

Sample Size and Model Prediction Accuracy in EQ-5D-5L Valuations Studies: Expected Out-of-Sample Accuracy Based on Resampling with Different Sample Sizes and Alternative Model Specifications

MDM Policy & Practice

2022, Vol. 7(1) 1–12

© The Author(s) 2022

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/23814683221083839

journals.sagepub.com/home/mppTonya Moen Hansen , Knut Stavem, and Kim Rand

Abstract

Background. National valuation studies are costly, with ~1000 face-to-face interviews recommended, and some countries may deem such studies infeasible. Building on previous studies exploring sample size, we determined the effect of sample size and alternative model specifications on prediction accuracy of modeled coefficients in EQ-5D-5L value set generating regression analyses. **Methods.** Data sets ($n = 50$ to ~1000) were simulated from 3 valuation studies, resampled at the respondent level and randomly drawn 1000 times with replacement. We estimated utilities for each subsample with leave-one-out at the block level using regression models (8 or 20 parameter; with or without a random intercept; time tradeoff [TTO] data only or TTO + discrete choice experiment [DCE] data). Prediction accuracy, root mean square error (RMSE), was calculated by comparing to censored mean predicted values to the left-out block in the full data set. Linear regression was used to estimate the relative effect of changes in sample size and each model specification. **Results.** Results showed that doubling the sample size decreased RMSE by on average 0.012. Effects of other model specifications were smaller but can when combined compensate for loss in prediction accuracy from a small sample size. For models using TTO data only, 8-parameter models clearly outperformed 20-parameter models. Adding a random intercept, or including DCE responses, also improved mean RMSE, most prominently for variants of the 20-parameter models. **Conclusions.** The prediction accuracy impact of further increases in sample size after 300 to 500 were smaller than the impact of combining alternative modeling choices. Hybrid modeling, use of constrained models, and inclusion of random intercepts all substantially improve the expected prediction accuracy. Beyond a minimum of 300 to 500 respondents, the sample size may be better informed by other considerations, such as legitimacy and representativeness, than by the technical prediction accuracy achievable.

Highlights

- Increases in sample size beyond a minimum in the range of 300 to 500 respondents provide smaller gains in expected prediction accuracy than alternative modeling approaches.
- Constrained, nonlinear models; time tradeoff + discrete choice experiment hybrid modeling; and including a random intercept all improved the prediction accuracy of models estimating values for the EQ-5D-5L based on data from 3 different valuation studies.
- The tested modeling choices can compensate for smaller sample sizes.

Corresponding Author

Tonya Moen Hansen, Division for Health Services, Norwegian Institute of Public Health, Postboks 222 Skøyen, Oslo, 0213, Norway; (tonyamoen.hansen@fhi.no).



Keywords

cross validation, EQ-5D, model misspecification, sample size, valuation study, regression models

Date received: October 14, 2021; accepted: February 8, 2022

Introduction

Health state values are estimated for instruments measuring health-related quality of life and can be used to measure and compare the utility of different health outcomes across patient groups and interventions. The EQ-5D-5L measures health using 5 dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), and with 5 severity levels per dimension, can define 3125 unique health states. Valuation studies for the EQ-5D-5L are guided by protocols developed by the EuroQol Group, the scientific nonprofit foundation that owns the EQ-5D. The EQ-VT protocol, developed as part of a research program to improve comparability and data quality in EQ-5D valuation studies,¹ provides standard procedures and requirements for EQ-5D valuation studies.² Following the EQ-VT protocol, time tradeoff (TTO) and discrete choice experiment (DCE) tasks are administered in computer-assisted one-to-one face-to-face interviews with a minimum of 1000 respondents.^{3,4}

A Norwegian valuation study for health states defined by the EQ-5D-5L was initiated in 2019, with data collection planned from the end of 2019 to mid-2020.⁵ Data collection was stopped roughly mid-way, in March 2020, due to social distancing measures following the COVID-19 pandemic, at which point 542 interviews had been completed. Faced with a considerably smaller sample size than the recommended 1000, the motivation for this study was to assess whether the data already collected

could be considered sufficient for estimating values for the Norwegian general population.

Even without the complication of an ongoing pandemic, completing 1000 face-to-face interviews can be challenging and costly, and many countries may deem such studies as infeasible. Interviewers must be trained extensively and followed up closely throughout data collection, as the data are prone to interviewer effects.⁶ Each interview takes considerable time, with an expected total interview time of 58 to 71 min.⁷ To represent the general population, studies often include some geographic aspect to their sampling strategy,^{8–10} as did the Norwegian study, which can require considerable travel.

Some cost-saving methods have been considered, such as increasing the number of direct valuations per respondent or increasing power by including data from other studies.^{11,12} In early EQ-5D valuation studies, the sample size varied significantly,¹³ and recommendations for sample size of 1000 in the EQ-VT protocol has been claimed to be “based on some assumption without support from empirical data and provided limited theoretical justification,” and sample size estimations suggest that far fewer than 1000 could be sufficient for estimating EQ-5D-3L values.¹⁴ Another simulation study showed stable estimates for EQ-5D-5L health state values using a VAS model with sample sizes greater than 500, given that 80 to 120 health states were directly valued.¹⁵

Most EQ-5D-5L value sets have used 20-parameter additive models based on TTO responses only to estimate values, but value sets can be modeled in different ways, for example, by including both TTO and DCE responses in a hybrid model. One may also include random effects, latent classes, allow for heteroscedasticity, or use different functional forms, which may influence the required sample size.^{12,13,16,17}

This study aims to provide empirically based background for discussion of sample size and modeling choices by comparing effects of increases in sample size and alternative modeling specifications on estimates of expected prediction accuracy based on data from published EQ-5D-5L valuation studies and data collected so far for the Norwegian valuation study.

Division for Health Services, Norwegian Institute of Public Health, Oslo, Norway (TMH); Health Services Research Unit, Akershus University Hospital, Norway (KS, KR); Institute of Clinical Medicine, University of Oslo, Norway (KS, KR); Department of Pulmonary Medicine, Medical Division, Akershus University Hospital, Norway (KS); Maths in Health B.V., Rotterdam, the Netherlands (KR). The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Stavem and Rand are members of the EuroQol Group, and Rand is the chairman of the group. The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Financial support for this study was provided entirely by a grant from the Norwegian Research Council (project No. 262673). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

Methods

In addition to the Norwegian data collected so far, we included data from the Netherlands⁹ and the United States⁷ EQ-5D-5L valuation studies. From each study, only respondents used in the final model and value set were included. All 3 studies valued the EQ-5D-5L using the EQ-VT protocol and had similar, although not identical, sampling strategies. Respondents in all 3 studies were administered the composite TTO (cTTO), a variant of the TTO using lead time for the valuation of health states considered worse than being dead, and DCE. The 86 directly valued health states were divided over 10 possible TTO blocks, and 28 DCE blocks, all designed to have a similar composition in health state severity. The same 86 unique health states, and 100 block/state combinations, were included in the block design of all 3 studies. Respondents were randomly assigned 1 block of each to value. Each block, TTO and DCE, respectively, included 10 TTO tasks and 7 DCE tasks.

Samples

The Norwegian (NO) study¹⁸ started data collection November 2019 and had completed 542 interviews by March 2020 when data collection was suspended due to restrictions imposed in response to the COVID-19 pandemic. Six geographic areas within 4 main regions of Norway were randomly selected with probability proportionate to the number of residents in each area. Quotas per area mirrored the proportion of residents per region and area. Stratified random sampling of locations from selected location types and use of quotas were applied to ensure representativeness according to age, sex, and education as well as increasing the number of respondents who may typically be hard to reach, for example, those in poor health, those with young children, and those with other ethnic backgrounds. Interviews were conducted at each location with computer-assisted personal interviews in accordance with EQ-VT version 2.1, which in addition to protocol defined in version 1.0 included TTO practice states during the introduction, quality control monitoring, a feedback module, and a dynamic question after the wheelchair example.⁴

The Netherlands (NL) study⁹ included TTO valuations from 979 respondents. A stratified sampling approach, with quotas for age, sex, and education based on the distribution of the general population in the Netherlands, was used. Respondents were randomly drawn from commercial panels until all quotas were met. Data collection was completed in selected cities, achieving geographical spread across the country. Data were collected in autumn 2012, with computer-assisted personal interviews in compliance with EQ-VT version 1.0.⁴

The United States (US) study⁷ included TTO valuations from 1062 respondents. Three different recruitment strategies were used: use of a web-based recruitment tool to contact potential respondents in recruitment areas, promotion of the study at student chapters near recruitment sites and community platforms, and on-site recruitment during data collection. Six metropolitan areas were selected, ensuring representativeness and sampling in all census regions. Data were collected between May and September 2017, with computer-assisted personal interviews in compliance with EQ-VT version 2.1.

Assessment of Prediction Accuracy

Resampling of data. Data were resampled with replacement at the level of individual study respondents, to create data sets with sample sizes ranging from 50 to the maximum number of respondents observed in each study (~ 1000 for NL and US, ~ 500 for NO), by intervals of 50. To reflect the design of the EQ-5D-5L valuation studies, in which respondents are conventionally assigned 1 out of 10 blocks of cTTO states for administration, resampling was balanced over TTO block. For each sample size in each study, we drew 1000 resamples. Each sampled respondent was attached their cTTO and DCE responses and assigned a pseudo-ID for use in mixed-effects modeling, ensuring a unique ID per observation even if the same respondent was resampled more than once.

In all 3 studies from which data were retrieved, each participant was administered 10 health states for cTTO valuation and 7 DCE state pairs. Thus, the sampling procedure would result in data sets with a minimum of 500 to a maximum of approximately 10 000 individual TTO values (approximately 5000 in the largest sample for the NO study) and 350 to approximately 7000 DCE choices.

Out-of-sample predictive accuracy by cross-validation. For each subsample, we used a cross-validation-based method to estimate out-of-sample predictive accuracy. The cross-validation method replicates that of previous cross-validations comparing models,¹⁷ by fitting a model on one part of the data and predicting values on the other, here by using a leave-one-out procedure at the level of TTO blocks, repeated until predicted values have been estimated for all 10 TTO blocks (Appendix Figure 2). To provide information about expected out-of-sample predictive accuracy as a function of sample size and the selected statistical model, we estimated the mean values for each TTO block/state combination from the full data set as the observed “true” value for comparison. As the cTTO values are left censored at -1 by the construction of the task, we used likelihood-based censored mean values throughout. Against this benchmark for comparison, all predictions were judged in

terms of the estimated root mean square error (RMSE) and summarized per sample size and model. The procedure followed these steps:

1. Respondents were sequentially left out by their assigned TTO block.
2. All candidate statistical models were fitted to the data from the remaining respondents.
3. Values for the 10 health states in the left-out TTO block were predicted for each candidate model.
4. Predicted values for health states in each left-out block was compared with observed censored mean values for the same block in the full sample and prediction accuracy estimated in terms of RMSE.
5. The previous steps were repeated until each block had been left out per sample, all models fitted to the rest of the data, and predicted values based on these models were compared with observed values.

Modeling Strategies Considered

A total of 8 candidate models were tested for comparison.

Two primary models were used; the “standard” additive 20-parameter model (Equation 1) and the constrained 8-parameter cross-attribute level-effect model (Equation 2):

1. 20-parameter model:

$$\begin{aligned} disutility = & \alpha + \\ & \beta_{MO2} \times MO_2 + \beta_{SC2} \times SC_2 + \beta_{UA2} \times UA_2 + \beta_{PD2} \times PD_2 + \beta_{AD2} \times AD_2 + \\ & \beta_{MO3} \times MO_3 + \beta_{SC3} \times SC_3 + \beta_{UA3} \times UA_3 + \beta_{PD3} \times PD_3 + \beta_{AD3} \times AD_3 + \\ & \beta_{MO4} \times MO_4 + \beta_{SC4} \times SC_4 + \beta_{UA4} \times UA_4 + \beta_{PD4} \times PD_4 + \beta_{AD4} \times AD_4 + \\ & \beta_{MO5} \times MO_5 + \beta_{SC5} \times SC_5 + \beta_{UA5} \times UA_5 + \beta_{PD5} \times PD_5 + \beta_{AD5} \times AD_5 + \varepsilon \end{aligned}$$

2. 8-parameter model:

$$\begin{aligned} disutility = & \alpha + \\ & (\beta_{MO} \times MO_2 + \beta_{SC} \times SC_2 + \beta_{UA} \times UA_2 + \beta_{PD} \times PD_2 + \beta_{AD} \times AD_2) \times \beta_{L2} + \\ & (\beta_{MO} \times MO_3 + \beta_{SC} \times SC_3 + \beta_{UA} \times UA_3 + \beta_{PD} \times PD_3 + \beta_{AD} \times AD_3) \times \beta_{L3} + \\ & (\beta_{MO} \times MO_4 + \beta_{SC} \times SC_4 + \beta_{UA} \times UA_4 + \beta_{PD} \times PD_4 + \beta_{AD} \times AD_4) \times \beta_{L4} + \\ & (\beta_{MO} \times MO_5 + \beta_{SC} \times SC_5 + \beta_{UA} \times UA_5 + \beta_{PD} \times PD_5 + \beta_{AD} \times AD_5) + \varepsilon \end{aligned}$$

These 2 primary models were used as TTO-only and TTO + DCE hybrid variants and both with and without random intercepts at the level of individual respondents. In the presentation of the results, models specifications are coded *T* for TTO only and *H* for hybrid models, *8* for 8-parameter models, and *20* for 20 parameters, and *r* indicating the inclusion of a random intercept, so that *T20r* denotes the 20-parameter TTO-only model with a random intercept, and *T20* denotes the same 20-parameter TTO-only model without a random intercept (Appendix Table 1). All candidate models and alternative model specifications have been used to value EQ-5D-5L health states for national EQ-5D value sets. The models included disutility per health state as the response and were right-censored at 2, the maximum disutility allowed in the TTO task.

In the hybrid models (Equation 3), the TTO likelihood was estimated precisely as for the TTO only models (i.e., a Tobit model of expressed disutility), right-censored at 2, the maximum disutility allowed in the cTTO task. The DCE likelihood was estimated using a conditional logit model. To account for the difference between the latent scale of the conditional logit model and the utility scale of the TTO model, the estimated value was multiplied by a nuisance parameter θ in the DCE side.

3. Hybrid model:

$$\underset{\alpha, \beta, \sigma, \theta}{\operatorname{argmax}} \{L(\alpha, \beta, \sigma, \theta; x)\} = \prod_{i=1}^n \begin{cases} \phi(x_i - f(\alpha_{tto}, \beta), \sigma) & x_i < 2, \text{ tto} \\ 1 - \Phi(x_i - f(\alpha_{tto}, \beta), \sigma) & x_i \geq 2, \text{ tto} \\ x_i \times \frac{\logit(\alpha_{dce} + \beta)}{\theta} + \\ (1 - x_i) \times \left(1 - \frac{\logit(\alpha_{dce} + \beta)}{\theta}\right) & \text{dce} \end{cases}$$

Table 1 Mean RMSE per Study and Sample Size (ss) for TTO-Only (T)/Hybrid (H) 20-/8-Parameter Models, with/without Random Intercept^a

Study	ss	T20	T20r	T8	T8r	H20	H20r	H8	H8r
Netherlands	50	0.178	0.164	0.130	0.132	0.127	0.138	0.114	0.126
	100	0.136	0.128	0.105	0.108	0.103	0.109	0.096	0.103
	200	0.112	0.106	0.091	0.092	0.089	0.092	0.086	0.089
	350	0.100	0.095	0.084	0.085	0.083	0.084	0.082	0.082
	500	0.095	0.091	0.081	0.081	0.080	0.080	0.080	0.079
	750	0.090	0.087	0.079	0.079	0.078	0.078	0.078	0.077
	950	0.088	0.085	0.078	0.078	0.077	0.077	0.078	0.076
	~ 990*	0.088	0.085	0.078	0.077	0.077	0.076	0.077	0.076
United States	50	0.209	0.189	0.154	0.157	0.146	0.169	0.132	0.154
	100	0.159	0.146	0.124	0.126	0.120	0.130	0.113	0.122
	200	0.128	0.119	0.108	0.107	0.105	0.110	0.102	0.104
	350	0.113	0.104	0.101	0.097	0.099	0.096	0.098	0.095
	500	0.107	0.097	0.098	0.092	0.097	0.090	0.096	0.090
	750	0.101	0.092	0.096	0.089	0.094	0.086	0.095	0.087
	950	0.099	0.090	0.095	0.087	0.094	0.084	0.094	0.086
	1000	0.099	0.089	0.094	0.087	0.093	0.084	0.094	0.086
	1050	0.098	0.089	0.094	0.086	0.093	0.084	0.094	0.085
	1100	0.098	0.088	0.094	0.086	0.093	0.083	0.094	0.085
1134	0.098	0.088	0.094	0.086	0.093	0.083	0.094	0.085	
Norway	50	0.174	0.163	0.136	0.138	0.137	0.151	0.126	0.141
	100	0.144	0.137	0.120	0.120	0.121	0.127	0.113	0.119
	200	0.128	0.120	0.110	0.109	0.112	0.113	0.106	0.107
	350	0.120	0.111	0.106	0.102	0.108	0.106	0.103	0.101
	500	0.117	0.108	0.104	0.100	0.106	0.103	0.101	0.099

RMSE, root mean square error; TTO, time tradeoff.

^ar indicates with random intercept. Values in bold indicate a mean RMSE \leq the mean RMSE for the 20-parameter TTO-only (T20) model at the maximum sample size.

* $n = 989$ for TTO-only models; $n = 992$ for hybrid models.

where $\text{logit}(x)$ is the standard logistic function $\frac{1}{1+e^{-x}} = \frac{1 + \tanh(\frac{x}{2})}{2}$; $f(\cdot)$ is the 8- or 20-parameter model; θ is a nuisance scaling parameter between TTO and DCE parameters, to account for the latent scale of the DCE; ϕ is the normal probability density function; Φ is the normal cumulative density function; σ is the variance parameter for the normal distribution; α is the intercept (separate for TTO and DCE); and β is a vector of parameters for f . For a similar model with random intercepts, a further parameter is fitted for the between-subjects variance.

In the absence of a natural gold standard for comparison, we used the mean level of prediction accuracy for the standard additive 20-parameter model with the largest sample size ($n = \sim 1000$) as a threshold to compare models at different sample sizes.

Effect of Sample Size and Model Choices

To assess the relative effect size of increases in sample size and each of the model choices (20- v. 8-parameter model,

adding a random intercept and including DCE data in a hybrid model), a linear regression model was applied to estimate the effect of increasing the sample size and each model choice on the RMSE per study using the following formula:

$$RMSE \sim \beta_1 \log_2(ss) + \beta_2 \text{par8} + \beta_3 \text{rand} + \beta_4 \text{hyb} + \beta_5 \text{study} + e$$

The sample size was included in the model as the binary logarithm (\log_2) of sample size, giving estimates interpreted as the effect of doubling sample size. Each of the model choices were coded as dummy variables (par8 indicating 8 parameter v. 20 parameter, coded 1 for 8 parameter; rand indicating the inclusion of a random intercept v. not, coded 1 for with random intercept; hyb indicating TTO + DCE hybrid model v. TTO-only model, coded 1 for hybrid model). Study was included in the model as a fixed effect (coded as a factor with values NL, US, NO).

All analyses were performed in R version 4.0.0.¹⁹

Results

The data from the NL and US study included 9787 and 10,620 TTO values, with a mean disutility per health state of 0.69 (SD 0.66) and 0.70 (0.70). The data collected so far in Norway included 5110 TTO values, with a mean disutility per health state of 0.59 (SD 0.60). The distribution of values was similar between the 3 studies, with slightly less clustering at disutility = 1 and a higher proportion of valuations with a disutility less than 0.5 in the NO data (Appendix Figure 1).

For all tested models, an increase in sample size has the greatest effect on the mean RMSE on sample sizes up to 500, with rapidly diminishing marginal improvement for further increases in sample size. Across all sample sizes and models, the mean RMSE varied from 0.178 to 0.077 for the NL data and from 0.209 to 0.084 for the US data (Table 1). The NO data, resampled from an original sample of only 510 respondents, had a max mean RMSE of 0.174 for samples of 50 respondents and reached a minimum mean RMSE of 0.099 for maximum sample size (~ 500).

For models including TTO data only, the 8-parameter models outperformed the more complex 20-parameter models for all sample sizes and all 3 studies in terms of out-of-sample RMSE (Figure 1), achieving the same expected prediction accuracy for samples of 350 to 500 respondents as for the standard 20-parameter model with ~ 1000 respondents. For the 20-parameter model, adding a random intercept also improved the mean prediction accuracy for all sample sizes and studies (Figure 2). For the NL data, there was no further improvement in mean RMSE for the 8-parameter models when adding a random intercept.

Using a hybrid model increased the prediction accuracy for all sample sizes compared with the corresponding TTO-only models, with the greatest observed difference for the 20-parameter models (Figure 1). For the 8-parameter models with a random intercept, the reduction in the mean RMSE from switching from a TTO-only to a hybrid model was less than 0,01 for all sample sizes.

The lowest sample size yielding similar mean RMSE as the standard 20-parameter TTO-only model was about 200 respondents using a hybrid 8-parameter model based on the NL data and 300 respondents using any hybrid models or 8-parameter random intercept TTO-only models based on the US data. However, smaller sample sizes do have much larger confidence intervals

and risk of significantly higher RMSE. It should also be noted that where the upper confidence interval limit indicates the extent of the error expected, the lower confidence interval has no practical interpretation in this context.

The differences in mean RMSE between the combined model choices based on ~ 1000 respondents were greater than the difference between sample sizes of 500 versus ~ 1000 using the same model. For instance, estimates from the regression model indicate that using, for example, an 8-parameter random intercept model, or a DCE + TTO hybrid model with a random intercept, could be expected to achieve similar or better prediction accuracy with a sample size of ~ 500 compared with a 20-parameter TTO-only model fitted to a sample size of ~ 1000.

The regression model estimating RMSE by change in sample size and model specification indicated that doubling the sample size decreased the RMSE by 0.012 ($p < 0.01$; Table 2). Switching to an 8-parameter model, adding a random intercept, or using a DCE + TTO model also decreased the RMSE by 0.007 ($p < 0.01$), 0.003 ($p < 0.01$), and 0.008 ($p < 0.01$), respectively. The study also had a significant effect on the RMSE.

The data included in the regression model comprised the RMSE for all tested models for each resample and sample size category (i.e., repeated measurements per sample). Although the regression coefficients are unbiased, the dependence between observations can lead to the underestimation of standard errors and thereby incorrect P values. Simulation modeling indicated that independent samples would have yielded slightly larger standard errors (0.1%–0.2%) and would not have changed the main findings.

Corresponding to the regression analyses, the main Figures 1 and 2 are presented with the logarithm of sample size as Figures 3 and 4 in the Appendix.

Discussion

The major finding of this study was that prediction accuracy, expressed as the RMSE, decreased with increasing sample size, and that different model specifications displayed substantial differences in prediction accuracy across sample sizes. The multiplicative 8-parameter model outperformed the more complex 20-parameter model for all sample sizes based on TTO data only for all 3 studies. For all models, the improvement in expected prediction accuracy tapered off quickly with increasing sample size, and there was minimal gain in terms of average expected prediction accuracy from increases in sample size above 300 to 500 respondents.

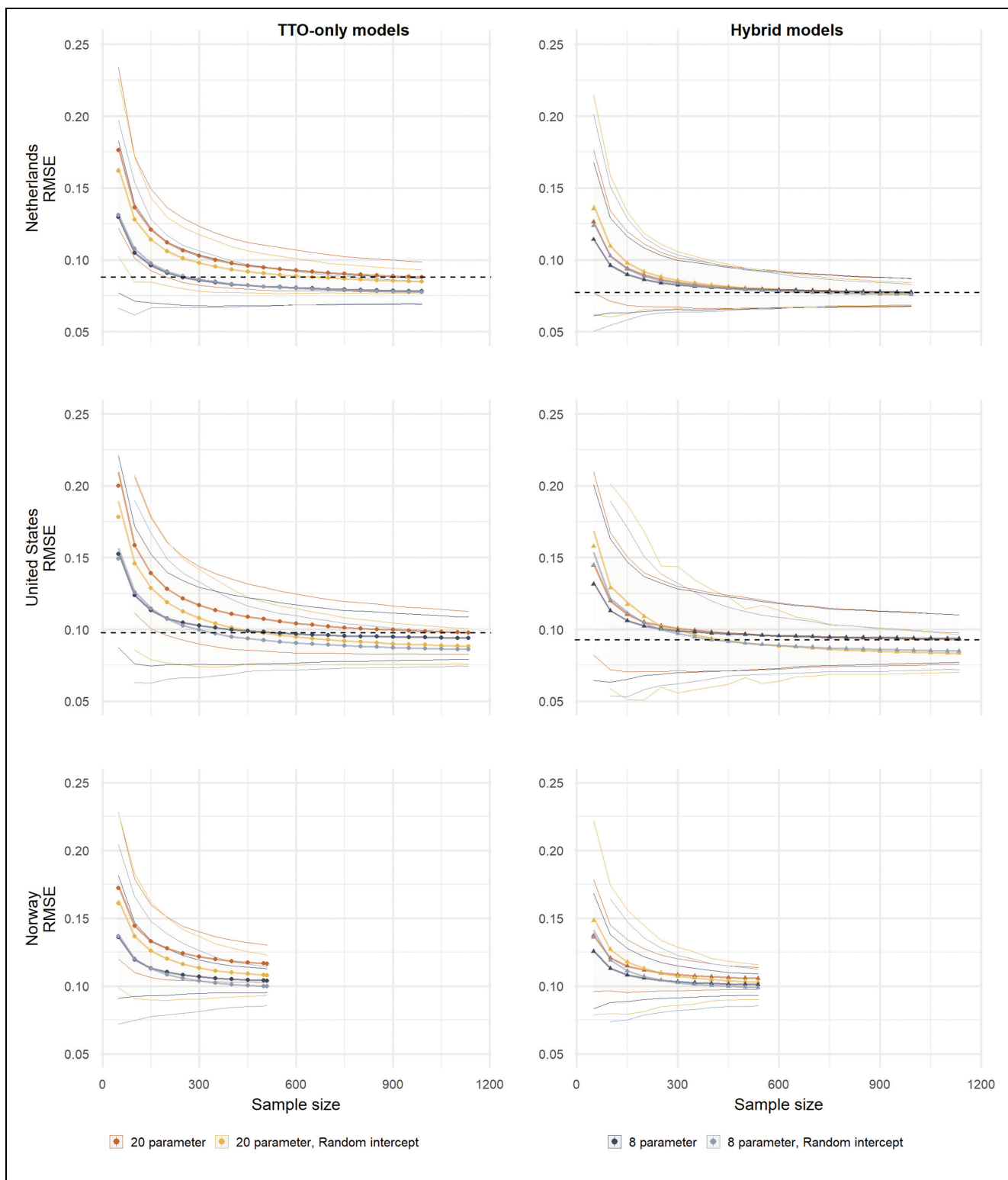


Figure 1 Root mean square error (RMSE) by sample size for variants of time tradeoff (TTO) data only and TTO + discrete choice experiment hybrid models. RMSE calculated from the predicted values compared with censored mean observed values for states included for direct valuation. The black dashed line indicates the mean RMSE for the 20-parameter model without a random intercept at the maximum sample size.

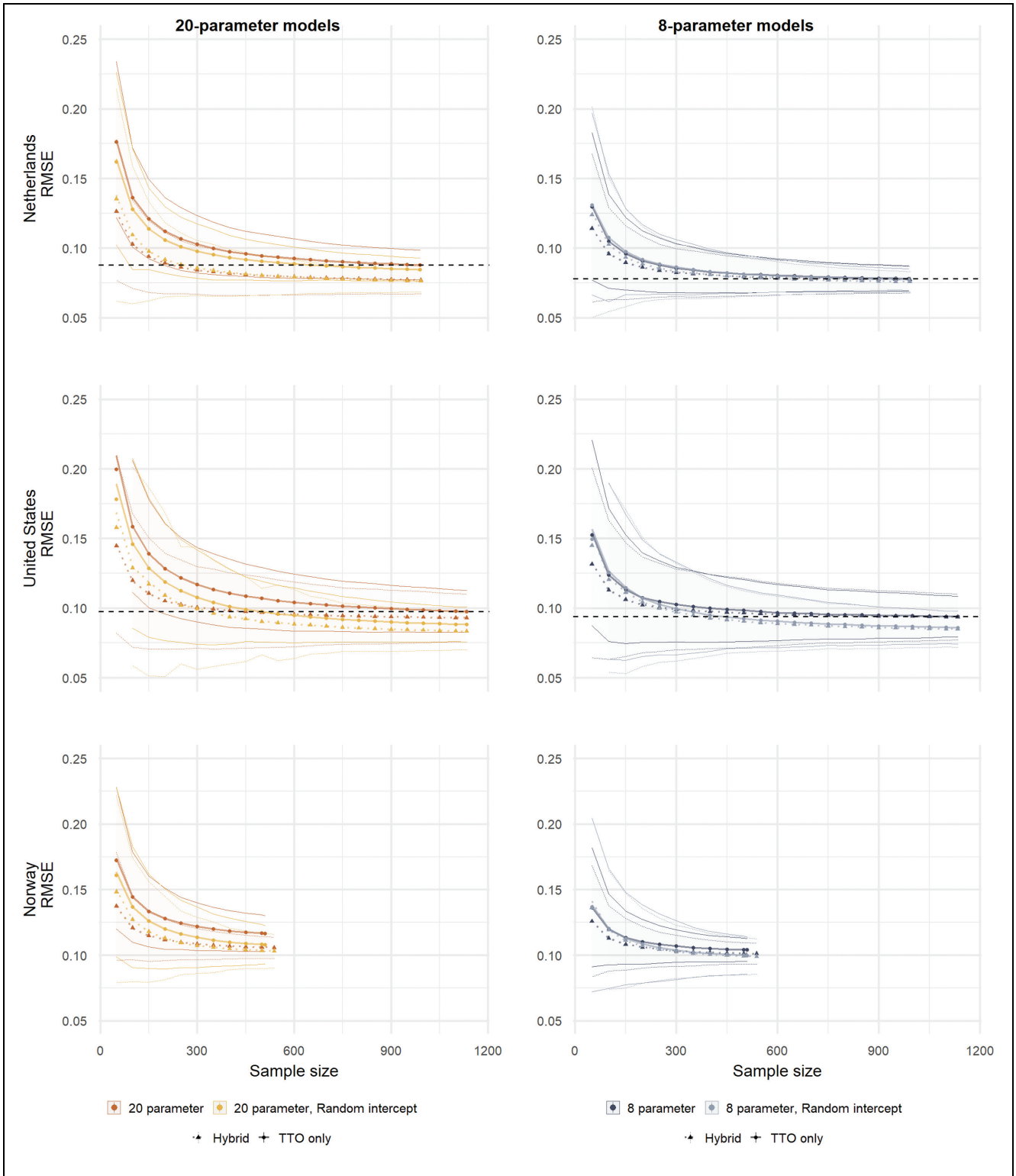


Figure 2 Root mean square error (RMSE) by sample size for variants of the 20-parameter additive and 8-parameter multiplicative models. RMSE calculated from predicted values compared with the censored mean observed values for states included for direct valuation. The black dashed line indicates the mean RMSE for the time tradeoff-only model without a random intercept at max sample size.

Table 2 Results from the Linear Regression Model Estimating the Effect of Doubling Sample Size (Binary Logarithm [\log_2] of Sample Size), using an 8-Parameter Model (8 Parameter), Adding a Random Intercept (Random), and including Discrete Choice Experiment Responses in a Hybrid Model (Hybrid) on Root Mean Square Error (RMSE), Overall and Per Study

	Dependent Variable: RMSE			
	Estimate (SE)	Netherlands	United States	Norway
$\log_2(\text{sample size})$	-0.012*** (0.00002)	-0.011*** (0.00003)	-0.013*** (0.00004)	-0.011*** (0.0001)
8 parameter	-0.007*** (0.00005)	-0.008*** (0.0001)	-0.005*** (0.0001)	-0.010*** (0.0001)
Random	-0.003*** (0.00005)	-0.001*** (0.0001)	-0.006*** (0.0001)	-0.002*** (0.0001)
Hybrid	-0.008*** (0.00005)	-0.009*** (0.0001)	-0.007*** (0.0001)	-0.007*** (0.0001)
Netherlands	0.204*** (0.0002)			
Norway	0.219*** (0.0002)			
United States	0.218*** (0.0002)			
Constant		0.196*** (0.0003)	0.227*** (0.0003)	0.212*** (0.0004)
Observations	432,000	160,000	184,000	88,000
R^2	0.975	0.532	0.440	0.421
Adjusted R^2	0.975	0.532	0.440	0.421
Residual standard error	0.016 ($df = 431,993$)			
F statistic	2,380,826.000*** ($df = 7; 431,993$)			

* $P < 0.1$; ** $P < 0.05$; *** $P < 0.01$.

A comparison with the results for NL and US data showed that the samples with a maximum sample size from the NO data achieve similar prediction accuracy using a 8-parameter multiplicative TTO model with random intercept (mean RMSE = 0.100) or 8-parameter multiplicative hybrid model with random intercept (mean RMSE = 0.099). Results suggest that modeling can compensate for the smaller sample size achieved by the Norwegian study in terms of prediction accuracy.

The same cross-validation method was used to compare models for estimating EQ-5D-5L health state values in a previous study using data from Spain, Singapore, and China.¹⁷ Our results support previous findings of the simpler 8-parameter model outperforming the standard 20-parameter model and that adding a random intercept at the level of the respondent may further improve predictive accuracy. This study indicates that these findings also hold across all sample sizes tested (50–1000).

The effect of sample size and threshold for the minimum number of respondents (300–500) is supported by previous studies exploring the sample size and prediction

accuracy of EQ-5D-3L values.^{12,13} A minimum of 500 respondents resulted in stable estimates for health states using a 20-parameter VAS main effects model using a random intercept,¹⁵ given direct valuations for 80–120 states. Two recent studies have shown that smaller designs with direct valuation of fewer states can suffice for estimation of EQ-5D-5L health states without compromising prediction accuracy.^{20,21} If the current “core” EQVT valuation study protocol is to be considered sufficiently good, our findings indicate that study costs could be lowered by reducing sample size, without substantial impact on prediction accuracy. Our results suggested that the impact of increases in sample size beyond 300 to 500 respondents is minimal using the EQ-VT design and the statistical model specifications currently employed.

It should be noted that the out-of-sample predictive accuracy of even the best-performing models with a maximum sample size in our analyses was of a magnitude similar to reported minimally important differences for the EQ-5D-5L.²² This reflects the ability of these models to reflect the nuances of the aggregated preferences for

the EQ-5D-5L health states and is only affected by the sample size to a limited extent. Although our results suggest that sample sizes could be reduced at limited cost to prediction accuracy, the similar magnitude of expected error to minimally important differences suggests that there is likely room for improvement both in terms of state design (e.g., by increasing the number of health states valued)¹³ and to the valuation methods and statistical models used to reflect population preferences.

The model specifications tested in this study have all been used to estimate national value sets for the EQ-5D-5L. The candidate models are, however, not exhaustive, and other designs and models have been suggested and used to estimate health states. Approaches such as adding state-level random effects have been shown to improve prediction accuracy,^{23,24} but such approaches have yet to be explored across sample sizes.

The studies included in the present analyses were all national data collections compliant with the EQ-VT protocol. Although adhering to different versions of the protocols, compliance with these protocols ensures that interviews were completed in a comparable manner, all using face-to-face personal interviews using digital software developed for the valuation of the EQ-5D-5L. All 3 studies used TTO and DCE and trained their interviewers as recommended by EuroQol. Newer versions of the protocols, as used by the US and NO study, included an extended introduction to the TTO task, with practice states and the adaptive lead-time TTO example, and more data quality controls to identify interviewer effects. However, the greater focus on data quality has not resulted in a higher prediction accuracy when modeling health state values, with consistently higher prediction errors from models based on the US and NO data compared with the NL data. The differences in the level of prediction error in these studies can be a result of differing response patterns and use of the scale; for instance, with more detailed introduction of the lead-time part of the task, respondents may be more inclined to use a larger portion of the total scale when valuing health states, resulting in greater variance in observed values.

Limitations

There is no defined threshold for an acceptable prediction accuracy of models estimating health state values. The chosen threshold in this study, the average predicted accuracy achieved with the “standard” 20-parameter model based on TTO data only, was used to compare the effect of each additional alternative modeling

specification across sample size and is not suggested to be a standard of acceptable prediction accuracy.

We compared models and model specifications by mean RMSE across samples per sample size. For smaller sample sizes, the RMSE per sample for all models naturally varied significantly, and the uncertainty and likelihood of a model achieving a much higher RMSE increased with smaller samples. We compared the predicted and observed values without regard for standard errors or parameter significance. In the final estimation of values, models are often assessed by the number of significant parameters and underlying assumptions of utility, such as monotonicity (i.e., that increases in severity levels equate to decreases in utility). When including health state values in cost-effectiveness analyses, the uncertainty of the predicted values and standard errors of the estimates are also relevant, both of which will naturally increase with decreasing sample size. These measures were, however, not assessed in this study.

Although the data included in the analyses were sampled from reputable and recent EQ-5D-5L studies, complying with the EQ-VT protocol using the standard design (direct valuation of 86 states), they were also chosen by convenience and with populations that were considered comparable with the Norwegian population. Given that different countries and cultures have shown different response patterns, findings from this study may not transfer to all settings.

Respondent characteristics affect the values given during TTO, and as such, the validity of the estimated value sets depends on whether the characteristics of the respondents in the sample are representative of the population they seek to represent.^{25,26} The resampling of data to differing sample sizes in this study did not take any respondent characteristics into account, and we cannot rule out that differences between study populations also affected the results. Despite this, the RMSE scores per study were similar, and the effect of each model specification on prediction accuracy was comparable across all 3 studies. Given that respondent characteristics can have a significant effect on health state values, different compositions of such characteristics in the different samples will add to the possible error and imprecision of predictions, with the smaller sample sizes naturally being most susceptible to more extreme combinations.

Implications of results

The implications of results of this study are 2-fold. For the Norwegian valuation study, for which data collection was stopped in March 2020 after the completion of only

542 interviews, the results of this study could support the estimation of values based on the data already collected. Given the context of the postponement, the comparability to previously collected data may be in question if data collection is resumed regardless. Following a prolonged interruption to data collection, new interviewers will need to be recruited and trained. Previously unexplored issues may also need to be addressed, such as potential shifts in population preferences due to the COVID-19 pandemic and changes in the health political climate.

Issues with the political legitimacy of values may, however, still be debatable. Geographical representation of all regions of Norway was a priority in the sampling strategy. Because of the sudden postponement, data were collected in only 2 of 4 regions. Although there is little prior knowledge suggesting that region would have a significant effect on health state preferences, cultural differences have been associated with health state values.²⁷ The use of values generated from preferences collected only in southern, and more urban and densely populated, regions of Norway may not be deemed politically acceptable as a national value set. Thus, while estimating values from a smaller sample size may be defensible in terms of the average expected prediction accuracy, modeling cannot properly adjust for inadequate representativeness in the achieved sample.

Conclusions

Sample size will always be a tradeoff between precision and costs. The more respondents, the greater the precision of estimates but also the greater the costs. Results from this study suggest that the expected gain in prediction accuracy from increasing sample sizes beyond 300 to 500 respondents is minimal and that the choice of model can compensate for a smaller sample size. Beyond this number of respondents, sample size considerations in the planning of national valuation studies may be better informed by considerations of legitimacy and representativeness than by the technical prediction accuracy achievable.


Acknowledgments

The authors would like to thank Andrew Garratt, principal investigator for the Norwegian Valuation Study project, who also secured the funding, and Ylva Helland, who made a substantial contribution to organizational aspects of the Norwegian data collection. Permission for use of the data from the US valuation study was graciously given by Professor Simon Pickard and for the NL valuation study data by Matthijs Versteegh and Elly Stolk.

Ethics

The Regional Committee for Medical and Research Ethics in Norway reviewed the Norwegian study and stated that their approval was not required. The Norwegian Institute of Public Health approved the Data Protection Impact Assessment for the study on September 30, 2019.

ORCID iD

Tonya Moen Hansen  <https://orcid.org/0000-0003-3150-4765>

Availability of Data

The Norwegian raw data cannot be shared due to privacy laws in Norway. Permission for use of the data from the US valuation study was graciously given by Professor Simon Pickard and for the NL valuation study data by Matthijs Versteegh and Elly Stolk.

Code Availability

R code can be shared upon reasonable request to the corresponding author.

Supplemental Material

Supplementary material for this article is available on the *MDM Policy & Practice* website at <https://journals.sagepub.com/home/mpp>.

References

1. Oppe M, Devlin NJ, van Hout B, et al. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value Health*. 2014;17:445–53. doi:10.1016/j.jval.2014.04.002
2. Oppe M, Rand-Hendriksen K, Shah K, et al. EuroQol protocols for time trade-off valuation of health outcomes. *Pharmacoeconomics* 2016;34:993–1004. doi:10.1007/s40273-016-0404-1
3. Oppe M, Hout B. The “power” of eliciting EQ-5D-5L values: the experimental design of the EQ-VT. *EuroQol Working Paper Series*. No. 17003. October 2017.
4. Stolk E, Ludwig K, Rand K, et al. Overview, update, and lessons learned from the international EQ-5D-5L valuation work: version 2 of the EQ-5D-5L valuation protocol. *Value Health*. 2019;22:23–30. doi:10.1016/j.jval.2018.05.010
5. Hansen TMH, Y, Augestad L.A., Rand K, Stavem K, Garratt A.M. Elicitation of Norwegian EQ-5D-5L values for hypothetical and experience-based health states based on the EuroQol valuation technology (EQ-VT) protocol. *BMJ Open*. 2020;10:e034683. doi:10.1136/bmjopen-2019-034683
6. Ramos-Goni JM, Oppe M, Slaap B, et al. Quality control process for EQ-5D-5L valuation studies. *Value Health*. 2017;20:466–73. doi:10.1016/j.jval.2016.10.012

7. Pickard AS, Law EH, Jiang R, et al. United States valuation of EQ-5D-5L health states using an international protocol. *Value Health*. 2019;22:931–41. doi:10.1016/j.jval.2019.02.009
8. Leidl R, Reitmeir P. An experience-based value set for the EQ-5D-5L in Germany. *Value Health*. 2017;20:1150–6. doi:10.1016/j.jval.2017.04.019
9. Versteegh M, Vermeulen K, Evers S, et al. Dutch tariff for the five-level version of EQ-5D. *Value Health*. 2016;19:343–52. doi:10.1016/j.jval.2016.01.003
10. Xie F, Pullenayegum E, Gaebel K, et al. A time trade-off-derived value set of the EQ-5D-5L for Canada. *Med Care*. 2016;54:98–105. doi:10.1097/mlr.0000000000000447
11. Chan KKW, Xie F, Willan AR, et al. Conducting EQ-5D valuation studies in resource-constrained countries: the potential use of shrinkage estimators to reduce sample size. *Med Decis Making*. 2018;38:26–33. doi:10.1177/0272989X17725748
12. Chan KKW, Pullenayegum EM. The theoretical relationship between sample size and expected predictive precision for EQ-5D valuation studies: a mathematical exploration and simulation study. *Med Decis Making*. 2020;40:339–47. doi:10.1177/0272989x20915452
13. Shams S, Pullenayegum E. Design and sample size considerations for valuation studies of multi-attribute utility instruments. *Stat Med*. 2020;39:3074–104. doi:10.1002/sim.8592
14. Gandhi M, Xu Y, Luo N, et al. Sample size determination for EQ-5D-5L value set studies. *Qual Life Res*. 2017;26:3365–76. doi:10.1007/s11136-017-1685-3
15. Bonsel G OM, Janssen M.. Optimization of the design of multi-attribute vignette studies: a simulation study based on the multinational EQ-5D-5L pilot studies. Presented at: EuroQol Plenary Meeting; Berlin, Germany; 2016.
16. Pullenayegum EM, Chan KK, Xie F. Quantifying parameter uncertainty in EQ-5D-3L value sets and its impact on studies that use the EQ-5D-3L to measure health utility: a Bayesian approach. *Med Decis Making*. 2016;36:223–33. doi:10.1177/0272989x15591966
17. Rand-Hendriksen K, Ramos-Goñi JM, Augestad LA, et al. Less is more: cross-validation testing of simplified nonlinear regression model specifications for EQ-5D-5L health state values. *Value Health*. 2017;20:945–52. doi:10.1016/j.jval.2017.03.013
18. Hansen TM, Helland Y, Augestad LA, et al. Elicitation of Norwegian EQ-5D-5L values for hypothetical and experience-based health states based on the EuroQol Valuation Technology (EQ-VT) protocol. *BMJ Open*. 2020;10:e034683. doi:10.1136/bmjopen-2019-034683
19. R_Core_Team. *R: A Language and Environment for Statistical Computing*. 2020. Available from: <https://www.R-project.org/>
20. Yang Z, Luo N, Oppe M, et al. Toward a smaller design for EQ-5D-5L valuation studies. *Value Health*. 2019;22:1295–302. doi:10.1016/j.jval.2019.06.008
21. Yang Z, Luo N, Bonsel G, et al. Effect of health state sampling methods on model predictions of EQ-5D-5L values: small designs can suffice. *Value Health*. 2019;22:38–44. doi:10.1016/j.jval.2018.06.015
22. Henry EB, Barry LE, Hobbins AP, et al. Estimation of an instrument-defined minimally important difference in EQ-5D-5L index scores based on scoring algorithms derived using the EQ-VT version 2 valuation protocols. *Value Health*. 2020;23:936–44. doi:10.1016/j.jval.2020.03.003
23. Kharroubi SA, O'Hagan A, Brazier JE. Estimating utilities from individual health preference data: a nonparametric Bayesian method. *J R Stat Soc Ser C (Appl Stat)*. 2005;54:879–95. doi:10.1111/j.1467-9876.2005.00511.x
24. Shams S, Pullenayegum E. Reducing uncertainty in EQ-5D value sets: the role of spatial correlation. *Med Decis Making*. 2019;39:91–9. doi:10.1177/0272989X18821368
25. Dolan P, Roberts J. To what extent can we explain time trade-off values from other information about respondents? *Soc Sci Med*. 2002;54:919–29. doi:10.1016/s0277-9536(01)00066-1
26. van Nooten F, Busschbach J, van Agthoven M, et al. What should we know about the person behind a TTO? *Eur J Health Econ*. 2018;19:1207–11. doi:10.1007/s10198-018-0975-1
27. Mahlich J, Dilokthornsakul P, Sruamsiri R, et al. Cultural beliefs, utility values, and health technology assessment. *Cost Effectiveness Resource Allocation : C/E* 2018; 16: 19. 2018/06/09. DOI: 10.1186/s12962-018-0103-1.