stratification weight of an interview was then calculated as the number of people that each interview represents divided by the Belgian adult population on 1 January 2017 (as in the original sampling) [3].

ESM2 – Data collection process

Data collection time frame, recruitment of interviewers and training

The data collection process ran from 1 May 2018 to 30 September 2020. The pace of the study, the number of interviews, as well as the impact of training and quality control is described in Fig. 1. As shown in this figure, the data collection time frame can be divided in 4 phases, each phase beginning with a new training of interviewers and pilot studies.

- Phase 1 (March 2018 to August 2018): In accordance with the EQ-VT protocol, which puts forward to deploy between 8 and 14 interviewers, a total of 11 interviewers were recruited at the start of the study. Interviewers were selected based on their experience with handling complex interviews. A one-day training as well as 5 pilot studies per interviewer were performed prior to the fieldwork. While this initial training of interviewers was performed at the beginning of March 2018, the actual data collection only started in May 2018 due to an unexpected long delay in obtaining a list of randomly sampled candidates from the National Register. During phase 1, nearly half of the interviews (44%) did not pass the quality control and 4 interviewers (as well as all their conducted interviews) were excluded.
- Phase 2 (September 2018 to August 2019): Because the quality of interviews was judged insufficient, an additional one-day training was organized and 4 new interviewers were enrolled and trained. It was also decided that each interviewer would receive a printed version of the EQ-5D-5L questionnaire in their language to help respondents to correctly locate the health state described on the screen in the EQ-5D-5L questionnaire and so to facilitate the assessment of a state's severity. During this phase, two interviewers did not pass the quality control and were excluded (as well as all their interviews).
- Phase 3 (September 2019 to June 2020): In May 2019, because the pace of the study was judged insufficient and because Profacts was no longer able to keep enough interviewers active in the field to successfully carry out the rest of the study, it was decided to temporarily halt the study and to recruit and train new interviewers (n=13). This decision was made in agreement with EuroQol. Because the limit of 14 interviewers required in the EQ-VT protocol was surpassed, a more intensive training and follow-up was pursued to reduce potential interviewers' bias as much as possible. This intensive training started with a one-day training, during which one pilot was done. Next, three additional pilot tests per interviewer were performed on family and friends (at the interviewer's convenience, e.g. at home). A second day of training then consisted in (1) a debriefing on the 4 conducted pilot tests, (2) an additional 6 pilot tests on a group of mock interviewees, which were monitored by the research team and (3) a final debriefing. The dropout rate on these pilots (i.e. the percentage of interviews flagged in the quality control process) had to be 30% or less before an interviewer was authorized to work in the field. In case the dropout rate exceeded the threshold, additional pilots were performed. During this phase, one interviewer did not pass the quality control and was excluded (as well as all his/her interviews).
- Phase 4 (July 2020 to September 2020): Finally, in March 2020, because of the COVID-19 pandemic and the imposed lock-down, it was decided, in consultation and in agreement with the EuroQol group, to stop the data collection before having reached the target of 1000 interviews and to only conduct a limited number of

interviews in the region of Brussels and in French-speaking Brabant (province Brabant-Wallon) in order to increase the representativeness of the sample (not enough interviews were performed in these two areas). These last interviews could only be conducted once the lockdown measures were released (i.e. from July 2020 onwards). Three interviewers working in these area were retrained during a 2 hours training to ensure they were ready to be deployed once again in the field after the break imposed by the lockdown. In order to carry out these last interviews in respect with the legal protective measures, 2 choices were proposed to the respondents: performing the interview at distance via Microsoft Teams (n= 5 interviews) or face-to-face (n= 43 interviews) by wearing a mask and respecting distancing and other protective rules. The resurgence in the number of COVID-19 cases and a second lockdown led to the termination of the study, in accordance with EuroQol. At that time, 916 interviews were performed that had passed the quality control process.

Throughout the study, a total of 22 interviewers were deployed in the field and conducted a total of 916 interviews that had passed the quality control process, leading to an average of 42 interviews per interviewer.

Fig. 1 Pace of the study, number of interviews and impact of training and quality control

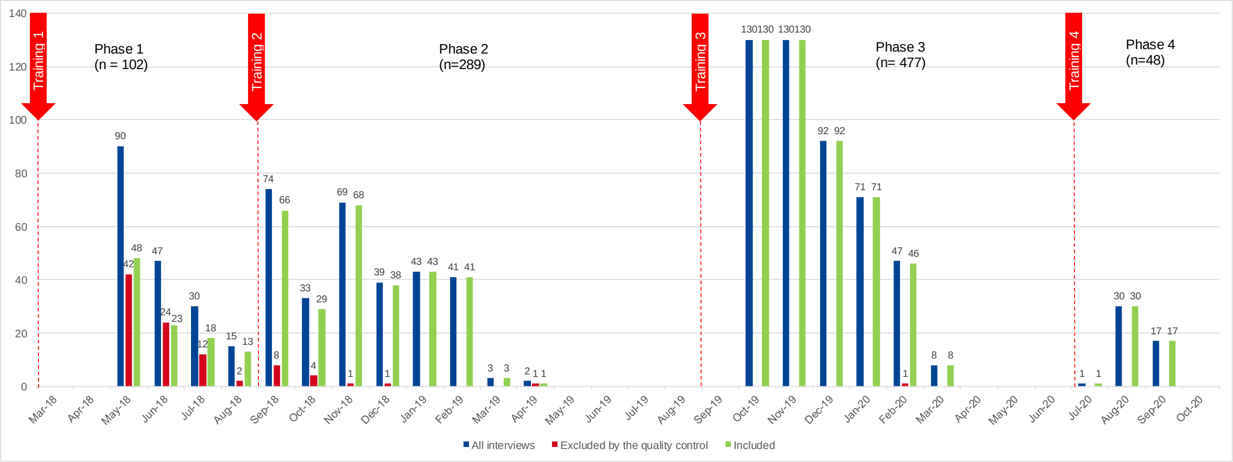
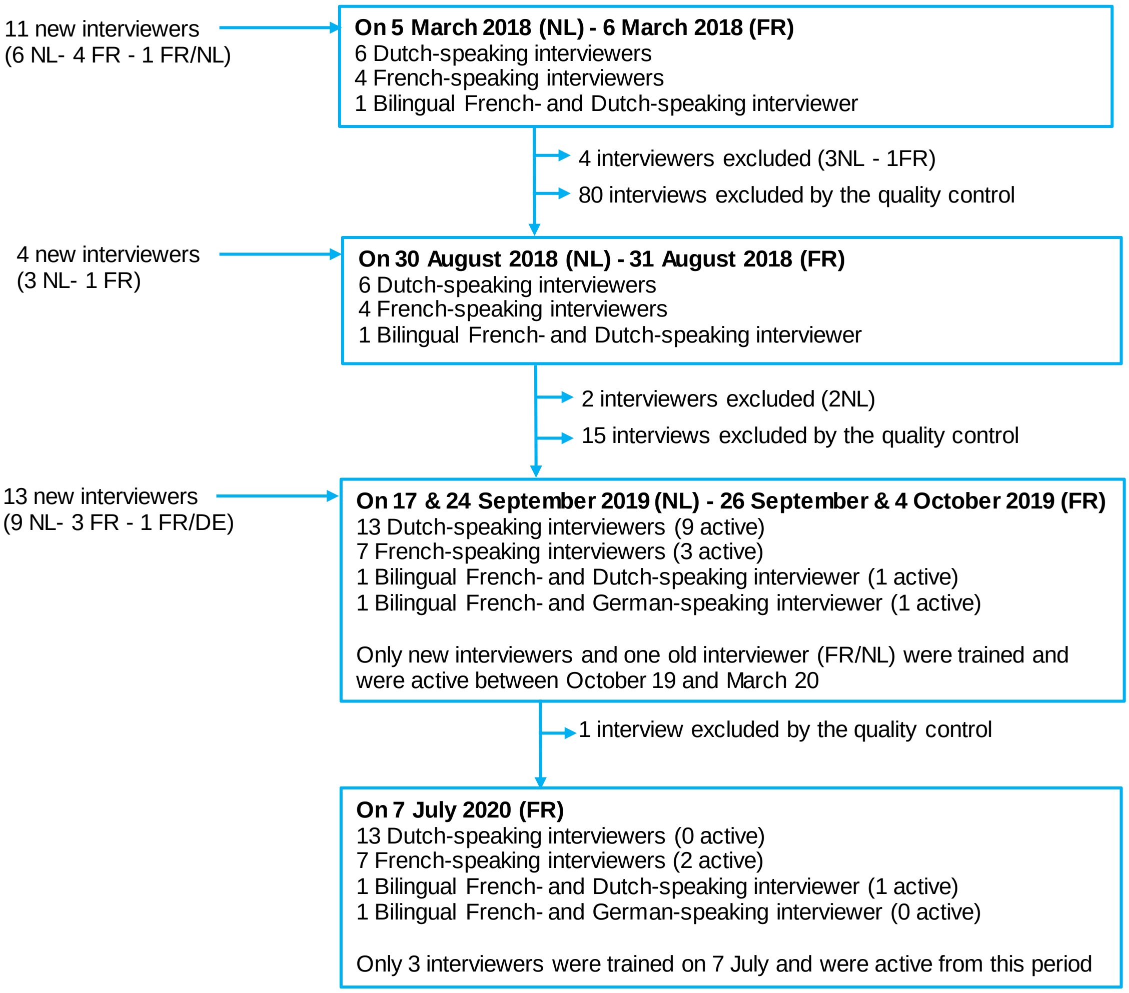


Fig. 2 Flow chart on the change over time in the number of interviewers



11 new interviewers
(6 NL- 4 FR - 1 FR/NL)

On 5 March 2018 (NL) - 6 March 2018 (FR)
6 Dutch-speaking interviewers
4 French-speaking interviewers
1 Bilingual French- and Dutch-speaking interviewer

4 interviewers excluded (3NL - 1FR)
80 interviews excluded by the quality control

4 new interviewers
(3 NL- 1 FR)

On 30 August 2018 (NL) - 31 August 2018 (FR)
6 Dutch-speaking interviewers
4 French-speaking interviewers
1 Bilingual French- and Dutch-speaking interviewer

2 interviewers excluded (2NL)
15 interviews excluded by the quality control

13 new interviewers
(9 NL- 3 FR - 1 FR/DE)

On 17 & 24 September 2019 (NL) - 26 September & 4 October 2019 (FR)
13 Dutch-speaking interviewers (9 active)
7 French-speaking interviewers (3 active)
1 Bilingual French- and Dutch-speaking interviewer (1 active)
1 Bilingual French- and German-speaking interviewer (1 active)

Only new interviewers and one old interviewer (FR/NL) were trained and were active between October 19 and March 20

1 interview excluded by the quality control

On 7 July 2020 (FR)
13 Dutch-speaking interviewers (0 active)
7 French-speaking interviewers (2 active)
1 Bilingual French- and Dutch-speaking interviewer (1 active)
1 Bilingual French- and German-speaking interviewer (0 active)

Only 3 interviewers were trained on 7 July and were active from this period

ESM3 – Four core model specifications

The EuroQol protocol does not specify a standard regression model to be used to generate utility values for an EQ-5D-5L value set. The regression models selected for the EQ-5D-5L value set vary between countries, reflecting the differences in underlying data characteristics, evaluation process and model selection criteria. Nonetheless, a set of four core regression specifications – one additive model and three multiplicative models – exists and serves as backbone for the approach followed in the current study [4-7].

First, the modelling strategy for the estimation of cTTO models is discussed. While cTTO was the main technique to determine utility values within the EQ-5D framework, it has been accompanied by a set of DCE questions in more recent valuations. DCE has a strong theoretical foundation in random utility theory and has been increasingly used to quantify stated preferences for health [8-10]. Therefore, also models estimated using the DCE data are presented as well as the hybrid approach, combining cTTO and DCE data in one model.

Modelling cTTO valuation

cTTO models can be described as follows

$$DU_{ij} = 1 - U_{ij} = I + \beta_{cTTO} X_j + \varepsilon_{ij} \qquad (1)$$

where U_{ij} and DU_{ij} are, respectively, the utility and disutility related to health state j reported by respondent i; I is the potential intercept in the regression; $\beta_{cTTO} X_j$ represents one of the four core regression specifications that we consider (see below) with coefficients (β_{cTTO}) and variables (X_j) representing the valuation of the different dimensions and levels; ε_{ij} is the error term.

Dependent variable

Disutilities rather than utilities were estimated as dependent variable [11]. Disutility is defined as the deviation in utility from the full health state with utility value 1. The reason for this choice is that utility values can be both positive and negative – they range from -1 to 1 – which complicates the estimation of coefficients. Disutility, on the other hand, is always positive and ranges from 0 to 2. Note that we treat the dependent variable as a continuous variable even though this is in reality not the case as respondents can only give 41 distinct values [6].

Four core regression specifications

Generally, there are four main regression specifications to model EQ-5D-5L value sets [4-7]. All specifications are made up of the same 20 binary variables, but of a different set of estimated coefficients. For each of the five dimensions – i.e. mobility (MO), self-care (SC), usual activities (UA), pain/discomfort (PD), anxiety/depression (AD) –, four binary variables were defined (x_{dl}) indicating whether or not the health state is characterized by a problem on dimension d at level l – i.e. levels 2, 3, 4 or 5, with level 1 being the reference level.

Additive 20-coefficients model **(ADD20).** One coefficient (β_{dl}) for each binary variable or a different disutility associated with each combination of dimension d and level l. The ADD20 is the most flexible specification with the highest number of estimated coefficients and can be mathematically described as follows [4, 12]:

$$\begin{aligned}\beta_{cTTO}X_j = \sum\nolimits_d \sum\nolimits_l \beta_{dl}x_{dl} = {} & \beta_{MO2}x_{MO2} + \beta_{MO3}x_{MO3} + \beta_{MO4}x_{MO4} + \beta_{MO5}x_{MO5} \\ & + \beta_{SC2}x_{SC2} + \beta_{SC3}x_{SC3} + \beta_{SC4}x_{SC4} + \beta_{SC5}x_{SC5} \\ & + \beta_{UA2}x_{UA2} + \beta_{UA3}x_{UA3} + \beta_{UA4}x_{UA4} + \beta_{UA5}x_{UA5} \\ & + \beta_{PD2}x_{PD2} + \beta_{PD3}x_{PD3} + \beta_{PD4}x_{PD4} + \beta_{PD5}x_{PD5} \\ & + \beta_{AD2}x_{AD2} + \beta_{AD3}x_{AD3} + \beta_{AD4}x_{AD4} + \beta_{AD5}x_{AD5}\end{aligned} \tag{2}$$

The remaining three specifications are multiplicative models, which are more restrictive, i.e. with fewer coefficients. Given their nonlinear nature, they are less frequently used [4, 5, 13].

Multiplicative 8-coefficients model **(MULT8).** There is one coefficient for each dimension (β_d), and one coefficient for each severity level (L_l) with level 1 and level 5 standardized at 0 and 1, respectively. The multiplication of a dimension and a level coefficient gives the disutility associated with each combination of dimension and level;

The dimension coefficient should be interpreted as the disutility of having a problem on dimension d at level 5. The level coefficients have a value between 0 and 1 and indicate the disutility for level l in proportion to level 5. There are coefficients for levels 2, 3 and 4, with levels 1 and 5 standardized at a value of, respectively, 0 and 1. It is assumed that the relative distance between the levels is identical for all dimensions. The mathematical formulation of the MULT8 specification is as follows:

$$\begin{aligned}\beta_{cTTO}X_j &= \sum\nolimits_l \left(\sum\nolimits_d \beta_d x_{dl}\right) L_l \\ &= (\beta_{MO}x_{MO2} + \beta_{SC}x_{SC2} + \beta_{UA}x_{UA2} + \beta_{PD}x_{PD2} + \beta_{AD}x_{AD2})L_2 \\ &+ (\beta_{MO}x_{MO3} + \beta_{SC}x_{SC3} + \beta_{UA}x_{UA3} + \beta_{PD}x_{PD3} + \beta_{AD}x_{AD3})L_3 \\ &+ (\beta_{MO}x_{MO4} + \beta_{SC}x_{SC4} + \beta_{UA}x_{UA4} + \beta_{PD}x_{PD4} + \beta_{AD}x_{AD4})L_4 \\ &+ (\beta_{MO}x_{MO5} + \beta_{SC}x_{SC5} + \beta_{UA}x_{UA5} + \beta_{PD}x_{PD5} + \beta_{AD}x_{AD5})\end{aligned} \tag{3}$$

Multiplicative 9-coefficients model **(MULT9)** extends the MULT8 specification with one additional coefficient, L_5. This coefficient allows to make a distinction at level 5 between the dimensions mobility, self-care and usual activities on the one hand and pain/discomfort and anxiety/depression on the other hand. The reason for this distinction is the difference in wording of level 5 in the different dimensions, described as “unable to” in the former three dimensions and “extreme” in the latter two dimensions. Hence, in MULT9, it is assumed that the relative distance between levels 1 to 4 is identical for all dimensions, but might differ between levels 4 and 5 for the first three and the last two dimensions. The mathematical formulation of the MULT9 specification is as follows:

$$\begin{aligned}\beta_{cTTO}X_j &= (\beta_{MO}x_{MO2} + \beta_{SC}x_{SC2} + \beta_{UA}x_{UA2} + \beta_{PD}x_{PD2} + \beta_{AD}x_{AD2})L_2 \\ &+ (\beta_{MO}x_{MO3} + \beta_{SC}x_{SC3} + \beta_{UA}x_{UA3} + \beta_{PD}x_{PD3} + \beta_{AD}x_{AD3})L_3 \\ &+ (\beta_{MO}x_{MO4} + \beta_{SC}x_{SC4} + \beta_{UA}x_{UA4} + \beta_{PD}x_{PD4} + \beta_{AD}x_{AD4})L_4 \\ &+ (\beta_{MO}x_{MO5} + \beta_{SC}x_{SC5} + \beta_{UA}x_{UA5}) + (\beta_{PD}x_{PD5} + \beta_{AD}x_{AD5})L_5\end{aligned} \tag{4}$$

Multiplicative 11-coefficients model **(MULT11)**. It is assumed that the difference between the first three and the last two dimensions not only affects the relative distance between levels 4 and 5 (as is the case for MULT9), but the relative distance between all levels. Hence for the two subgroups of dimensions a separate set of level coefficients is defined, i.e. LE_l for the “extreme”-group and LU_l for the “unable to”-group. As for the MULT8 specification, level 1 and level 5 are standardized for each set at 0 and 1, respectively. It can be mathematically specified as follows:

$$\begin{aligned}\beta_{cTTO}X_j = {} & (\beta_{MO}x_{MO2} + \beta_{SC}x_{SC2} + \beta_{UA}x_{UA2})LU_2 + (\beta_{PD}x_{PD2} + \beta_{AD}x_{AD2})LE_2 \\ & + (\beta_{MO}x_{MO3} + \beta_{SC}x_{SC3} + \beta_{UA}x_{UA3})LU_3 + (\beta_{PD}x_{PD3} + \beta_{AD}x_{AD3})LE_3 \\ & + (\beta_{MO}x_{MO4} + \beta_{SC}x_{SC4} + \beta_{UA}x_{UA4})LU_4 + \beta_{PD}x_{PD4} + \beta_{AD}x_{AD4})LE_4 \\ & + (\beta_{MO}x_{MO5} + \beta_{SC}x_{SC5} + \beta_{UA}x_{UA5}) + (\beta_{PD}x_{PD5} + \beta_{AD}x_{AD5})\end{aligned} \tag{5}$$

As the mathematical formulations in Equations (2) to (5) make clear, the multiplicative specifications are constrained variants of the more flexible additive model, with fewer degrees of freedom.

Each of these four core specifications can be further adjusted: with/without *intercept*; with/without *random effects* (a respondent-specific component in the error term, that can be interpreted as individual variation around the intercept); with/without *heteroscedasticity* (a correction for increasing variability in reported cTTO values as health states worsen); with/without *censoring* (correction for respondents who would like to value health states below the minimum cTTO value capped at -1 by design).

Modelling DCE valuation

In the DCE task, respondents compared a pair of health states A and B and chose the better one, i.e. the health state with the higher utility value. These choices give information on the relative preference of one health state over another. Contrary to the cTTO valuation, the DCE valuation did not provide direct utility values that are anchored to a scale where a value of 1 represents full health and a value of 0 represents dead, but rather relative values between levels and dimensions [12].

Let us consider the following utilities for health states A and B in pair p evaluated by respondent i:

$$U_{ipA} <?> U_{ipB} \tag{6}$$

As for cTTO valuation, disutilities are modelled rather than utilities. The disutilities can be specified in a similar way as in Equation (1) using the same four core specification.

$$DU_{ipA} = 1 - U_{ipA} = I + \beta_{DCE}X_{pA} + \varepsilon_{ipA} <?> DU_{ipB} = 1 - U_{ipB} = I + \beta_{DCE}X_{pB} + \varepsilon_{ipB} \tag{7}$$

The comparison of disutilities can be translated in a binary choice variable (C) indicating whether disutility is highest for health state A ($C = 1$) or for health state B ($C = 0$)

$$\begin{gathered}\text{If } DU_{ipA} - DU_{ipB} = \beta_{DCE}(X_{pA} - X_{pB}) + (\varepsilon_{ipA} - \varepsilon_{ipB}) > 0 \ \text{ then } C_{ip} = 1 \\ \text{Else if } DU_{ipA} - DU_{ipB} = \beta_{DCE}(X_{pA} - X_{pB}) + (\varepsilon_{ipA} - \varepsilon_{ipB}) \leq 0 \ \text{ then } C_{ip} = 0\end{gathered} \tag{8}$$

The DCE model uses this binary choice variable C_{ip} as dependent variable, while difference between the binary dimension-level variables for health states A and B are used as independent variables. Under the assumption that the errors follow an extreme value distribution, the coefficients can be estimated by a logit model.

Note that by taking the difference between health states A and B, the intercept has disappeared, because we assumed that the intercept is the same for all health states (see Equation (1)). If the intercept is significant, this would signify that there is a systematic disutility difference between health states selected as choice A and health states selected as choice B, i.e. that the intercept in health state A is significantly different from the intercept in health state B [6].

For the DCE valuation, the four core specifications are estimated without correction for heteroscedasticity and random effects (it can be shown that the random effect cancels out as was the case for the intercept). By design, the DCE data are not censored.

Hybrid model

In order to maximize the available information in the estimation and hence improve accuracy, a hybrid model can be used to estimate simultaneously a single set of coefficients on the cTTO and DCE data. For a detailed description of the hybrid model, we refer the interested reader to Ramos-Goñi et al. [14]. The hybrid model has been frequently selected as final model to generate EQ-5D-5L utility values [12, 15-19].

The main underlying assumption in the hybrid model is that the coefficients in the DCE model (β_{DCE}) can be rescaled to match the coefficients from the cTTO model (β_{cTTO}). Following Ramos-Goñi et al. [14], proportionality between both coefficients is assumed, or put differently, the same scaling parameter applies to all coefficients:

$$\beta_{DCE} = \frac{\beta_{cTTO}}{\theta} \Rightarrow \beta_{HYB} = \beta_{cTTO} = \beta_{DCE}\theta \tag{9}$$

Although the coefficients from the hybrid model (β_{HYB}) are estimated on both cTTO and DCE data, they have exactly the same interpretation as the coefficients from the cTTO model in terms of utility decrements on a scale where value 1 represents full health and value 0 represents dead. Moreover the rescaling parameter θ can also be used to rescale coefficients from a DCE-only model and make them more comparable to the coefficients from the cTTO-only or hybrid model.

The cTTO part of the hybrid model can be specified in various ways – with/without random effect, intercept, correction for heteroscedasticity, censoring etc., while the DCE part of the hybrid model is specified without constant term. Of course the same core regression specification – ADD20, MULT8, MULT9, or MULT11 – were applied to both the cTTO and DCE part of the hybrid model.

ESM4 – Model selection criteria

A large range of regression models were evaluated and compared with the aim to select one final model that is best able to predict utility values for all health states defined by the EQ-5D-5L descriptive system based on the cTTO and DCE responses for a small set of these health states. Model selection was based on logical consistency of the coefficient estimates, goodness of fit, predictive accuracy and theoretical considerations.

In a first stage, logical consistency, goodness of fit and predictive accuracy were assessed for the models using only cTTO data to get a ranking of best potential specifications. In a second stage, theoretical considerations as well as the results from the DCE regression models were used to consider the use of a censored model and/or a hybrid model.

Logical consistency

The estimated coefficients of the final model must be logically consistent. Coefficients are considered logically consistent if the disutility in a health dimension does not decrease with the severity level. This implies the following:

- **ADD20-model**: $I \geq 0$ and for all dimensions (d) $\beta_{d5} \geq \beta_{d4} \geq \beta_{d3} \geq \beta_{d2} \geq 0$
- **MULT8-model**: $I \geq 0$, for all dimensions (d) $\beta_d > 0$, and for the levels $1 \geq L_4 \geq L_3 \geq L_2 \geq 0$
- **MULT9-model**: $I \geq 0$, for all dimensions (d) $\beta_d > 0$, and for the levels $\min(1, L_5) \geq L_4 \geq L_3 \geq L_2 \geq 0$
- **MULT11-model:** $I \geq 0$, for all dimensions (d) $\beta_d > 0$, and for the levels $1 \geq LU_4 \geq LU_3 \geq LU_2 \geq 0;\ 1 \geq LE_4 \geq LE_3 \geq LE_2 \geq 0$

Goodness of fit

Goodness of fit refers to the ability of an estimated model to fit the observed data. It was evaluated using the mean absolute error (MAE), a frequently used measure for predictive accuracy, and the Bayesian information criterion (BIC), a measure that combines goodness of fit and model complexity [20, 21]. The goodness of fit of the estimated regression models was evaluated by first ranking the models by their performance on each measure (MAE and BIC) and then summing up both ranks.

The **mean absolute error** is computed as the average of the sum of the absolute values of the difference between the predicted and observed utility value for a health state (see Equation (10)). Note that the MAE was not assessed at the observation level (i.e. a health state evaluated by a respondent), but at the population mean of the health state, as population-level health state utilities are the main outcome of the data analysis. Therefore, in the computation of the MAE, the weighted average observed utility of a health state is compared to the predicted value derived from the regression model. A lower MAE is favoured as it indicates a better match between observed and predicted population values.

$$MAE_M = \frac{1}{N}\sum_j \left|\frac{\sum_i w_i U_{ij}}{\sum_i w_i I_{ij}} - \widehat{U}_{Mj}\right| \tag{10}$$

Where MAE_M is the MAE value of regression model M; N equals the number of observed health states that are taken into account in the computation of the MAE, i.e. 86 health states in case of the cTTO data; w_i is the population weight of respondent i; I_{ij} is a dummy indicating that respondent i evaluated health state j; U_{ij} is the utility and disutility related to health state j reported by respondent; and $\widehat{U}_{Mj}$ represent the predicted utility value for health state j based on the estimated coefficients in regression model M.

The ability of the model to have a good fit with the data is important, but a narrow focus on goodness of fit may be misleading as the inclusion of additional coefficients always increases a model's fit. An over-fitted model is, however, not able to distinguish the factors driving the utility values from random variation and does not produce trustworthy predictions. Therefore, also a second measure was used, the **Bayesian information criterion**. It presents a trade-off between the goodness fit of the model (evaluated by the likelihood value) and the complexity or parsimony of the model (evaluated by the number of estimated parameters) and reduces the risk of selecting overfitted models. A model with a lower BIC value is preferred and reflects a better fit and/or fewer estimated coefficients. Only models using the same input data can be compared to each other using the BIC value. Hence, for example cTTO models and hybrid models cannot be compared [10, 19]. The BIC is related to another often used measure, the Akaike's information criterion (AIC), but the BIC penalizes more heavily for additional parameters [20].

$$BIC_M = -2\ln(L_M) + k_M ln(n) \tag{11}$$

Where BIC_M is the BIC of regression model M; $\ln(.)$ is the natural logarithm function; L_M is the likelihood value of regression model M; k is the number of estimated coefficients of regression model M; and n is the number of observations used in the estimation.

Predictive accuracy

The ability of a model to predict unobserved values is a vital element of the model's performance and thus predictive accuracy is considered a more important selection criterion than goodness of fit.

Cross-validation techniques were used to assess the predictive accuracy of the cTTO-only models. In this case, the modelling data were split into two parts: a training set and a validation set. First, the regression model was fitted on the training data. Next, out-of-sample utility values were predicted for health states that were withheld from the training data but observed in the validation set. Finally, the MAE was calculated for these health states (see above). In addition, the logical consistency of each of the fitted models in the cross-validation was assessed.

Based on Rand-Hendriksen et al. [4], three different cross-validations were used, using a different split between training and validation data.

1. **Leave-one-state-out:** one health state was left out of the modelling dataset to create a training dataset with 85 health states and to predict the utility of the left-out state. Hence, the majority of the data was available to fit the model. This subdivision was replicated 86 times, sequentially removing each of the observed health

states once. The final MAE value for the leave-one-state-out cross-validation is the average of the MAE resulting from the 86 replications.

2. **Leave-one-block-out:** in the cTTO task, the evaluated health states were grouped in ten blocks, randomly assigned to the respondents. Hence health states in a particular block were all valued by the same set of individuals and are not fully independent. In case only one state is left out, information on how these respondents valued other health states was still observed in the training data potentially increasing the accuracy of out-of-sample predictions. Therefore a second cross-validation technique consisted in sequentially leaving out each of the ten blocks of health states to create the training data, effectively reducing the data available for model fitting by about 10%. For each of the 10 replications, this procedure removed all information of a subset of respondents. The MAE was not calculated for all health states in the left out block, as the worst state (55555) was included in all blocks and the mildest health state (with misery index 6) was included in at least one other block. Hence, the predictions for these states were not assessed. The final MAE value for the leave-one-block-out cross-validation is the average of the MAE resulting from the 10 replications.

3. **Leave-random-block-out:** the downside of the leave-one-block-out cross-validation is that it can only be replicated 10 times. Therefore, a third cross-validation was carried out, in which new blocks were randomly generated. Using random Latin squares, the original 10 blocks were mixed to create 10 new random pseudo-blocks so that each new pseudo-block consisted of one health state from each of the original blocks, including the worst state and one of the 5 mildest states. This process of creating new blocks was replicated 10 times, leading to 100 pseudo-blocks that can be sequentially left out. As in the leave-one-state-out cross-validation, this technique does not lead to the exclusion of full interviews. The final MAE value for the leave-random-block-out cross-validation is the average of the MAE resulting from the 100 replications.

The predictive accuracy of the estimated regression models was evaluated by summing up the MAE resulting from each of the three different cross-validation techniques and next ranking the models from low to high on the combined MAE.

Theoretical considerations

Some model features may be desirable from a theoretical point of view, but may reduce the model's performance in terms of predictive accuracy or goodness of fit as measured above. We attempt to substantiate the choices based on theoretical considerations with suggestive evidence. These choices relate to the need to correct for heteroscedasticity, the presence of censored data and the use of the hybrid model.

1. **Heteroscedasticity**. A feature found in several valuation studies is that the variability in reported utility values increases with worsening health states. This is indicative of heteroscedasticity in the data and can be taken into account by modelling the variance of error term.

2. **Censored data.** The minimum reported cTTO utility value is capped at -1 by design. It is possible that some respondents would like to value certain health states even lower. If that is the case, the (dis)utility values are censored, i.e. we observe a capped value instead of the real value.

While there are arguments in favor of censoring, and a number of valuation studies have accounted for it, it is impossible to discern from the data whether or not respondents want to value health states lower than they currently do when given the opportunity. Alternative explanations exists. One possibility is that respondents are (nearly) indifferent between health states once a critical level of ill health is reached and hence might value multiple health states at the lowest possible utility value of -1.

The possible censored nature of the data can be taken into account in the estimation. This would deliberately lead to a lower valuation of the more severe health states than observed and reduce the performance of the model in terms of goodness of fit or predictive accuracy. Hence, the choice to correct for censoring is typically a theoretical consideration. Nonetheless, as DCE data are uncensored by design, a comparison of results from DCE-only and cTTO-only models can give some further indication.Hence, the level of between both can provide suggestive evidence on the need for treating the cTTO data as censored. Versteegh et al. [21] propose to explore the "*DCE fit*", i.e. the mean absolute difference between the utility values for all 3125 health states predicted by a DCE-only model and those generated by a cTTO-only model once with and once without censoring. To generate the DCE value set, the rescaling factor θ from the hybrid model with the same core specification was used to make both value set more comparable. The cTTO-only model (censored or uncensored) with the lowest DCE fit is preferred and steers the choice for censoring.

3. **The hybrid model**. The hybrid model has been proposed as a pragmatic compromise to combine stated preferences from cTTO and DCE valuation techniques and increase precision [10, 16, 18, 22]. The hybrid model has been selected in multiple recent valuation studies as preferred model [12, 13, 15-19, 23-30].

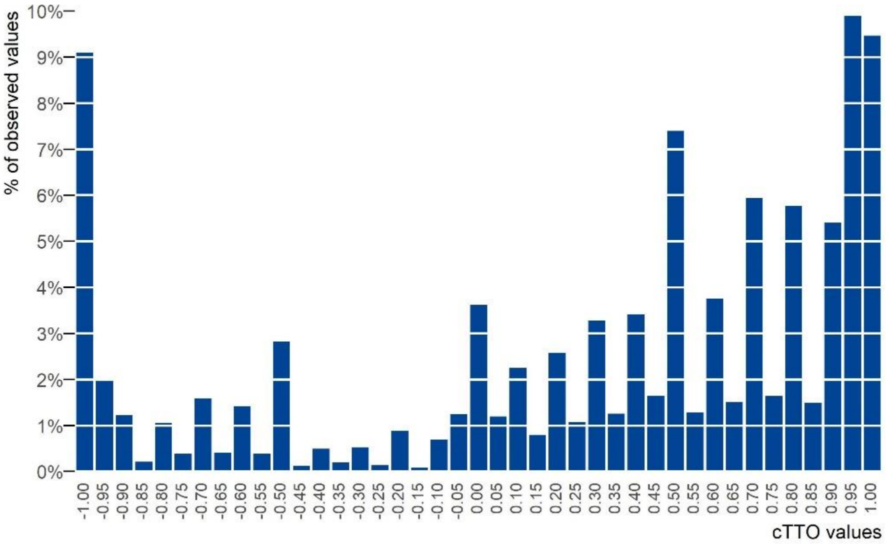
   While the desirability of the hybrid model is difficult to assess, we do argue that it can be of added value when the same value-function underlies the responses in both cTTO and DCE. In case strong agreement between the results from the DCE-only and cTTO-only model was found, this would justify the use of the hybrid model.

ESM5 – Distribution of cTTO values

The distribution of observed cTTO values is presented in Fig. 3. The full range of potential cTTO values was observed in the data. Overall 25.2% of cTTO values were valued worse than dead. About 9.1% of the cTTO tasks resulted in the lowest possible value of -1, indicating that respondents exhausted all available lead time. 161 respondents (18%) valued multiple heath states at -1.

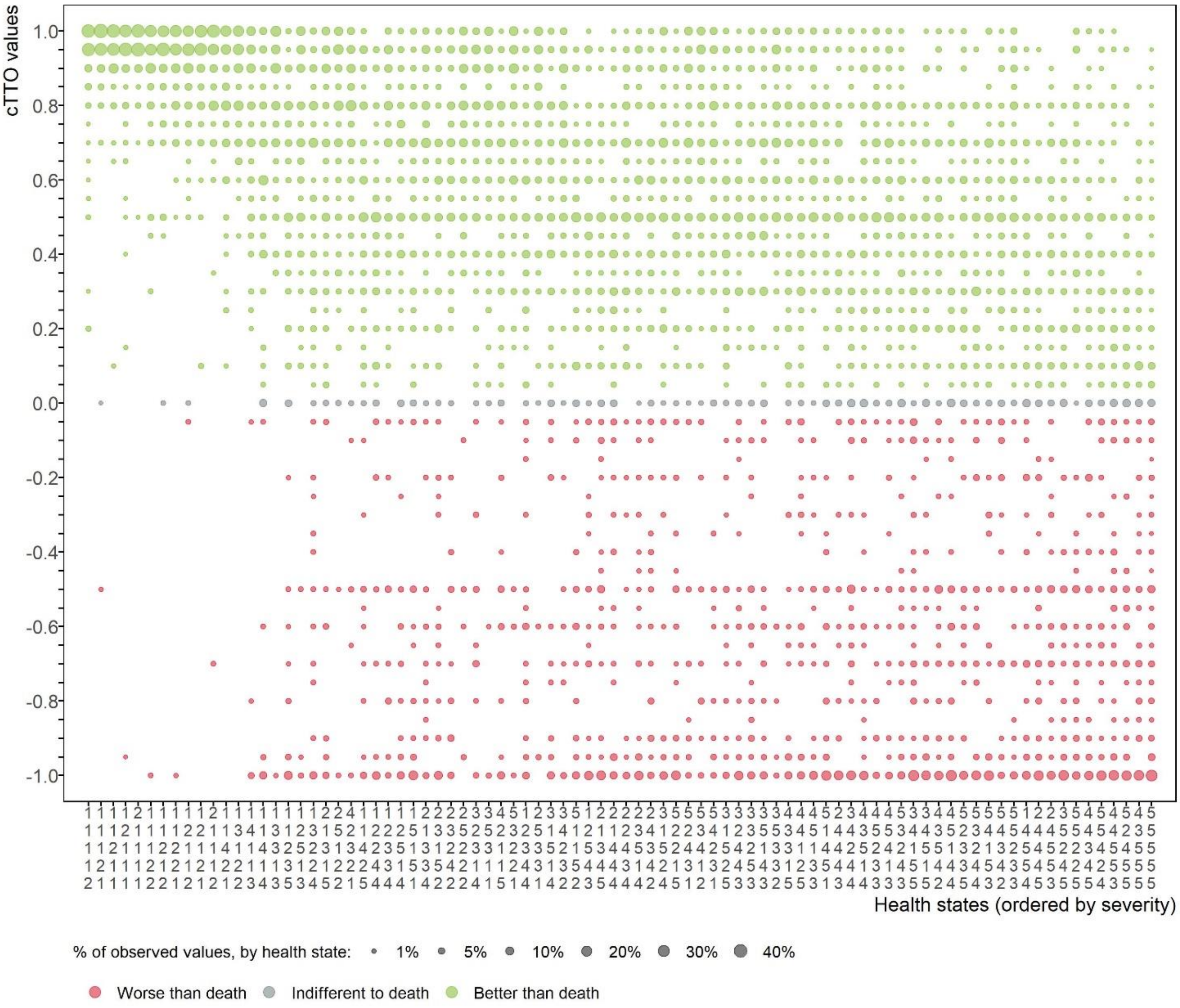
Clustering was also observed for other key values, such as 0.5, 0.95 and 1. About 9.5% and 9.9% of the responses have a cTTO value of 1 and 0.95, respectively, implying no or a very limited willingness to trade-off life-years to avoid health problems. Fig. 3 further indicates digit preference among respondents with a higher proportion of responses for round numbers. Responses at -0.5, 0 and 0.5 represent, respectively, 2.8%, 3.6% and 7.4% of all observations.

Fig. 3 Distribution of observed cTTO values



The results in Fig. 3 mask important variation observed in the valuation of health states. Not only the worst state that was valued at -1 nor did only mild health states receive a value of 1. The individual variation in cTTO values by health states is presented in Fig. 4. A higher percentage of responses for a specific combination is visualized by the size and transparency of the bubble. Positive values are presented in green, negative values in red and a value of 0 in grey. For the 5 mildest health states, the cTTO values are heavily concentrated at 1 or 0.95, with a high level of agreement between the respondents. The divergence in preferences gradually increases and from severity 11 onwards, health states systematically receive negative valuations, while below this threshold, negative values occur sporadically. At severity level 17 or higher, a lower share of high cTTO values is observed (between 0.5 and 1).

Fig. 4 Distribution of observed cTTO values for the 86 health states



Interviewer effects

In spite of a strict interview protocol, quality control process and training of the interviewers, differences in the cTTO valuation between interviewers were observed.

Fig. 5 presents the difference in cTTO valuation at the interviewer level. Each interviewer is represented by a unique symbol while the color of the symbol is related to the number of conducted interviews.

The results indicate that there is particularly important variation in the fraction of responses clustered at 1 and at -1 and in the fraction of responses valued worse than dead (cTTO value < 0). Fig. 3 showed that at the upper end of the distribution, there was a clustering of values at both 0.95 and 1. When taking both values together, the results in Fig. 5 indicate much less variability between interviewers in the fraction of responses having one of these values. The fraction of responses valued at -1 ranges from 0% to 32% across interviewers, while the fraction of negatively valued health states ranges from 4% to 50%. The upper and lower ends of the ranges were not determined by interviewers with lower numbers of conducted interviews.

The fraction of excluded interviews, the fraction of interviews flagged in the feedback modules and the percentage of responses clustered at 0.5, 0 or -0.5 appear to be quite consistent across interviewers.

Summary statistics

In Table 1, summary statistics (mean, median, variance, p10, p25, p75, p90) are provided for all 86 health states of the cTTO valuation as well as the unconscious state. The unconscious state was valued lowest (mean at -0.454), while the health state with slight problems in walking about and no problems in the other dimensions – state 21111 – was valued most closely to full health (mean at 0.958). The 5 mildest states have the highest cTTO values, ranging from 0.923 to 0.958. The mean cTTO value for the worst state (55555) and the unconscious state are nearly identical, but the variability in valuation is lower for the unconscious state. In addition to these two states, 14 other states were negatively valued on average.

Fig. 5 Differences in cTTO valuation by interviewer

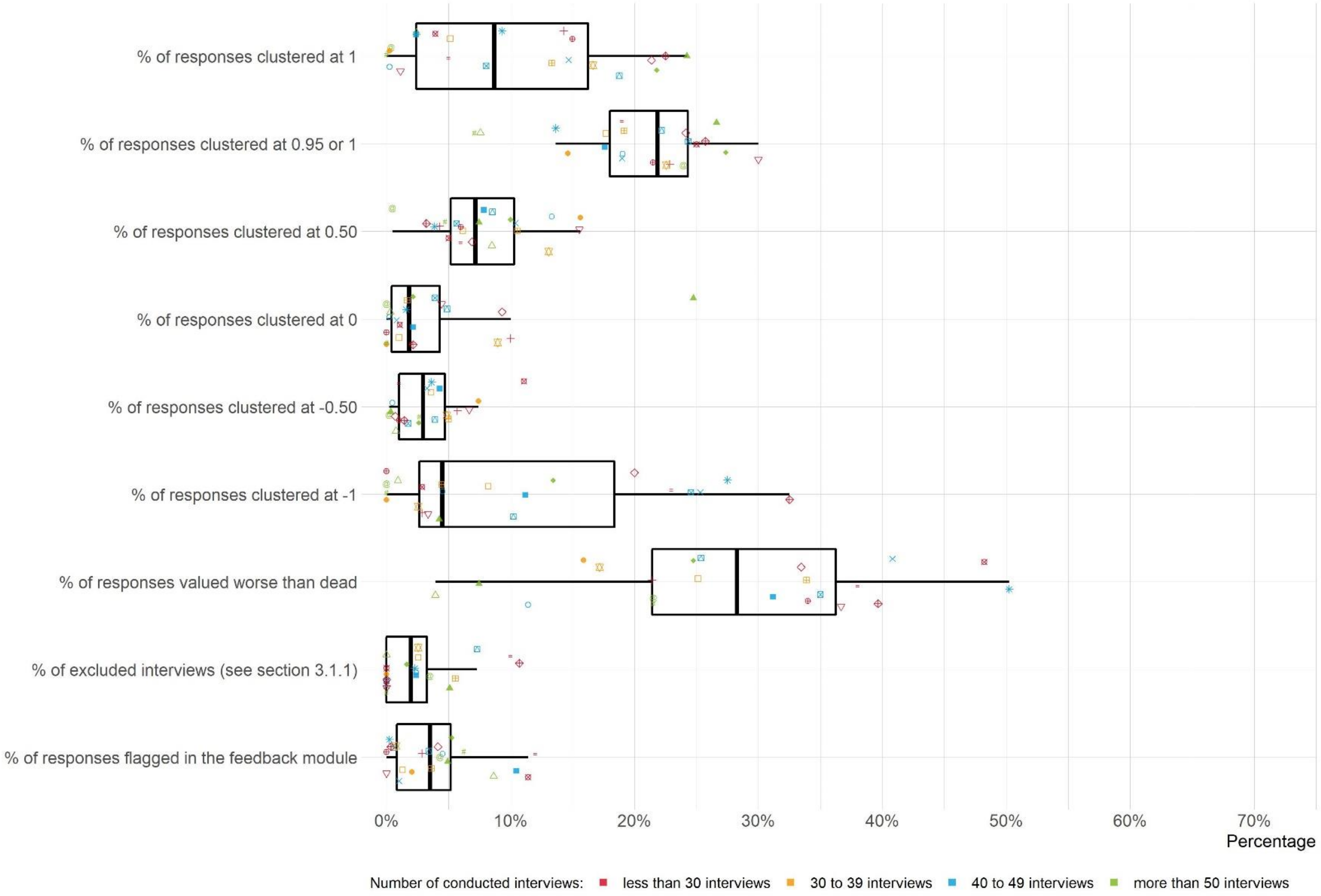


Table 1 Summary statistics for the 86 health states of the cTTO valuation and the unconscious state

Profile	N	Mean (standard error)	P10	P25	Median	P75	P90	Variance
11112	171	0.926 (0.011)	0.85	0.95	0.95	1.00	1.00	0.019
11121	169	0.945 (0.011)	0.90	0.95	0.95	1.00	1.00	0.020
11122	89	0.874 (0.026)	0.70	0.90	0.95	1.00	1.00	0.061
11211	190	0.941 (0.008)	0.90	0.95	0.95	1.00	1.00	0.011
11212	88	0.896 (0.017)	0.75	0.90	0.95	1.00	1.00	0.027
11221	93	0.892 (0.023)	0.75	0.85	0.95	1.00	1.00	0.046
11235	93	0.265 (0.065)	-1.00	0.00	0.50	0.75	0.85	0.385
11414	79	0.428 (0.063)	-0.60	0.30	0.60	0.80	0.95	0.317
11421	92	0.828 (0.020)	0.60	0.75	0.85	0.95	1.00	0.037
11425	94	0.376 (0.055)	-0.55	0.20	0.50	0.80	0.95	0.285
12111	182	0.923 (0.013)	0.80	0.95	0.95	1.00	1.00	0.030
12112	88	0.881 (0.019)	0.75	0.85	0.95	0.95	1.00	0.032
12121	96	0.903 (0.015)	0.75	0.90	0.95	1.00	1.00	0.022
12244	92	0.219 (0.056)	-0.95	0.00	0.35	0.50	0.80	0.291
12334	82	0.479 (0.063)	-0.50	0.40	0.70	0.90	0.95	0.324
12344	90	0.178 (0.062)	-0.80	-0.15	0.35	0.60	0.85	0.336
12513	90	0.651 (0.043)	0.20	0.50	0.70	0.95	1.00	0.159
12514	93	0.278 (0.065)	-0.95	0.00	0.50	0.75	0.90	0.389
12543	96	0.184 (0.064)	-1.00	-0.25	0.40	0.60	0.90	0.395
13122	94	0.869 (0.013)	0.70	0.80	0.90	0.95	1.00	0.016
13224	89	0.376 (0.061)	-0.70	0.20	0.55	0.80	0.90	0.324
13313	92	0.793 (0.027)	0.50	0.70	0.90	0.95	1.00	0.069
14113	89	0.681 (0.045)	0.30	0.55	0.80	0.95	1.00	0.180
14554	90	-0.034 (0.061)	-1.00	-0.60	0.10	0.40	0.65	0.323
15151	89	0.131 (0.073)	-1.00	-0.50	0.40	0.70	0.85	0.464
21111	180	0.958 (0.004)	0.90	0.95	0.95	1.00	1.00	0.004
21112	90	0.888 (0.022)	0.70	0.80	0.95	1.00	1.00	0.040
21315	89	0.446 (0.062)	-0.60	0.25	0.65	0.90	1.00	0.335
21334	82	0.441 (0.064)	-0.70	0.35	0.70	0.80	0.95	0.338
21345	88	-0.002 (0.063)	-1.00	-0.50	0.15	0.50	0.70	0.364
21444	79	0.093 (0.063)	-0.95	-0.10	0.25	0.50	0.70	0.317
31525	92	0.269 (0.057)	-0.70	0.10	0.40	0.60	0.90	0.306
32314	82	0.444 (0.063)	-0.65	0.20	0.65	0.90	0.95	0.328
32443	96	0.183 (0.066)	-1.00	-0.15	0.45	0.65	0.85	0.425
33253	82	0.177 (0.072)	-0.90	-0.50	0.45	0.70	0.90	0.424
34155	96	-0.247 (0.061)	-1.00	-0.90	-0.10	0.30	0.50	0.358
34232	89	0.560 (0.050)	0.00	0.50	0.70	0.85	0.95	0.221
34244	88	-0.061 (0.063)	-1.00	-0.55	0.00	0.40	0.80	0.371
34515	93	0.005 (0.069)	-1.00	-0.70	0.20	0.55	0.70	0.431
35143	79	0.236 (0.066)	-0.90	0.00	0.45	0.65	0.80	0.347
35245	93	-0.151 (0.065)	-1.00	-0.90	0.00	0.40	0.65	0.391
35311	89	0.670 (0.041)	0.25	0.55	0.80	0.90	1.00	0.146
35332	94	0.605 (0.043)	0.30	0.50	0.70	0.90	0.95	0.175
42115	94	0.281 (0.059)	-0.60	0.00	0.45	0.70	0.80	0.322
42321	89	0.670 (0.039)	0.30	0.60	0.80	0.90	0.95	0.134
43315	89	0.257 (0.065)	-0.95	0.05	0.45	0.70	0.90	0.369
43514	88	0.150 (0.067)	-1.00	-0.10	0.30	0.60	0.80	0.411
43542	96	0.056 (0.067)	-1.00	-0.50	0.35	0.55	0.80	0.434
43555	89	-0.217 (0.062)	-1.00	-0.80	0.00	0.30	0.55	0.334
44125	90	0.165 (0.061)	-0.95	-0.10	0.30	0.55	0.80	0.319
44345	90	-0.086 (0.063)	-1.00	-0.60	0.00	0.40	0.55	0.341
44553	88	-0.231 (0.064)	-1.00	-0.90	-0.10	0.20	0.50	0.373
45133	96	0.339 (0.064)	-1.00	0.30	0.60	0.75	0.90	0.405
45144	93	-0.183 (0.066)	-1.00	-0.80	0.00	0.30	0.70	0.394
45233	92	0.383 (0.058)	-0.80	0.25	0.50	0.75	0.90	0.316
45413	94	0.317 (0.057)	-0.70	0.15	0.50	0.70	0.90	0.306
51152	94	0.113 (0.061)	-0.90	-0.40	0.30	0.50	0.70	0.347
51451	93	-0.030 (0.072)	-1.00	-0.80	0.15	0.50	0.80	0.467
52215	96	0.055 (0.066)	-1.00	-0.50	0.30	0.50	0.75	0.424
52335	89	0.121 (0.064)	-1.00	-0.50	0.25	0.60	0.80	0.358
52431	89	0.481 (0.059)	-0.70	0.40	0.70	0.90	1.00	0.309
52455	92	-0.196 (0.058)	-1.00	-0.70	0.00	0.20	0.50	0.321

Profile	N	Mean (standard error)	P10	P25	Median	P75	P90	Variance
22434	94	0.405 (0.052)	-0.30	0.30	0.50	0.70	0.90	0.253
23152	88	0.215 (0.068)	-1.00	-0.20	0.50	0.70	0.90	0.427
23242	82	0.345 (0.067)	-0.90	0.25	0.55	0.75	0.95	0.369
23514	96	0.163 (0.071)	-1.00	-0.50	0.50	0.70	0.90	0.489
24342	82	0.221 (0.066)	-0.80	-0.10	0.45	0.65	0.80	0.356
24443	89	0.164 (0.067)	-1.00	-0.05	0.35	0.60	0.75	0.394
24445	89	-0.215 (0.062)	-1.00	-0.80	0.00	0.25	0.50	0.334
24553	94	0.032 (0.061)	-0.95	-0.50	0.20	0.50	0.70	0.341
25122	92	0.710 (0.034)	0.35	0.60	0.80	0.95	1.00	0.110
25222	79	0.668 (0.044)	0.35	0.55	0.80	0.90	1.00	0.152
25331	79	0.689 (0.041)	0.40	0.50	0.80	0.95	1.00	0.136
31514	79	0.264 (0.065)	-0.90	0.00	0.40	0.65	0.80	0.342
31524	89	0.335 (0.065)	-0.95	0.20	0.50	0.70	0.95	0.367
53221	90	0.690 (0.039)	0.30	0.60	0.80	0.95	1.00	0.128
53243	79	0.170 (0.066)	-0.95	-0.05	0.35	0.60	0.80	0.352
53244	79	-0.042 (0.066)	-1.00	-0.60	0.15	0.35	0.60	0.352
53412	82	0.498 (0.060)	-0.50	0.40	0.65	0.90	0.95	0.295
54153	89	0.031 (0.071)	-1.00	-0.65	0.15	0.60	0.70	0.446
54231	93	0.441 (0.057)	-0.55	0.30	0.60	0.80	0.95	0.298
54342	90	0.183 (0.056)	-0.70	-0.10	0.30	0.50	0.75	0.270
55225	82	0.056 (0.076)	-0.95	-0.65	0.20	0.70	0.95	0.478
55233	92	0.347 (0.062)	-0.70	0.20	0.50	0.80	0.95	0.362
55424	88	-0.060 (0.064)	-1.00	-0.65	0.10	0.40	0.60	0.383
55555	892	-0.453 (0.017)	-1.00	-1.00	-0.55	0.00	0.20	0.270
Unconscious	189	-0.454 (0.035)	-1.00	-1.00	-0.50	0.00	0.05	0.233

ESM6 – Selection process and full regression results

Taking all selection criteria into consideration, the hybrid version of the multiplicative 8-coefficient model with intercept for the cTTO data, with random effects and correction for heteroskedasticity was selected as preferred model.

Selection process

Logical consistency

For all cTTO-only models as well as the hybrid models, the ADD20 specification had an inconsistent ordering of levels 2 and 3 in the usual activities dimension, with disutility of level 2 exceeding disutility of level 3. Also, the DCE-only model with the ADD20 specification shows inconsistencies for level 2 and 3 in the dimensions mobility and usual activities. The high level of flexibility in the estimation of the coefficients in the ADD20 models comes at a cost in terms of consistency. Therefore, the ADD20 models as such are not further considered. Alternatively, an adjusted version of the ADD20 specification enforcing consistency in the usual activities dimension was estimated; in what follows, it is referred to as adjusted-ADD20.

Illogically ordered coefficients were found in almost none of the multiplicative specifications for the cTTO-only, DCE-only and hybrid models estimated on the modelling dataset.

Goodness of fit

All cTTO-only models were evaluated in terms of goodness of fit. Table 2 presents the 15 models with the best performance, ordered by the sum of the ranks in BIC value and MAE value. The following conclusions can be drawn.

1. Multiplicative models perform better in terms of BIC, while additive models perform better in terms of MAE.
2. Models with intercept (12 out of 15 models), random effects (12 out of 15 models) and with correction for heteroscedasticity (10 out of 15 models) rank highest and are thus preferred. The inclusion of random effects leads in particular to improvements in the likelihood value and hence BIC value, implying that the model is a more likely representation of the underlying data. A correction for heteroscedasticity leads to both a better BIC and MAE value. The inclusion of an intercept has only a minor effect on the BIC value, but utility values predicted by a model with intercept are generally in closer agreement with the observed utility values leading to lower MAE value.
3. Models that accounted for censoring do not perform very well with respect to goodness of fit. This was expected as the predicted values of a censored model deliberately deviate from the observed values, under the assumption that some respondents would prefer to value certain health states lower than allowed in the cTTO valuation.

Predictive accuracy

Table 3 presents a summary of the cross-validation results for the 15 cTTO-only models with the lowest sum of MAE of the three different techniques. Models that were also among the best performers in terms of goodness of fit are indicated in bold. The overlap is substantial with 8 out of 15 models. Models for which 5% or more of the cross-validation estimations showed inconsistent orderings are indicated in red in Table 3. The following conclusions can be drawn.

1. The MULT8 specification with intercept, random effects and correction for heteroscedasticity is the best performing model in terms of predictive accuracy, with a high rank in each of the cross-validations. It ranked 6th in terms of goodness of fit.
2. Most models in Table 3 have an intercept. The impact of including random effects is less clear, with 8 out of 15 models having random effects, but 4 of them are ranked in the top 5. Only 5 models in Table 3 correct for heteroscedasticity and none for censoring.
3. The logical consistency of all cross-validation estimations was assessed and revealed a number of inconsistencies. Various alternative versions of the adjusted-ADD20 specification had illogically ordered coefficients, in particular with regard to levels 2 and 3 as well as 4 and 5 in the self-care dimension and level 1 and 2 in the pain/discomfort dimension. It was decided to not further enforce consistency in these dimensions and to also discard the adjusted-ADD20.

Table 2 Evaluation of goodness of fit

	Specification	Options	Sum of ranks	BIC value	BIC rank	MAE value	MAE rank
1	Adjusted-ADD20	Intercept, random effects, correction for heteroscedasticity	11	10 541	7	0.052	4
2	MULT11	Intercept, random effects, correction for heteroscedasticity	15	10 513	5	0.055	10
3	Adjusted-ADD20	No intercept, random effects, correction for heteroscedasticity	15	10 542	8	0.054	7
4	MULT9	Intercept, random effects, correction for heteroscedasticity	21	10 497	2	0.056	19
5	Adjusted-ADD20	Intercept, random effects	24	10 657	15	0.054	9
6	MULT8	Intercept, random effects, correction for heteroscedasticity	24	10 492	1	0.056	23
7	Adjusted-ADD20	No intercept, random effects	29	10 658	16	0.055	13
8	Adjusted-ADD20	Intercept, correction for heteroscedasticity	30	12 354	28	0.051	2
9	MULT11	Intercept, correction for heteroscedasticity	32	12 321	21	0.055	11
10	MULT8	Intercept, correction for heteroscedasticity	34	12 308	18	0.055	16
11	MULT11	Intercept, random effects	38	10 638	13	0.057	25
12	Adjusted-ADD20	No intercept, random effects, correction for heteroscedasticity and censoring	39	12 390	31	0.054	8
13	Adjusted-ADD20	Intercept, random effects, correction for heteroscedasticity and censoring	43	12 390	29	0.054	14
14	MULT9	Intercept, random effects	45	10 631	10	0.057	35
15	MULT8	Intercept, random effects	46	10 626	9	0.058	37

Note: MAE = Mean Absolute Error, BIC = Bayesian Information Criterion

Table 3 Evaluation of predictive accuracy

	Specification	Options	Sum of MAE	MAE state out	rank state out	MAE block out	Rank block out	MAE random blocks	Rank random blocks
1	**MULT8**	**Intercept, random effects, correction for heteroscedasticity**	**0.192**	**0.062**	**2**	**0.066**	**1**	**0.064**	**4**
2	MULT8	Intercept, random effects	0.193	0.063	5	0.066	3	0.064	6
3	MULT8	Intercept	0.194	0.064	11	0.066	4	0.063	3
4	**Adjusted-ADD20**	**No intercept, random effects**	**0.194**	**0.063**	**4**	**0.068**	**8**	**0.063**	**1**
5	**MULT9**	**Intercept, random effects**	**0.195**	**0.064**	**8**	**0.067**	**5**	**0.064**	**7**
6	**MULT9**	**Intercept, random effects, correction for heteroscedasticity**	**0.195**	**0.062**	**1**	**0.068**	**11**	**0.065**	**12**
7	MULT9	Intercept	0.196	0.065	15	0.067	6	0.064	5
8	Adjusted-ADD20	Intercept, random effects	0.197	0.065	13	0.068	9	0.064	8
9	Adjusted-ADD20	No intercept	0.197	0.066	19	0.068	12	0.063	2
10	**MULT11**	**Intercept, random effects**	**0.197**	**0.064**	**12**	**0.069**	**14**	**0.065**	**11**
11	**MULT11**	**Intercept, correction for heteroscedasticity**	**0.198**	**0.064**	**10**	**0.066**	**2**	**0.068**	**35**
12	MULT11	Intercept	0.198	0.066	18	0.068	13	0.064	10
13	**MULT8**	**Intercept, correction for heteroscedasticity**	**0.198**	**0.064**	**7**	**0.070**	**17**	**0.065**	**16**
14	Adjusted-ADD20	Intercept	0.199	0.066	27	0.068	10	0.064	9
15	**Adjusted-ADD20**	**Intercept, random effects, correction for heteroscedasticity**	**0.200**	**0.066**	**24**	**0.067**	**7**	**0.066**	**22**

Theoretical considerations

Additional considerations were used to substantiate the choice to (not) correct for heteroscedasticity, to (not) correct for censoring and to (not) use the hybrid model.

- **Heteroscedasticity**

A visual inspection of the variability in the error terms by the predicted values of the estimated models clearly showed that the error term is heteroscedastic. An observation that was corroborated by two formal tests of heteroscedasticity, the White's test and the modified Breusch-Pagan test [31]. Both tests were performed for the cTTO-only models in each of the four specifications – ADD20, MULT8, MULT9 and MULT11 – using an intercept, but without random effects or censoring and confirm the presence of heteroscedasticity ($p<0.001$).

Hence, irrespective of the performance in terms of goodness of fit or predictive accuracy, there is a clear argument to favour a model that corrects for heteroscedasticity.

- **Censoring**

DCE data are uncensored by design. Under the assumption that responses in the cTTO and DCE task were driven by the same underlying preferences, an assessment of the agreement between the predictions of the DCE-only model and the cTTO-only models with and without correction for censoring, provides suggestive evidence on the need of treating the cTTO data as censored.

Table 4 presents the results of such assessment for the 2 best performing cTTO-only models in term of predictive accuracy that account for heteroscedasticity and have a strong goodness of fit. It concerns the MULT8 and MULT9 model with intercept, random effects and correction for heteroscedasticity, ranked, respectively, 1st and 5th for predictive accuracy and 6th and 4th on goodness of fit. Other specifications have also been evaluated (not presented here), leading to the same conclusions.

The DCE fit in Table 4, which was calculated as the mean absolute difference between the utility values of the health states predicted by the DCE-only model and the cTTO-only models, indicates that predicted values from uncensored cTTO-only models are in better agreement with the predicted values from the DCE-only models than the predicted values from censored cTTO-only models. The same conclusion holds when comparing the DCE fit of the censored and uncensored hybrid models. This provides suggestive evidence that treating cTTO data as censored does not improve the estimation of the underlying preferences and that an uncensored model is preferable.

- **Hybrid model**

Fig. 6 demonstrate the high level of correspondence between the predicted values from the uncensored cTTO-only and DCE-only MULT8 and MULT9 models. The point estimates of the coefficients of the corresponding cTTO-only, DCE-only and hybrid models are given in Table 5. The coefficients of the DCE-only models were rescaled using the rescaling factor of the hybrid model.

The preference ranking of the dimensions is consistent between the DCE-only and cTTO-only model. Disutility is lowest and not significantly different for the dimensions self-care and usual activities, mobility has a slightly higher level of disutility, the disutility associated with anxiety and depression is substantially higher and the most important source of disutility is the dimension pain and discomfort. The results from the MULT8 and MULT9 model are highly similar.

The level coefficients are slightly different. In the DCE-only model, there is a higher utility decrement going from level 1 to 2 and from level 4 to 5, and a lower decrement moving from level 2 to 3 and from level 3 to 4, compared to the results from the cTTO-only model. In both models, the transition from level 3 to 4 is accompanied with the sharpest drop in utility.

The high level of agreement between the cTTO-only and DCE-only results, in particular in the preference ranking of the dimensions, warrants the use of a hybrid model in which the coefficients can be estimated with higher precision.

Table 4 Evaluation DCE fit

	MULT8 – intercept, random effects, correction for heteroscedasticity		MULT9 – intercept, random effects, correction for heteroscedasticity	
	Not censored cTTO-only	Censored cTTO-only	Not censored cTTO-only	Censored cTTO-only
DCE fit	0.040	0.045	0.041	0.045
Negatively valued states (out of 3 125 states)	498 (15.9%)	584 (18.7%)	505 (16.1%)	596 (19.1%)

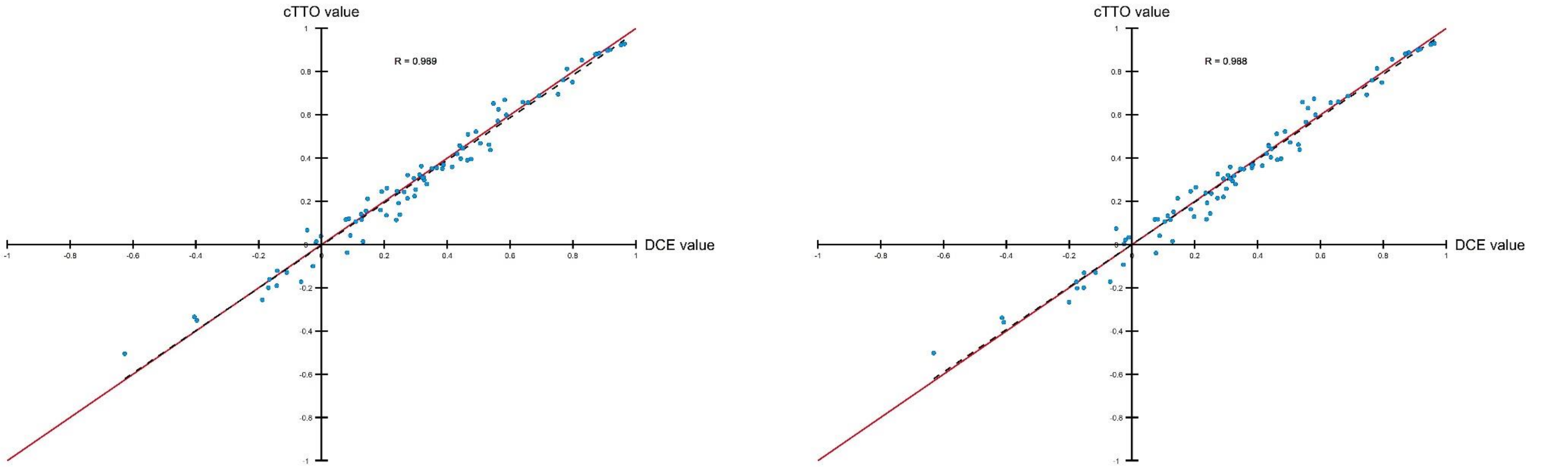
Table 5 Coefficients of cTTO-only, DCE-only and hybrid models of MULT8 and MULT9 with intercept, random effects and correction for heteroscedasticity

MULT8 – intercept, random effects, correction for heteroscedasticity				MULT9 – intercept, random effects, correction for heteroscedasticity			
Coefficients	cTTO-only	DCE-only	Hybrid	Coefficients	cTTO-only	DCE-only	Hybrid
$Intercept$	0.055		0.038	$Intercept$	0.054		0.037
β_{MO}	0.214	0.265	0.227	β_{MO}	0.210	0.264	0.222
β_{SC}	0.162	0.199	0.166	β_{SC}	0.161	0.200	0.169
β_{UA}	0.160	0.200	0.181	β_{UA}	0.157	0.200	0.179
β_{PD}	0.476	0.507	0.482	β_{PD}	0.455	0.491	0.456
β_{AD}	0.438	0.456	0.439	β_{AD}	0.416	0.443	0.415
L_2	0.100	0.178	0.139	L_2	0.102	0.184	0.144
L_3	0.254	0.236	0.258	L_3	0.268	0.242	0.266
L_4	0.847	0.706	0.788	L_4	0.885	0.726	0.824
				L_5	1.056	1.037	1.061

MULT9 not statistically different from MULT8

The distinguishing factor between the MULT8 and MULT9 specification is the L_5 coefficient relaxing the proportionality assumption between the levels in all dimensions by allowing the level 5 coefficient in the dimensions pain/discomfort and anxiety/depression to differ from the level 5 coefficient in the dimensions mobility, self-care and usual activities. The level 5 coefficient for the latter three dimensions is standardized at 1. In other words, the MULT8 specification can be seen as a special case of the MULT9 model with L_5 constrained to 1. As can be seen in Table 5, the L_5 coefficient is, only marginally different from 1. A likelihood ratio test on the cTTO-only, DCE-only and hybrid models does not reject the null hypothesis (at p<0.05) that the constrained MULT8 model provides as good a fit for the data as the more flexible MULT9 model. In addition, a Wald test indicates that the L_5 coefficient is not statistically different from 1 in the cTTO-only, DCE-only and hybrid model (at p<0.05).

Fig. 6 Agreement between DCE predicted values and cTTO predicted values for the 86 health states used in the valuation



A. **MULT8 specification – intercept, random effects, correction for heteroscedasticity, not censored**

B. **MULT9 specification – intercept, random effects, correction for heteroscedasticity, not censored**

Note: The diagonal is presented by the red line and indicates an equal value for predictions from DCE-only and cTTO-only models. The dashed black line is the linear regression line between the values from the cTTO-only and DCE-only values. The R-value is the correlation between both values.

Full regression results

Table 6 Coefficients and bootstrapped standard errors of the preferred model

	Preferred model			
	Coefficient value	Standard error	T-statistic	P value
Model				
$\boldsymbol{Intercept}$ **(cTTO part only)**	0.038	0.0148	2.55	0.0054
$\boldsymbol{\beta_{MO}}$	0.227	0.0102	22.19	<0.0001
$\boldsymbol{\beta_{SC}}$	0.166	0.0108	15.32	<0.0001
$\boldsymbol{\beta_{UA}}$	0.181	0.0098	18.37	<0.0001
$\boldsymbol{\beta_{PD}}$	0.482	0.0138	34.84	<0.0001
$\boldsymbol{\beta_{AD}}$	0.439	0.0132	33.13	<0.0001
$\boldsymbol{L_2}$	0.139	0.0154	9.02	<0.0001
$\boldsymbol{L_3}$	0.258	0.0158	16.34	<0.0001
$\boldsymbol{L_4}$	0.788	0.0157	50.26	<0.0001
Heteroscedasticity ($\boldsymbol{\ln \sigma_{res}}$)				
$\boldsymbol{Intercept}$	-1.158	0.0665	-17.41	<0.0001
$\boldsymbol{\beta_{MO}}$	0.095	0.0340	2.78	0.0027
$\boldsymbol{\beta_{SC}}$	0.040	0.0324	1.25	0.1064
$\boldsymbol{\beta_{UA}}$	0.027	0.0301	0.90	0.1833
$\boldsymbol{\beta_{PD}}$	0.123	0.0346	3.54	0.0002
$\boldsymbol{\beta_{AD}}$	0.148	0.0489	3.03	0.0013
$\boldsymbol{L_2}$	-0.204	0.2190	-0.93	0.1764
$\boldsymbol{L_3}$	0.695	0.2439	2.85	0.0022
$\boldsymbol{L_4}$	1.229	0.3153	3.90	<0.0001
Rescaling factor				
$\boldsymbol{\ln(\theta)}$	-0.727	0.0209	-34.83	<0.0001

Notes: MO = Mobility, SC = Self-care, UA = Usual activities, PD = Pain/discomfort, AD = Anxiety/depression; L_x = Severity level x; cTTO = composite time trade-off. Standard errors of the coefficients were derived using the Rao-Wu bootstrap with 1000 replications [1].

Comparison with competing models

Table 7 presents a summary of key characteristics of the value set and of competing models that ranked high throughout the selection process, showing little divergence between the value sets. The highest impact relates to the choice of a hybrid model, i.e. changed preference ranking of the dimensions self-care and usual activities, and censoring, i.e. a higher fraction of negative values in censored models.

Table 7 Comparison of key characteristics of the value sets produced by different high-performing models

	Preferred model	cTTO-only MULT8, intercept, random effects, correction for heteroskedasticity	cTTO-only MULT8, intercept, random effects, correction for censoring and heteroskedasticity	hybrid MULT8, intercept, random effects, correction for censoring and heteroskedasticity	hybrid MULT9, intercept, random effects, correction for heteroskedasticity	hybrid ADD20, enforced logical consistency, intercept, random effects, correction for heteroskedasticity
% health states valued worse than dead	15.0%	15.9%	18.7%	18.0%	15.2%	15.3%
Preference ranking of dimensions (ordered from highest to lowest utility loss at level 5)	PD AD MO UA SC	PD AD MO SC UA	PD AD MO SC UA	PD AD MO UA SC	PD AD MO UA SC	PD AD MO UA SC
Minimum value (state 55555)	-0.532	-0.505	-0.606	-0.630	-0.531	-0.526
Maximum value (except full health)	0.939 (state 12111)	0.929 (state 11211)	0.935 (state 11211)	0.950 (state 12111)	0.939 (state 12111)	0.954 (states 11211, 11311)

Notes: MULT8 = multiplicative 8-coefficients model; MULT9 = multiplicative 9-coefficientys model; ADD20 = additive 20-coefficients model; cTTO = composite time trade-off; MO = Mobility, SC = Self-care, UA = Usual activities, PD = Pain/discomfort, AD = Anxiety/depression.

ESM 7 – Variable definitions Belgian EQ-5D-5L value set

The Belgian EQ-5D-5L value set can be found in CSV format in a separate attachment. Information on the variable definitions is given in the Table below.

Table 8 Variable definition value set

Variable name	Type	Definition
state	Numeric	EQ-5D-5L code of the health state. The code consists of 5 digits, where each digit represents the severity level of a dimension. By convention, the order of dimensions is mobility, self-care, usual activities, pain/discomfort and anxiety/depression.
state_string	Character	EQ-5D-5L code of the health state as string (see above) and additionally the label "unconscious" for the state of unconsciousness.
value	Numeric	Utility value related to the EQ-5D-5L health state. Utility values are expressed on a scale where 0 is the value for 'dead' and 1 is the value for 'full health'. Negative values are possible for health states considered worse than dead.
mo	Numeric	The severity level of dimension mobility.
sc	Numeric	The severity level of dimension self-care.
ua	Numeric	The severity level of dimension usual activities.
pd	Numeric	The severity level of dimension pain/discomfort.
ad	Numeric	The severity level of dimension anxiety/depression.

References

[1] Heeringa SG, West BT, Berglund PA. Applied survey data analysis (2nd ed.). Boca Raton: Chapman and Hall/CRC 2017. https://doi.org/10.1201/9781315153278

[2] Demarest S, Van der Heyden J, Charafeddine R, Drieskens S, Gisle L, Tafforeau J. Methodological basics and evolution of the Belgian health interview survey 1997–2008. Archives of Public Health. 2013; 71(1):24. https://doi.org/10.1186/0778-7367-71-24

[3] Statistics Belgium. Population by place of residence, nationality, marital status, age and sex. 2017 [Geraadpleegd 1/03/2021]; Beschikbaar via: https://statbel.fgov.be/en/open-data/population-place-residence-nationality-marital-status-age-and-sex-0

[4] Rand-Hendriksen K, Ramos-Goñi JM, Augestad LA, Luo N. Less is more: cross-validation testing of simplified nonlinear regression model specifications for EQ-5D-5L health state values. Value in Health. 2017; 20(7):945-52. https://doi.org/10.1016/j.jval.2017.03.013

[5] Luo N, Liu G, Li M, Guan H, Jin X, Rand-Hendriksen K. Estimating an EQ-5D-5L value set for China. Value in Health. 2017; 20(4):662-9. https://doi.org/10.1016/j.jval.2016.11.016

[6] Feng Y, Devlin N, Shah K, Mulhern B, van Hout B. New methods for modelling EQ-5D-5L value sets: an application to English data. London: Office of Health Economics; 2016.

[7] Oppe M, Ramos-Goñi JM, van Hout B. Modeling EQ-5D-5L valuation data. In: Busschbach JJ, editor. 29th Scientific Plenary Meeting of the EuroQol Group; 2012 September 13-15 2012; Rotterdam; 2012. p. 61-91.

[8] Soekhai V, de Bekker-Grob EW, Ellis AR, Vass CM. Discrete choice experiments in health economics: past, present and future. PharmacoEconomics. 2019; 37(2):201-26. https://doi.org/10.1007/s40273-018-0734-2

[9] OECD. Discrete choice experiments. *Cost-benefit analysis and the environment: further developments and policy use*. Paris: OECD Publishing 2018.

[10] Ramos-Goñi JM, Pinto-Prades JL, Oppe M, Cabases JM, Serrano-Aguilar P, Rivero-Arias O. Valuation and modeling of EQ-5D-5L health states using a hybrid approach. Medical Care. 2017; 55(7):e51-e8. https://doi.org/10.1097/MLR.0000000000000283

[11] Dolan P. Modeling Valuations for EuroQol Health States. Medical Care. 1997; 35(11):1095-108.

[12] 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 economics. 2018; 27(1):7-22. https://doi.org/10.1002/hec.3564

[13] Shafie AA, Vasan Thakumar A, Lim CJ, Luo N, Rand-Hendriksen K, Md Yusof FA. EQ-5D-5L valuation for the Malaysian population. PharmacoEconomics. 2019; 37(5):715-25. https://doi.org/10.1007/s40273-018-0758-7

[14] Ramos-Goñi JM, Craig B, Oppe M, van Hout B. Combining continuous and dichotomous responses in a hybrid model. Rotterdam: EuroQol Research Foundation; 2016.

[15] Andrade LF, Ludwig K, Ramos-Goñi JM, Oppe M, de Pouvourville G. A French value set for the EQ-5D-5L. PharmacoEconomics. 2020; 38(4):413-25. https://doi.org/10.1007/s40273-019-00876-4

[16] Ludwig K, Graf von der Schulenburg JM, Greiner W. German value set for the EQ-5D-5L. PharmacoEconomics. 2018; 36(6):663-74. https://doi.org/10.1007/s40273-018-0615-8

[17] Ramos-Goñi JM, Craig BM, Oppe M, Ramallo-Fariña Y, Pinto-Prades JL, Luo N, et al. Handling data quality Issues to estimate the Spanish EQ-5D-5L value set using a hybrid interval regression approach. Value in Health. 2018; 21(5):596-604. https://doi.org/10.1016/j.jval.2017.10.023

[18] Purba FD, Hunfeld JAM, Iskandarsyah A, Fitriana TS, Sadarjoen SS, Ramos-Goñi JM, et al. The Indonesian EQ-5D-5L value set. PharmacoEconomics. 2017. https://doi.org/10.1007/s40273-017-0538-9

[19] Hobbins A, Barry L, Kelleher D, Shah K, Devlin N, Ramos-Goñi JM, et al. Utility values for health states in Ireland: a value set for the EQ-5D-5L. PharmacoEconomics. 2018; 36(11):1345-53. https://doi.org/10.1007/s40273-018-0690-x

[20] Van de Voorde C, Van den Heede K, Beguin C, Bouckaert N, Camberlin C, de Bekker P, et al. Required hospital capacity in 2025 and criteria for rationalisation of complex cancer surgery, radiotherapy and maternity services. Health Services Research (HSR). Brussel: Belgian Health Care Knowledge Centre (KCE); 2017. Report No.: 289.

[21] Versteegh MM, Vermeulen KM, Evers SMAA, de Wit GA, Prenger R, Stolk EA. Dutch tariff for the five-level version of EQ-5D. Value in Health. 2016; 19(4):343-52. https://doi.org/10.1016/j.jval.2016.01.003

[22] Pickard AS, Law EH, Jiang R, Pullenayegum E, Shaw JW, Xie F, et al. United States valuation of EQ-5D-5L health states using an international protocol. Value in Health. 2019; 22(8):931-41. https://doi.org/10.1016/j.jval.2019.02.009

[23] Shiroiwa T, Ikeda S, Noto S, Igarashi A, Fukuda T, Saito S, et al. Comparison of value set based on DCE and/or TTO data: scoring for EQ-5D-5L health states in Japan. Value in Health. 2016; 19(5):648-54. https://doi.org/10.1016/j.jval.2016.03.1834

[24] Ferreira PL, Antunes P, Ferreira LN, Pereira LN, Ramos-Goñi JM. A hybrid modelling approach for eliciting health state preferences: the Portuguese EQ-5D-5L value set. Quality of Life Research. 2019; 28(12):3163-75. https://doi.org/10.1007/s11136-019-02226-5

[25] Lin H-W, Li C-I, Lin F-J, Chang J-Y, Gau C-S, Luo N, et al. Valuation of the EQ-5D-5L in Taiwan. PLOS ONE. 2018; 13(12):e0209344. https://doi.org/10.1371/journal.pone.0209344

[26] Golicki D, Jakubczyk M, Graczyk K, Niewada M. Valuation of EQ-5D-5L health states in Poland: the first EQ-VT-based study in central and eastern Europe. PharmacoEconomics. 2019; 37(9):1165-76. https://doi.org/10.1007/s40273-019-00811-7

[27] Welie AG, Gebretekle GB, Stolk E, Mukuria C, Krahn MD, Enquoselassie F, et al. Valuing health state: an EQ-5D-5L value set for Ethiopians. Value in Health Regional Issues. 2020; 22:7-14. https://doi.org/10.1016/j.vhri.2019.08.475

[28] Pattanaphesaj J, Thavorncharoensap M, Ramos-Goñi JM, Tongsiri S, Ingsrisawang L, Teerawattananon Y. The EQ-5D-5L valuation study in Thailand. Expert Review of Pharmacoeconomics & Outcomes Research. 2018; 18(5):551-8. https://doi.org/10.1080/14737167.2018.1494574

[29] Mai VQ, Sun S, Minh HV, Luo N, Giang KB, Lindholm L, et al. An EQ-5D-5L Value Set for Vietnam. Quality of Life Research. 2020; 29(7):1923-33. https://doi.org/10.1007/s11136-020-02469-7

[30] Jensen CE, Sørensen SS, Gudex C, Jensen MB, Pedersen KM, Ehlers LH. The Danish EQ-5D-5L Value Set: A Hybrid Model Using cTTO and DCE Data. Applied Health Economics and Health Policy. 2021; 19(4):579-91. https://doi.org/10.1007/s40258-021-00639-3

[31] SAS Institute Inc. SAS/STAT®14.1 user’s guide: the MODEL procedure. Cary, NC: SAS Institute Inc. 2015.