Department of Economics and Related Studies

Thomas Clemmet

Supervisor: Alex Thompson, PhD

Validating Methods for Estimating Age- and Sex-Adjusted Baseline Utilities

MSc Health Economics

Wordcount: 9,945

Submitted 13th of September 2023

Acknowledgement

I’m very grateful to my supervisor, Alex Thompson, and the wider team at the Manchester Centre for Health Economics for their support and feedback throughout this project.

Abstract

Age- and sex-adjusted HSUV baselines are sometimes required for HTA submissions, especially since NICE's 2022 introduction of absolute/proportional QALY shortfalls to determine severity modifiers. Research on estimating these baselines has been sparse and has not considered a broad range of models. I implement a stratified k-fold cross-validation procedure to test the performance of a selection of baseline utility models: an ALDVMM, a range of polynomial models and a range of RCS models. My results fail to support the DSU's recommendation of the ALDVMM. Polynomial models are useful but higher-order versions produce counterintuitive predictions for older people. RCS models avoid this problem, with the optimal number of knots (i.e. model flexibility) being around 7. However, the differences in models' baselines seem unlikely to be decisive in economic evaluations, and there are significant difficulties in estimating HSUVs for older people which further research ought to thoroughly investigate.

See https://github.com/tomclemmet/baseline-utilities for the code used in this project.

Abbreviations

AIC	Akaike's Information Criterion
ALDVMM	Adjusted Limited Dependent Variable Mixture Model
DSU	Decision Support Unit
HRQoL	Health-Related Quality of Life
HSE	Health Survey for England
HSUV	Health State Utility Value
HTA	Health Technology Appraisal
MAE	Mean Absolute Error
ME	Mean Error
MVH	Measuring and Valuing Health (Study)
NICE	National Institute for Health and Care Excellence
OLS	Ordinary Least Squares
PROM	Patient-Reported Outcome Measure
QALE	Quality Adjusted Life Expectancy
QALY	Quality Adjusted Life Year
RCS	Restricted Cubic Spline
RCT	Randomised Controlled Trial
RMSE	Root Mean Squared Error

Contents

Tables

Figures

1 Introduction

Decision analytic models submitted in health technology appraisals (HTAs) involve making predictions about the health benefits of treatments on a typical patient. Ideally, such predictions are made using randomised controlled trials (RCTs), but in practice it is not always possible to conduct these, especially when estimating a continuous outcome like health-related quality of life (HRQoL) rather than a binary outcome like effectiveness. In some cases, this leads to the need for age- and sex-adjusted baseline utilities. For example, to understand the value of an intervention which cures a chronic long-term condition, researchers will need to predict the utility of a cured patient over their lifetime. Doing this with an RCT would be prohibitively costly and time-consuming. Instead, modellers could assume that these patients simply follow the trajectory of HRQoL observed in the general population. NICE (2022b, p.195) highlight this point in their most recent guidance, stating that:

> *"In some circumstances adjustments to utility values may be needed, for example for age or comorbidities. If baseline utility values are extrapolated over long time horizons, they should be adjusted to reflect decreases in health-related quality of life seen in the general population"*

The 2022 guidance also introduced severity modifiers. These change how quality-adjusted life-years (QALYs) are calculated based on either proportional or absolute QALY shortfall relative to *"the total QALYs that the general population with the same age and sex distribution would be expected to have*" (NICE, 2022b, p.163). Over the last two decades the need for accurate baseline utilities has often been highlighted in the wider health utilities literature (Manca, Hawkins and Sculpher, 2005; Ara and Brazier, 2011b; Ara and Wailoo, 2011; Ara, Brazier and Zouraq, 2017), particularly in

work on estimating utilities for joint health state utilities (Ara and Brazier, 2011a; Ara and Wailoo, 2013; Thompson, Sutton and Payne, 2019; Thompson, 2019). They have also been applied in numerous NICE technology appraisals (e.g. NICE, 2015, 2022a).

Given that there is a significant need for general population health state utility value (HSUV) baselines in HTA submissions, the baselines produced must be both accurate and consistent across submissions. Otherwise, predictions of health benefits may not reflect real-world use, partly because modellers would be able to choose whichever baseline suited their treatment best. The NICE guidance contains few details concerning baselines, simply recommending that they should be sourced from “recent and robust” data (NICE, 2022b, p.72,164). In addition, published research on this issue has been fairly sparse, and the research that does exist has focussed on a narrow range of methods. In this paper, I present an analysis of different methods for calculating HSUV baselines using the Health Survey for England (HSE) dataset. I explore several issues, the most significant being the optimal degree of model flexibility to use. I implement a stratified k-fold cross-validation procedure, an approach popular in predictive modelling that has been previously implemented in similar HSUV-related contexts. While the results are somewhat difficult to interpret due to significant unexplained variation in the data, the analysis suggests that previous research may not have fully appreciated complex non-linearities in the relationship between age, sex and HSUVs which require more flexible models to detect.

2 Literature Review

2.1 Non-Parametric Approaches

Research producing age- and sex-adjusted HSUV baselines for the general population has been fairly scarce. An R Shiny web application (https://shiny.york.ac.uk/shortfall Schneider et al., 2023) has been published which reports some of the baselines that have been produced and uses them to calculate remaining QALYs and absolute/ proportional QALY shortfalls for a given population and treatment. However, different models and datasets have been used to produce each baseline, demonstrating a lack of consistency. In this section, I will review these baselines and the studies that produced them.

The simplest way of producing population baselines for HSUVs is to simply take sample means for people of different ages (or age groups) and sexes. This approach can be described as non-parametric as it does not involve any assumptions about the form of the relationship between HSUVs and their predictors. However, while an absence of assumptions provides a 'clean slate' for analysis, it can be a limitation if non-parametric results violate common sense principles. For example, one would expect the difference in baselines from one year to the next to be small, but random variation may mean that the sample means in adjacent years are significantly different. Nonetheless, non-parametric approaches provide a useful starting point for methods in this field and have been applied in a couple of papers.

The first published HSUV population norms for the UK came from the 'Measuring and Valuing Health (MVH) study, which informed the EQ-5D-3L value set in the UK (Dolan, 1997) and were reported by Kind et al. (1999). Mean HSUVs are reported for males and females in 10-year age bands from 'under 25' to 'over 75'. The use of age bands

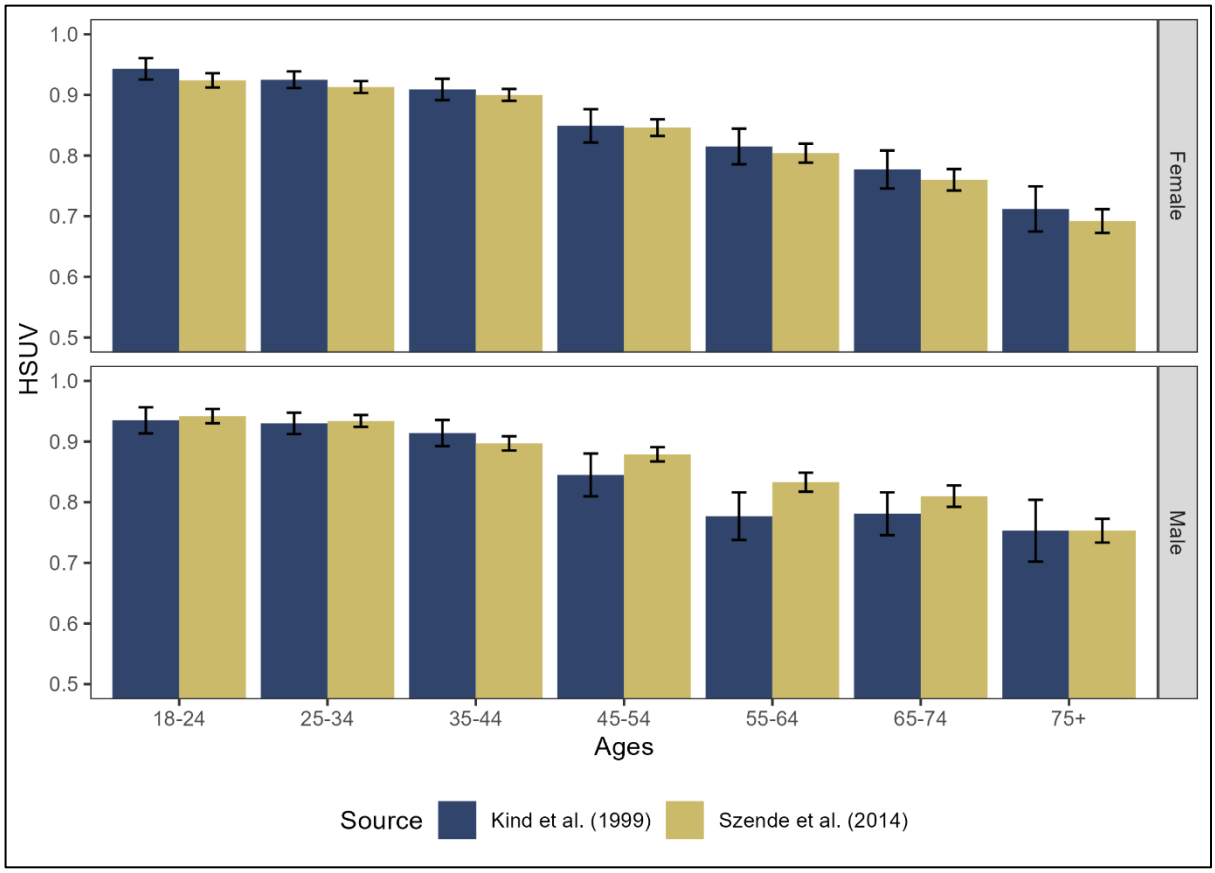

Figure 1: Published mean HSUVs with 95% confidence intervals

is presumably to optimise sample sizes (the study had only 3,395 participants). Without making it explicit by specifying a functional form, the authors implicitly adhere to the idea that the relationship between age and baseline HSUVs is smooth and monotonic by using large age bands. Using narrower age bands might produce erratic baselines which violate this principle. In a similar context in Norway, researchers even choose to remove the results from a non-conforming age band and used a weighted average from neighbouring categories to maintain monotonicity (Stavem et al., 2018; Norwegian Institute of Public Health, 2021). Unfortunately, grouping age into categories means that the model will struggle to accurately predict HSUVs in subjects for which the exact age is known. For example, Kind et al.'s results imply that the HSUV baseline remains constant after the age of 75 when in reality one would expect a decline.

Szende et al. (2014) calculate HSUV baselines for the same age groups using a different dataset. The 2010 Health Survey for England (HSE) was used, a sample that improves upon the MVH data in that it is more recent and significantly larger (n = 14,763), meaning that the confidence intervals for the estimates of the means are narrower. The two sets of baselines produced are shown in Figure 1, with 95% confidence intervals shown by the error bars[1]. The results of the two studies are broadly similar, especially for women. For men, the later study is a little more optimistic about the baseline HSUVs, though the estimates mostly remain within the confidence intervals of the other study's results.

Before moving on to other methodological approaches, it is worth noting that age-group means have recently been published for the newer five-level (5L) version of the EQ-5D using data from the 2017 and 2018 HSE (McNamara et al., 2023). Studying trends in the EQ-5D-5L is currently limited by the lack of an approved value set (NICE, 2019), so the current practice is to map 5L data to 3L HSUVs using a published algorithm (NICE, 2022b, p.74). NICE itself admits that mapping is an imperfect solution, and their preferred algorithm takes age and sex as arguments (Hernández Alava, Pudney and Wailoo, 2023), so age- and sex-adjusted baselines using 5L data are difficult to interpret. For this reason, this paper will focus only on the 3L, though when an approved UK value set for the EQ-5D-5L is published it may be of interest to repeat this study using the newer metric.

[1] Confidence intervals for the sample means (here, and throughout the paper) are calculated using the standard error formula $\widehat{\sigma_{\bar{x}}} = \frac{\widehat{\sigma_x}}{\sqrt{n}}$ and applying the central limit theorem to assume a normal distribution.

2.2 Parametric Predictors

HSUV baseline models explicitly intended to predict have largely been parametric. Parametric approaches choose a specific mathematical function for the relationship between variables and then choose the parameters of the function such that it 'fits' the data. This approach makes its assumptions explicit – the smooth monotonicity that non-parametric methods often aim for is built into the chosen functional form. This class of estimators will be the main focus of this study.

The first published parametric age/sex model for general population HSUVs can be found in Ara & Brazier (2010). This model forms a small part of the paper, the primary focus of which is how different approaches to combining HSUVs affect a decision analytic model, and is therefore not discussed in detail. The model specification is quadratic[2], with a variable added in for a simple adjustment by sex. The model was fitted using ordinary least squares (OLS) regression on a pooled dataset drawn from the 2003 and 2006 rounds of the HSE (n = 26,679). Despite its simplicity and the lack of detail in the paper, this model (or versions of it) has on occasion been used in HTA submissions to NICE when general population baseline HSUVs have been required (Hill et al., 2019; NICE, 2013a, 2013b, 2021). For this reason, I refer to the quadratic model as the status quo approach to baseline utility adjustment by age.

A more complex parametric model is fitted in a paper by Hernández Alava et al. (2022) using the 2014 round of the HSE (n = 7,085). Adjusted limited dependent variable mixture models (ALDVMMs) (Hernández Alava, Wailoo and Ara, 2012) are fitted separately for each sex. This more sophisticated model was designed to correct for

[2] $EQ5D = \beta_0 + \beta_1 \cdot male + \beta_2 \cdot age + \beta_3 \cdot age^2$

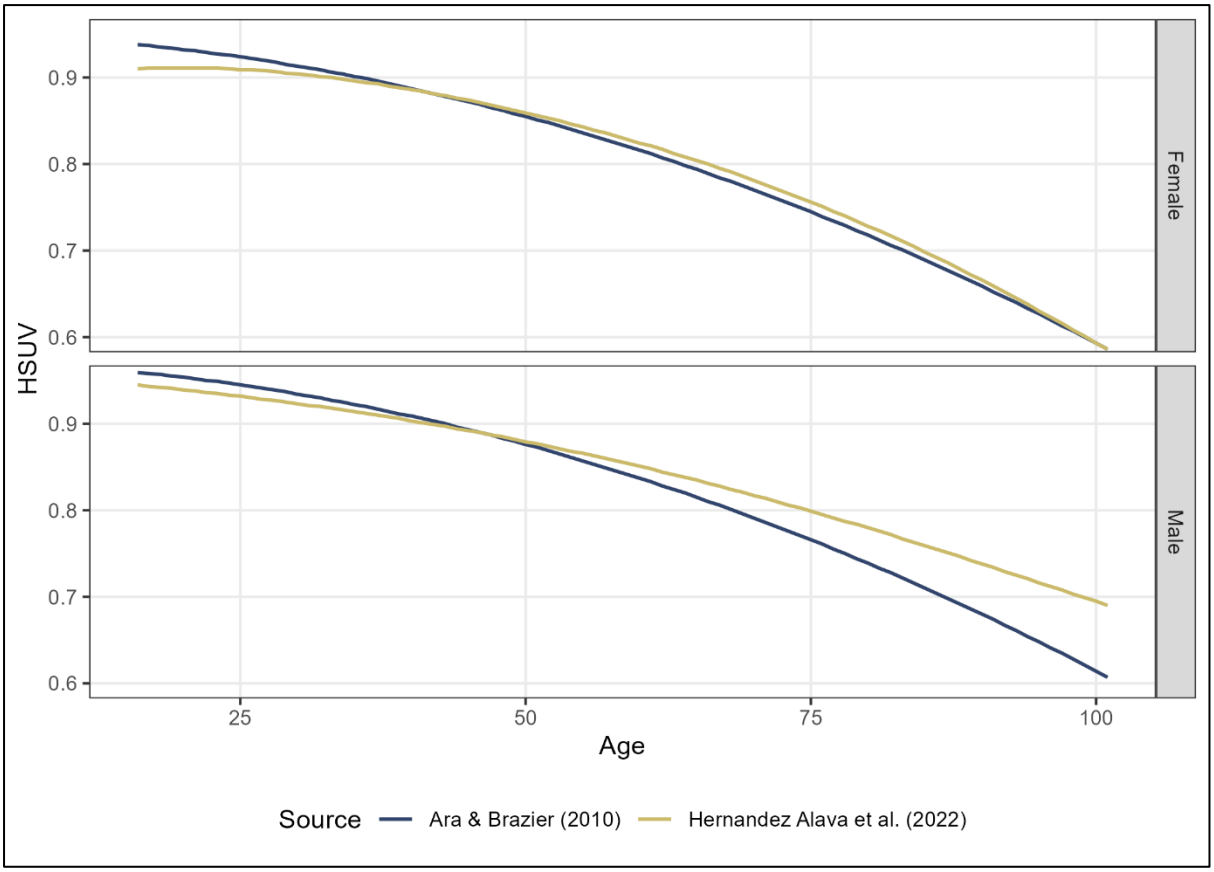

Figure 2: Published parametric baseline HSUVs

the unusual distribution of EQ-5D-3L data, specifically multi-modality and limits in range. It works by combining a number of tobit-style models, the membership of which is determined by a separate multinomial logit model. In this instance, the authors prefer three-component models with age and age squared as variables in each component and age as a variable in the component membership model. The ALDVMMs were fitted by maximum likelihood estimation.

The predictions of both the ALDVMM and the OLS model are visualised in Figure 2. Despite the different methods and datasets used, the baselines are fairly similar, with the most significant differences emerging among older males. The differences in results between OLS and ALDVMM approaches were the focus of a 2022 ISPOR EU poster (Chang-Douglass, Bungey and Dakin, 2022). This poster compared the

baselines produced by each model using both the 2014 HSE and the 2003-2014 HSE rounds pooled, finding that in general the baselines produced were very similar.

2.3 Model Selection

Parametric methods are demanding in their requirement for researchers to choose a functional form and a method for fitting their models, and these choices can have a big influence on results. For this reason, it is worth critiquing the modelling decisions in these two parametric-baseline papers and the processes by which they were made.

Both papers' modelling decisions have advantages and disadvantages. The Ara & Brazier model's simplicity is appealing, and the functional form allows for some flexibility through the inclusion of the age-squared term. On the other hand, Hernández Alava et al. run models separately for each sex, allowing for more complex differences between sexes to emerge. ALDVMMs are also more theoretically robust as they properly account for the distribution of the variable they are intended to predict. In addition, they are more flexible with regard to the relationship between age and HSUVs due to the multiple components, each with age and age-squared terms. However, the ALDVMM approach has also been shown to be difficult to fit as the solution to the maximum likelihood procedure is sensitive to both starting parameters and the likelihood optimisation method (Hernández Alava and Wailoo, 2015). This may hamper reproducibility as subtle decisions by researchers can significantly affect results.

It is also worth noting that the two models are trained using different data. Hernández Alava et al. (2022) use data which is more recent but has several times fewer observations than the data of Ara & Brazier (2010). This point about sample size is particularly important for baselines for those over the age of 80 as the HSE sample

sizes in these groups are relatively small. Further analysis is needed to better understand this trade-off between recency and sample size.

Both papers' approaches have advantages and disadvantages that are not easy to resolve, so it is also important to look at the process by which modelling decisions were made. As already mentioned, the Ara & Brazier (2010) model was not the focus of the paper, and there is accordingly no mention of why the variables or functional form were chosen. Hernández Alava et al. (2022) are more transparent – the component membership model is chosen based on improvements in Akaike and Bayesian information criteria, and four-component models were rejected after signs of overfitting emerged. However, the choice of the quadratic form for the component models is not justified, and the optimisation approach is not reported.

But what should this process of model selection look like? To answer this question, it is necessary to take a step back and think about predictive modelling in comparison to explanatory modelling, the latter of which is far more prevalent in economics (Shmueli, 2010). Explanatory modelling aims to identify a causal link between specified factors and an outcome, whereas predictive modelling aims only to identify the mathematical function that best relates factors to an outcome.

These differences in aims have important implications for the choice of variables and models (Sainani, 2014). Explanatory modelling begins with theory, which is used to specify a null hypothesis. Models are then designed such that the null hypothesis can be empirically tested, given that the relevant assumptions for the model hold. Choosing a model using measures of performance may play some role, but the testability of the null hypothesis is crucial. Over-using the data may even be considered a form of manipulation if researchers run a series of models with slightly

different variables/features and simply choose whichever has the most desirable results.

In contrast, the functional forms of predictive models are not restricted by the need for meaningful hypothesis testing, and the choice of model ought to be fundamentally data-led. Additionally, predictive modellers ought to be wary of in-sample performance metrics like the R^2 as these may be blind to overfitting. Instead, performance should be assessed using data that the model has not 'seen'. Theory plays only a small role in model selection, restricted to contexts where there is limited data, for example.

The data-led principles of predictive modelling have fostered an explosion in machine learning techniques in recent decades (Sammut et al., 2010; Steyerberg, 2019), moving far beyond the relatively simple models generally applied in explanatory modelling (Shmueli, 2010; Taagepera, 2008). In health economics, the literature on HSUV mapping largely falls into the predictive modelling category (e.g. Gray, Wailoo and Hernández Alava, 2018) and there have been numerous applications within health sciences more generally (e.g. Chekroud et al., 2021). However, the same cannot be said of the very limited literature on baseline HSUVs. There has been a lack of model validation using 'unseen' data, and there has been limited attention given to more flexible models for the relationship between age and HSUVs.

In the last 20 years, significant evidence has emerged in the psychology literature that there may be a midlife low in mental well-being, creating a U-shaped relationship between health and age (Blanchflower and Oswald, 2019). If this trend were to generalise across other dimensions of health and to less specialised/sensitive patient-reported outcome measures (PROMs) like the EQ-5D-3L, then a similar relationship might emerge in HSUV baselines if sufficiently flexible models were fitted. Guthrie et

al. (2023) implement such a model in a population with no history of heart disease, finding that a 5^{th}-degree polynomial function was preferred by AIC. While this finding may not transfer to the general population, it certainly motivates further exploration of this issue.

If the purpose of modelling was solely explanatory then flexibility would not be a significant issue – a monotonic, concave relationship is consistent with common sense intuitions (i.e. theory) about the deterioration of quality of life with age and is well suited to hypothesis testing. This may explain some of the choices in previous research – from Kind et al. (1999) ensuring monotonicity by using large age bands, to Ara & Brazier (2010) employing a quadratic function to ensure that such a relationship is produced. However, since the function of baseline models is to predict, the degree of flexibility of these models ought to be determined by data as far as possible. In this paper, I will empirically study the degree of flexibility that the data implies in the relationship between age, sex and HSUVs.

3 Data

3.1 The Health Survey for England

This study uses data from the Health Survey for England (HSE) (NatCen Social Research, 2023), a repeated cross-sectional dataset collected yearly where participants from the general population are asked questions about their health. The specific questions vary somewhat year by year as researchers focus on different issues. The three-level version of the EQ-5D survey has been conducted on and off as part of the HSE, making the dataset a useful tool for HSUV-related analysis. As we have already seen, it has already been used several times to estimate baseline utilities.

I accessed the HSE dataset from the UK Data Service website for every year that the EQ-5D-3L questionnaire was conducted (2003-2006, 2008, 2010-2012, and 2014). All manipulation and analysis of this dataset was conducted using the R programming language (v4.3.1, R Core Team, 2023), and the code used can be found at https://github.com/tomclemmet/baseline-utilities. Observations with missing values for relevant variables were removed, following the approach taken elsewhere in the literature (e.g. Ara and Brazier, 2010; Hernández Alava et al., 2022). The 'eq5d' package (Morton and Nijjar, 2023) was used to calculate HSUVs from responses to the survey using NICE's approved value set (Dolan, 1997).

While the HSE has been used several times in the HSUV literature, authors have differed in their choice of which round to use. Higher sample sizes give more confidence to results, and this problem is particularly critical in the older population where the number of observations tends to be lower. The more rounds of the HSE that are pooled the greater the sample size, but the older the data used. This is why authors

Year	Sample Size			Age		HSUVs	
	Total	Over 75s	Over 90s	Mean	Std. Dev.	Mean	Std. Dev.
2003	13,753	1,219	63	47.7	18.2	0.861	0.223
2004	6,114	593	26	48.9	18.2	0.851	0.235
2005[3]	9,211	1,601	87	54.5	19.8	0.840	0.237
2006	12,926	1,250	73	48.9	18.3	0.857	0.232
2008	14,113	1,430	81	48.8	18.6	0.851	0.235
2010	7,332	746	44	49.3	18.4	0.849	0.230
2011	7,517	727	37	49.1	18.3	0.825	0.244
2012	7,294	740	49	49.8	18.5	0.853	0.234
2014[4]	7,085	694	40	49.7	18.3	0.856	0.232

Table 1: Summary statistics by HSE round (2003-2014)

have used different samples for their analysis, with the NICE DSU view (Hernández Alava et al., 2022) being that the most recent HSE round with the EQ-5D-3L (currently 2014) should be used. Given the importance and difficulty of achieving reasonable sample sizes in the older population, this recommendation ought to be properly tested. Table 1 reports summary statistics for all HSE rounds where the EQ-5D-3L has been collected since 2000. From this, we can see that each round has roughly similar characteristics[3], and in particular, there is no obvious improvement in mean HSUVs over time as one might expect.

3.2 Choosing the Rounds

To investigate the detrimental effects of using non-recent data (as discussed in section 2.3), I compared the performance of models using older and newer data in the newest HSE round where the exact age of all subjects is published – 2012. The 2003 round was used as the older sample (n = 13,753) and the 2010/11 rounds pooled were the newer sample (n = 14,811). Figure 3 plots the mean HSUVs across different ages and

[3] An over 65s sample boost for the 2005 round gives this data different characteristics.
[4] In the 2014 dataset age is top-coded at 90 to preserve the anonymity of participants. Following the approach of Hernández Alava et al. (2022) I assigned the average age of those aged over 90 in the 2013 round (92.5) to those aged over 90 in 2014.

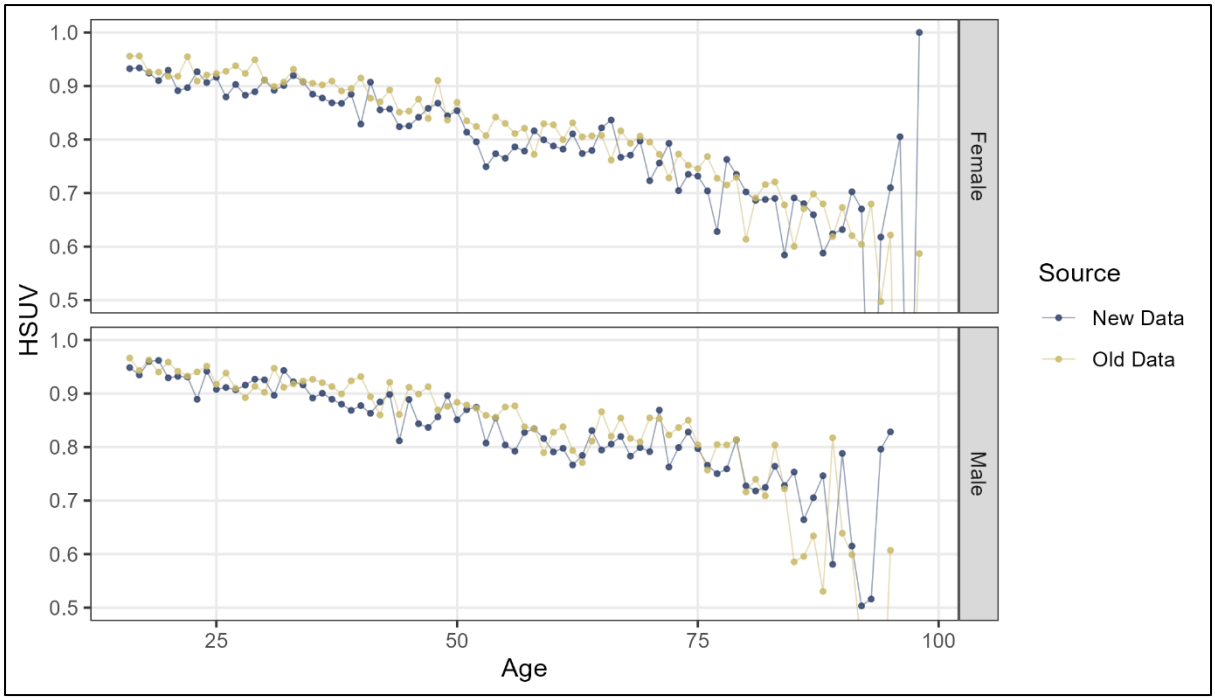

Figure 3: Mean HSUVs from Older and Newer HSE Rounds

Source	RMSE	MAE	ME
New	0.226515	0.158606	0.018282
Old	0.225935	0.153316	-0.001143

Table 2: Error statistics for models trained on old and new data

sexes for each sample, and from visual inspection there is no clear difference between the two. I fitted simple quadratic OLS models to the old and new data, running models separately for each sex (details of which can be found in Appendix A1). I then used these models to predict HSUVs in the 2012 dataset and calculated error statistics[5] (Table 2).

Interestingly, across all three error statistics, the model based on older data outperforms the model based on newer data. Despite being a decade out-of-date, the 'old' model has a lower RMSE and MAE, and its ME is several times smaller in magnitude. In fact, the positive ME means that the 'new' model tends to over-estimate HSUVs – the opposite of what we might expect if there was a long-term trend in public

[5] RMSE = root mean square error, MAE = mean absolute error, ME = mean error.

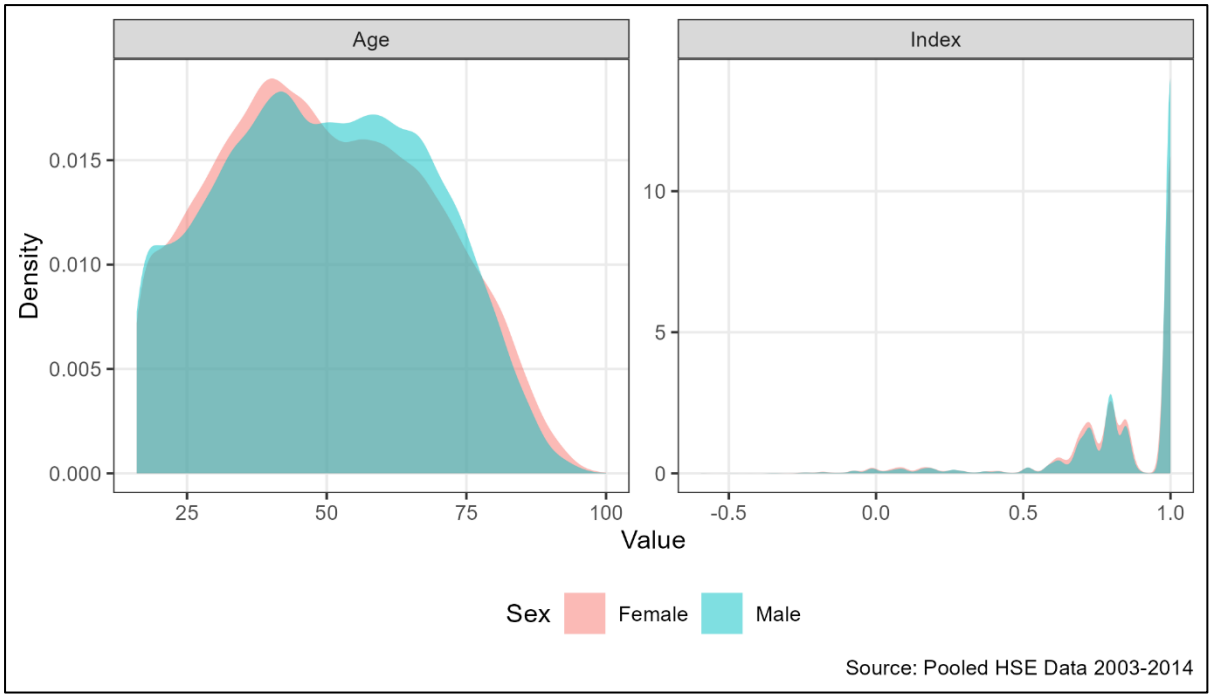

Figure 4: Kernel density plots for age and HSUVs

health which made older data unsuitable. These results fail to prove that using older HSE rounds would be detrimental to the analysis, and their counterintuitive nature suggests that they may be a product of random variation.

3.3 The Pooled Dataset

Following the results of this preliminary analysis, for this project I chose to pool all of the rounds of the HSE from 2003 to 2014 (n = 85,245). This sample has been used in research before (Chang-Douglass, Bungey and Dakin, 2022), and the concerns about recency in pooling such a large dataset seem to be outweighed by the benefits of maximising the sample size in older ages. For comparison, the 2014-only sample used by Hernández Alava et al. (2022) has only 40 subjects aged over 90, but the pooled sample has 366, giving more confidence to the baselines produced for this range. However, this number remains small relative to the overall sample size. Figure 4 illustrates this point by showing the distribution of ages in the pooled dataset, showing the rapid drop-off in sample sizes in the older ages. It also reports the distribution of

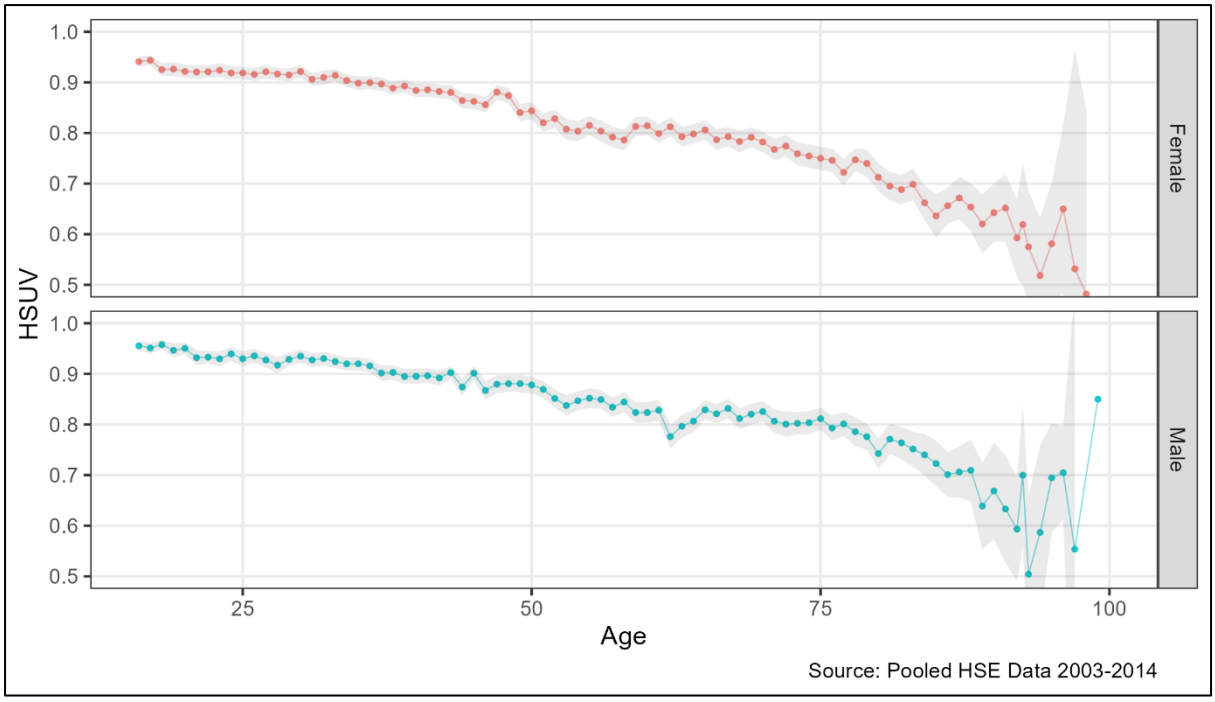

Figure 5: Mean HSUVs from the pooled dataset with 95% confidence intervals

the HSUVs, which features the usual peaks and troughs that the ALDVMM approach was built to solve.

Figure 5 plots the mean HSUVs, along with 95% confidence intervals, across different ages, with the plot separated by sex. The benefit of the large dataset is that despite the large unexplained variation in the data, confidence intervals are fairly small for the majority of ages (up to approximately 90). While the relationship between HSUVs and ages is fairly linear, there may be subtle non-linearities which ought to be appreciated. Firstly, there is a sharper decline in old age, which previous quadratic models have taken account of. Secondly, there is a subtle pattern of declines and plateaus of HSUVs in middle age. This second feature has not been appreciated by previous baseline models, though it may be consistent with results elsewhere in life sciences (Blanchflower and Oswald, 2019). This demonstrates the need for a robust study of more flexible age/HSUV models which test whether such models' ability to capture these nonlinearities is useful.

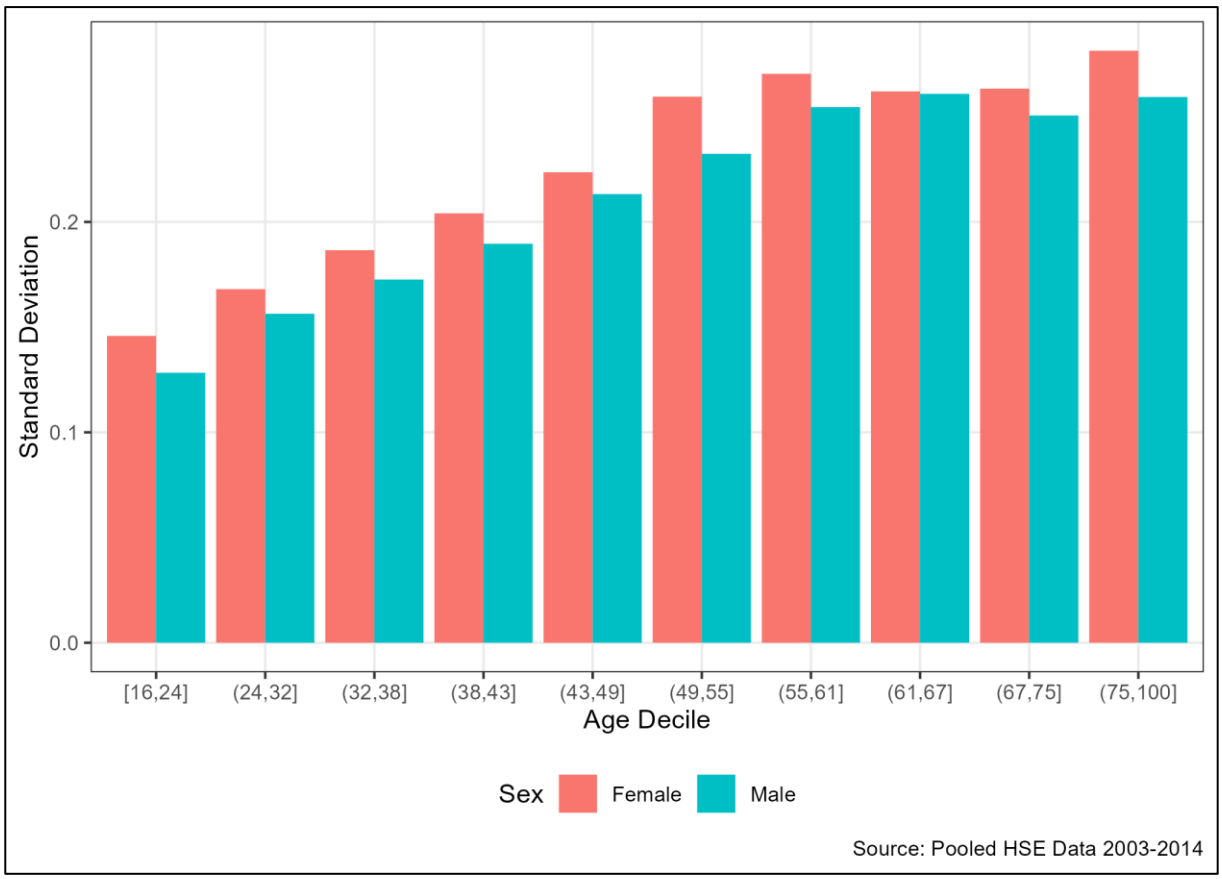

Figure 6: Standard deviations of HSUVs for each age decile

One final point to make about the sample in the context of age- and sex-adjusted baseline HSUVs is that there is a significant amount of variation in the data which cannot be explained by these two variables alone. This is illustrated by Figure 6, which shows the data's standard deviation in different age deciles for each sex. It makes sense that this is the case – health-related quality of life might systematically differ based on income, ethnicity or geographical location, for example, and there is a significant random element involved. This means that a predictive model based only on age and sex will inevitably be underpowered (the following section will return to this point and its implications).

4 Methods

4.1 Models

My approach to drawing methodological conclusions about baseline HSUVs was to select a series of models, specify and implement a testing procedure, and interpret the results. This section will discuss the models chosen to test. The approach to model selection had two main objectives: first, to replicate the methods already used in the literature, and second, to allow for differing degrees of model flexibility. By re-applying existing methods using the pooled HSE data, it is possible to compare novel models to existing ones with the biases of differing datasets removed, and by including more flexible models we can see if the evidence supports a more flexible continuous predictor for HSUVs based on age.

One crucial method to include is the ALDVMM – the most recent recommendation by the NICE DSU (Hernández Alava et al., 2022) for constructing baseline HSUVs. The approach is a significant contribution to EQ-5D-3L modelling and is particularly suited to mapping applications where there is a growing body of literature supporting its use (Gray, Wailoo and Hernández Alava, 2018). However, its usefulness in predicting baselines, which are expected HSUVs and therefore will not follow the same distribution as the EQ-5D-3L, is not well understood. The complexity of ALDVMMs mean that a full analysis of their capabilities is beyond the scope and complexity of this project, but I will nonetheless include it as a model in my analysis. I replicated the DSU model as closely as possible, using the ‘aldvmm’ R package (Pletscher, 2023) to estimate a three-component model with age and age squared in each component model and age in the component membership model.

I also included a range of polynomial functions. These are linear functions whose terms are a given variable raised to increasing powers. More formally, in this paper I use the term 'nth-degree polynomial function' to mean the function $f(x, n) = \sum_{k=0}^{n} \beta_k x^k$. For example, a 2nd-degree polynomial function is $f(x, 2) = \beta_0 + \beta_1 x + \beta_2 x^2$. Polynomial functions can be fitted straightforwardly by OLS as they are linear in their parameters. As the order of the highest polynomial increases these functions allow for more and more flexibility, and the curve fitted becomes more and more 'wiggly' (Steyerberg, 2019, p.180). These functions are the classical approach to non-linear continuous predictors (e.g. Grambsch and O'Brien, 1991), making them an obvious choice for this analysis. Including a range of polynomial functions allows for the analysis of increased flexibility, as well as replicating work previously done (Ara and Brazier, 2010; Guthrie et al., 2023). I included polynomial functions for $k \in [0, 10]$.

While polynomial functions are a widely used and well-understood tool, higher-degree polynomials may be prone to overfitting if sample sizes are small, particularly in the tails of a distribution (Lever, Krzywinski and Altman, 2016). To an extent, this overfitting is what I am interested in testing, but if models are highly unstable in the tails then this may limit the degree of flexibility that can be explored in other parts of the distribution.

Another, more sophisticated, class of continuous non-linear predictor is restricted cubic spline functions (RCSs) (see Croxford, 2016), an approach that has its origins in a clinical prediction context (Harrell, Lee and Pollock, 1988) and is easy to implement in R (via Harrell, 2023). Spline functions involve fitting curves individually between specified points, or 'knots', in the distribution of the explanatory variable such that the curves are continuous at each knot. In the RCS case these functions are cubic – the simplest polynomial form to allow for a 'kink' in the curve – but the 'restricted'

element of RCSs is that linear functions, rather than cubic, are fitted at the tails of the distribution. This final feature makes RCSs a useful complement to simple polynomial functions by guaranteeing reasonably stable behaviour in the tails of the age distribution, allowing for very flexible models to be considered without producing unusual results. At the time of writing, I am not aware of any previous attempts to use RCS functions in this specific context, though they have been used in other health economics settings (for example, to model progression-free survival in Gibson et al., 2017).

I fitted RCS functions with a range of knots (from 3 to 10 inclusive) to allow for different degrees of model flexibility to be considered. The positions of the knots were based on quantiles of age[6], following Harrel (2015).

4.2 Cross-Validation

Previous work on predictive modelling for baseline HSUVs has lacked model validation using 'unseen' data. In-sample performance metrics may prescribe overly complex, overfitted models, and even parameter-penalising metrics like AIC and adjusted R^2 may not fully correct for this. Overfitted models tend to produce inaccurate out-of-sample predictions, making them unsuitable for implementation. By using unseen data researchers can simulate a model's real-world application, allowing for a true assessment of a model's capabilities.

There is a rich literature on approaches to validating prediction models (see Steyerberg, 2019, pp.329–342). The optimal approach is external validation, where a

[6] For 3 knots, quantiles for knots were 0.1, 0.5, and 0.9. For 4-6 knots the upper and lower quantiles were 0.05 and 0.95 with the area between divided equally between the remaining knots. For > 6 knots the same rule applied, but with upper and lower knots at 0.025 and 0.975.

model is tested using a different dataset collected from a plausibly related population. This provides a very realistic (and demanding) simulation of a model's real-world usage, making it an excellent test for overfitting. However, in practice it may be difficult to find a suitable dataset to use.

In this study, I take an internal validation approach. Internal validation involves using the same dataset to test and train a model, but not necessarily the same sample (Steyerberg, 2019, pp.331–335). A simple version of this would simply be to split the dataset into separate training and testing sets, but this approach makes any results highly dependent on the samples chosen. One popular approach that largely resolves this issue is k-fold cross-validation (Geisser, 1975). In k-fold cross-validation the data is split into k random folds, the model is trained using all but one of the folds, and it is then assessed on its ability to predict values in the final fold. This process is repeated for each combination of folds, meaning that every observation has at one point been in both the training and testing samples. The case where k = n is called 'leave one out cross validation' or 'the jackknife', but in practice values of k = 10 or k = 5 are frequently used for computational ease and to reduce the variance of results across folds (James et al., 2013, pp.181–184).

There are many extensions and variations on k-fold cross-validation methods (see Arlot and Celisse, 2010), but for this dissertation I will use the standard form where k = 10, with a minor adjustment. Previous sections have identified the problem that the HSE data has very few observations for older people. This means that when splitting the sample, there would be a worry that some 'folds' contain very few people from these underrepresented ages. A solution for this problem is stratified cross-validation (Sammut et al., 2010, p.928), where the folds are stratified random samples so that the distribution of ages in each fold is similar to the distribution in the original dataset.

I conducted stratified 10-fold cross-validation, stratifying age by decile, with the entire procedure carried out separately for each sex.

The majority of models discussed in section 4.1 lend themselves well to k-fold cross-validation as they can be easily adjusted using a single parameter (degree of polynomial or number of knots) and are straightforward to fit. However, ALDVMMs can be adjusted in several different fundamental ways (e.g. number of components, component models, membership model) and have a likelihood function that is challenging to maximise. Hernández Alava and Wailoo (2015) note this in the documentation for the model's Stata command, recommending that researchers fit a range of models using different components/optimisation methods/starting values and choose whichever model performs best. They warn that arbitrarily choosing these settings might result in a poorly fitted model or the likelihood function failing to maximise. However, conducting this comprehensive trial-and-error procedure would be very computationally intensive within k-fold cross-validation, as every model requires fitting in 20 times (10 folds for each sex).

To avoid this, I tested a variety of settings for the ALDVMM using the whole dataset and compared the results in terms of AIC. I kept the three-component structure used by the DSU model and tested the available optimisation methods using several starting values, both random and specifically chosen. I found that across both sexes the best ALDVMM fits were given by using the 'nlminb' optimisation method (representing PORT optimisation routines, see Gay, 1990) and setting the starting parameters to similar values to the Ara & Brazier (2010) quadratic model[7].

[7] In each component model: $\boldsymbol{X\beta} = 1 - 0.01 \cdot Age - 0.001 \cdot Age^2$, $\ln(\sigma) = -2$. In the component membership model all parameters $= 0$ (i.e. equal probability of membership).

These decisions meant that the ALDVMM could be included in the analysis, but they were far from optimal. Each training set will have subtly different characteristics, perhaps meaning that different optimisation/starting value decisions may produce an improved fit. The analysis is also limited by only considering the DSU approach – models with a different functional form or a different number of components may have been more appropriate in some training sets. As a result, all ALDVMM results ought to be treated with caution.

4.3 Interpretation

The cross-validation procedure's raw output is a set of cross-validation iterations, each with a set of prediction errors. There are many different ways of quantifying model performance depending on the specific context, but I draw on recent applications of the k-fold cross-validation method in the field of HSUVs for my approach. Thompson et al. (2019) conducted a study of different methods for predicting HSUVs using subjects' reported (combinations of) long-term conditions. They assess model performance using the root mean square error (RMSE) and the mean absolute error (MAE), both measures combining accuracy and precision, with RMSE giving more weight to larger errors than MAE. They also calculate the mean error (ME) which is simply an estimation of model bias and can be positive or negative, indicating whether the model tends to under- or over-predict HSUVs. All three metrics are widely used, applicable to a wide range of models and well understood (Walther and Moore, 2005). RMSEs/MAEs/MEs can be calculated for individual folds, and then the mean can be taken to produce an overall measure of performance.

It is sometimes difficult to know how to interpret these different error statistics. For example, Thompson et al. (2019) use a degree of judgment in their interpretation,

weighing up models' strengths and weaknesses to reach a conclusion. On the other hand, Davison et al. (2018) formalise model selection by proposing a ranking algorithm that combines the results from all three error statistics[8]. This approach allows no freedom for interpretation, but it may miss nuances in the data. For example, the method forces a ranking onto models even if the differences between them are very small and/or not significant, which does not seem desirable. In this specific paper, the authors also make more informal arguments and conduct sensitivity analysis for their algorithm weighting to justify their results, but these two HSUV-based papers nevertheless illustrate a dichotomy in approaches to interpreting these kinds of results.

Before deciding how formal an approach to take to model selection, it is worth noting that there are a couple of difficulties with this specific context of baseline utilities which give reason for caution. The first problem is that all the models are underpowered, no matter how complex they are with respect to age. This is because there is a large amount of variation in the data which cannot be explained by age, as shown in Figure 6. The implication of this is that all models will perform fairly badly in terms of RMSE/ MAE, and the differences in these statistics between models may be small even if their predictions look quite different. With this problem in mind, it would be very useful to have some idea of the distribution of the RMSE and MAE so that small but significant differences could be distinguished from those that are random. However, this is not straightforward for k-fold cross-validation, with Bengio & Grandvalet (2004) showing that there is no unbiased estimator for the variance of k-fold results. There is an active literature looking at estimators for k-fold standard errors (e.g. Yousef, 2021), but in the

[8] Each model is given a ranking for RMSE, MAE and ME respectively. The number-rank from each statistic is added up to form an overall score for each model (with a lower score being better). This final score is used to create an overall model ranking.

absence of a consensus I will not report standard errors for results. This is the approach usually taken in k-fold cross-validation papers (e.g. Jung et al., 2020).

These difficulties with interpretation suggest that nuanced judgment might be required in the interpretation of results. For this reason, I will not try to formalise model selection using an algorithm, rather relying on a more holistic analysis of each model's strengths and weaknesses. Taking this approach may mean that this project fails to clearly recommend a specific model, or that another researcher might draw a different conclusion from the same results. However, these limitations would not be solved by arbitrarily specifying a more formal model selection criteria for a complex range of results.

I have discussed the difficulties with *statistical* significance in this analysis: the concept of one result being genuinely different to another rather than a product of random variation. However, another important issue is *qualitative* significance: the concept of one result being meaningfully different to another. While applications of baseline utilities will vary depending on the context, it would be useful to explore whether different baseline modelling choices are likely to affect whether or not a treatment is deemed cost-effective. Hernández Alava et al. (2022) conduct such an analysis, generally finding that the differences between baselines have little effect. However, while their analysis benefits from access to confidential HTA submissions, the baselines they compare use different datasets *and* different methods, making interpretation difficult.

I will conduct a similar analysis by comparing the baselines produced using different models but trained using the same pooled HSE dataset. While I do not have access to confidential HTA submissions to 'try out' different baselines, it is still possible to

produce quality-adjusted life expectancy (QALE) estimates for different ages/sexes, following the example of the Schneider et al. (2023) R Shiny application. QALEs are useful when calculating absolute/proportional QALY shortfalls, or in any other context where a now-healthy patient's remaining QALYs are required. The QALE estimates were produced using ONS life tables from 2018-2020 (Office for National Statistics, 2021). For each age-sex pair the survival rate in each remaining year of life was multiplied by the model-in-question's baseline utility and then discounted by the NICE discount rate of 3.5% (NICE, 2022b, p.81). A 0% discount rate was also applied for comparison. The benefit of this analysis is that more complex models may perform better but produce similar enough baselines to simpler models that the choice between the two is inconsequential in practice.

4.4 Key Questions

To help structure the interpretation of results, I will finish this section by specifying three key questions that the results of this analysis will help to inform (based on the issues in the literature discussed in section 2).

The first question is "Are ALDVMMs a useful tool for capturing the relationship between utilities and age?". Previous research has shown ALDVMM predictions to be similar to quadratic OLS models (Chang-Douglass, Bungey and Dakin, 2022), though the published predictions of the Hernández Alava et al. (2022) model are significantly different from that of Ara & Brazier (2010) for older men[9]. Performing k-fold cross-validation means that the two models are repeatedly trained on the same dataset with

[9] Although in this case the ALDVMM was trained on the 2014 HSE, and the quadratic OLS model was trained on the 2003/06 HSE

a different subsample being used each time, allowing for more robust conclusions to be drawn about the similarity of the two methods.

The second question is “Do RCS functions outperform polynomial functions?”. As has already been discussed, higher-order polynomial functions may perform badly in the tails of the age distribution, but RCS functions allow for a similar degree of flexibility while guaranteeing linear behaviour in the tails. K-fold cross-validation might reveal whether this adjustment gives RCS predictors improved out-of-sample prediction accuracy.

The final, and perhaps most important, question is “What degree of model flexibility is optimal?”. This is something which has not been studied so far in the literature, and which the k-fold cross-validation approach is well suited to studying through the use of ‘unseen’ data. By testing a range of polynomial and RCS models the cross-validation procedure should reveal at what point increased model flexibility is no longer beneficial.

5 Results

5.1 Visualisations

The pooled HSE dataset was separated by sex, and then each half was randomly divided into 10 stratified folds, yielding 20 test/training data splits for cross-validation. For each split 20 models were run: polynomial functions from 0^{th} (i.e. the sample mean) to 10^{th} degree (labelled Poly-#), RCS functions with 3-10 knots (labelled RCS-#) and an ALDVMM.

Figure 7 shows the predictions of each model across each of the folds, clearly illustrating the properties of each model. One point that can immediately be made is that for both the polynomial and RCS models the curves become more 'wiggly' as the degree of complexity (either order of polynomial or number of knots) increases. This is what we would expect, and shows that the procedure has worked as intended. For the most part, predictions are very similar across folds, which is to be expected given that the training sets have large sample sizes of 33,912 and 42,898 for men and women respectively, but there are two key exceptions.

First, the ALDVMM is noticeably more erratic than other models in its predictions for the oldest and youngest in the sample. This demonstrates the concern about using this model in cross-validation as the optimal model-fitting approach may differ across folds, so taking a one-size-fits-all approach as I have done may fail to produce suitable predictions in some cases. While there are occasions where the ALDVMM model roughly matches the quadratic model (consistent with Chang-Douglass, Bungey and Dakin, 2022), in other cases it does not, and it is difficult to say whether these differences are genuine or a result of poor fitting. Second, there are erratic predictions

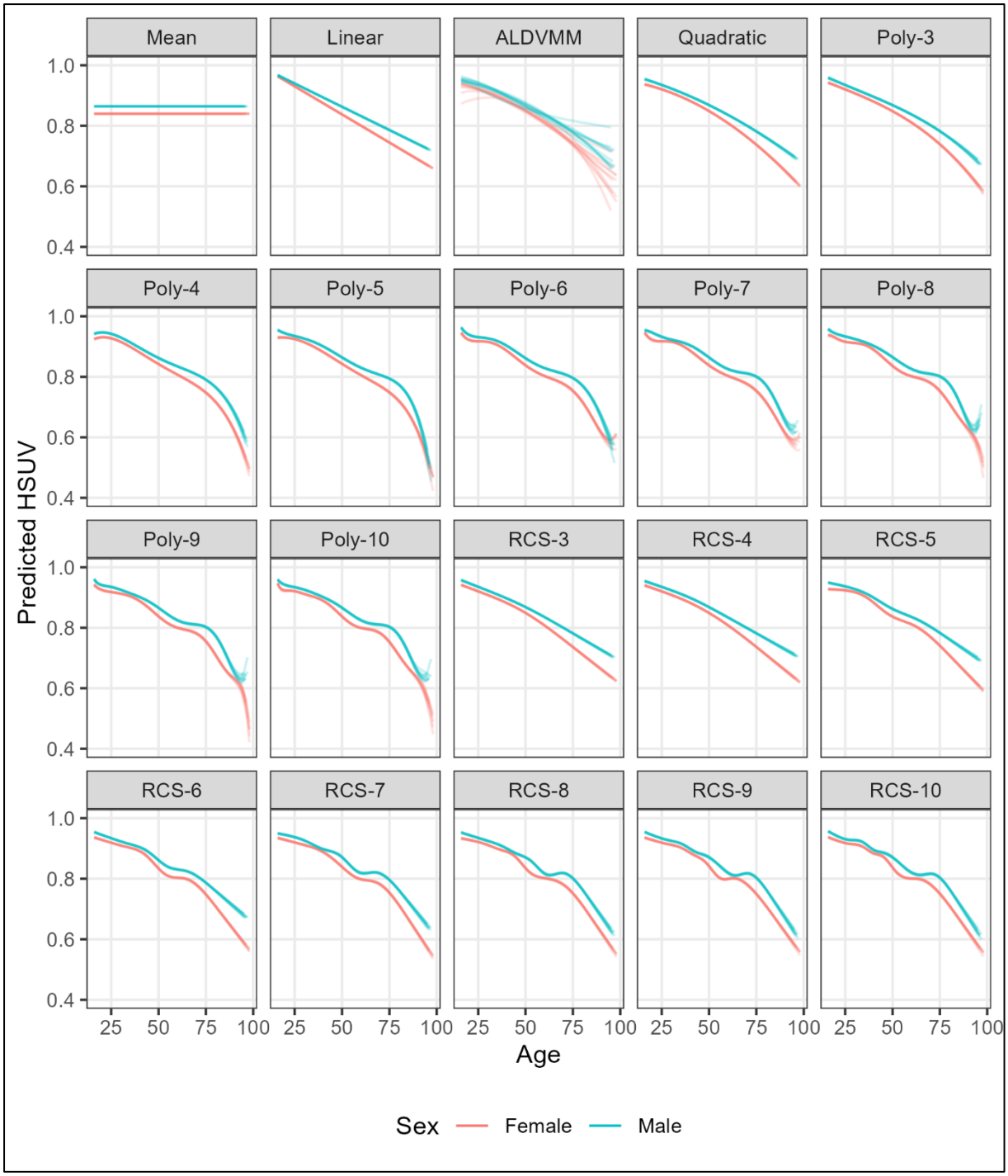

Figure 7: Visualisation of model predictions across all folds

for the very oldest in the higher-order (6th and above) polynomial models. The small sample size in this group means that there may be significant variation across training sets which these models pick up, leading to unstable behaviour. These predictions are not a plausible representation of quality of life, making these functions unsuitable candidates for predictors. 6th-order and above polynomial models are therefore taken

no further in this analysis. Note that, by design, the RCS models face no such problems, allowing for greater flexibility to be tested elsewhere in the distribution without any unusual results in the tails.

5.2 RMSEs and MAEs

MEs, RMSEs and MAEs were calculated for each model for each fold, and then averaged across the folds to produce an overall measure of performance. The full results for all three statistics are presented in Appendix B1, and Figure 8 plots the RMSEs across all the remaining models, using the linear model as a baseline. The RMSE results are encouraging for the analysis as although the differences are small, clear patterns emerge relative to model complexity. This is crucial as it suggests that despite the small magnitude of the differences between model RMSEs, these differences are statistically significant and not a product of random chance.

Starting with the ALDVMM, the RMSE of this model fails to improve upon the Linear model and falls significantly behind the quadratic model which has previously produced similar predictions. While these results may not be robust, they do highlight the poor reproducibility of ALDVMMs in this context.

Looking at the other polynomial models, there is a rough downward trend in RMSEs as the degree of polynomial increases, with a quintic form model performing best for this group. This downward trend is also seen in the RCS model, though the RMSE

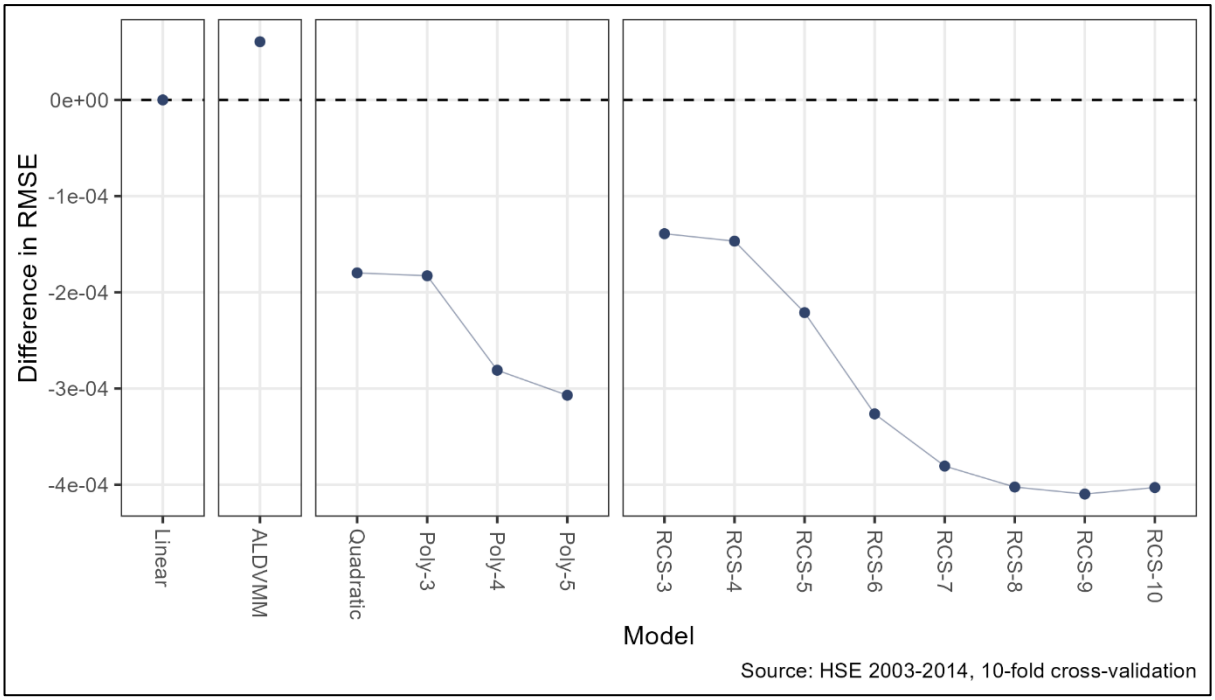

Figure 8: Mean RMSEs across all folds relative to the linear model

plateaus for models with more than approximately 7 knots. The RCS-7 model's RMSE is 2.01E-04 lower than for the literature-preferred quadratic model, meaning that it performs twice as well relative to the linear model. Higher knot RCS models show subtle indications of overfitting as the 10-knot model has a slightly higher RMSE than the 9-knot model, but it is difficult to say if this is significant. Applying the common-sense idea that all else being equal a simpler model ought to be preferred, the RMSE analysis suggests that the 7-knot RCS model is optimal for modelling HSUV baselines.

The overall RMSE analysis is useful, but it does hide any differences in accuracy based on age – poor predictive capabilities in older people may not show up. To address this, I conducted subgroup analysis in the age deciles of the data[10], calculating error statistics as previously (see Appendix B2 for full results). Figure 9

[10] Subgroup analysis by sex revealed similar trends to the overall analysis, although the gains in RMSE in the female data were significantly higher than in the male data.

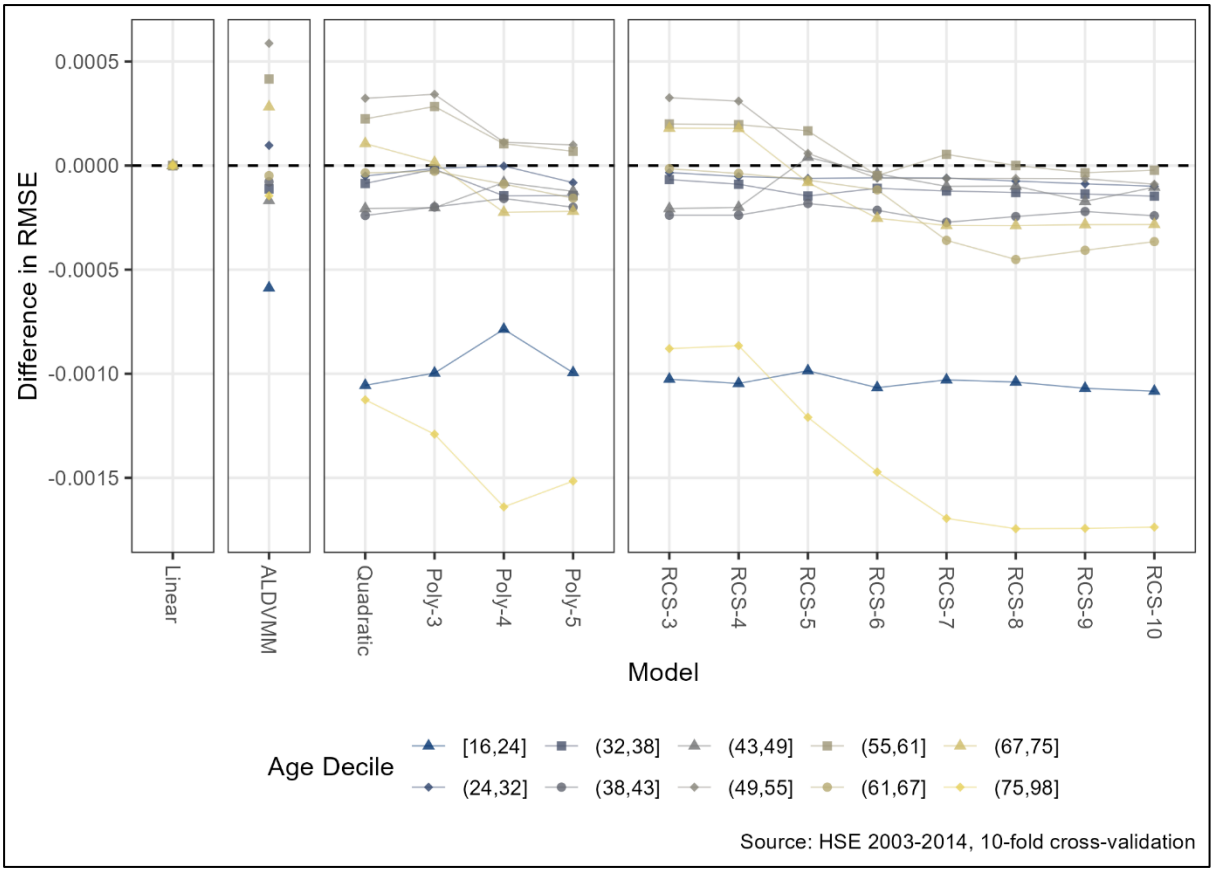

Figure 9: Mean RMSEs across all folds, separated by age decile, relative to the linear model within each decile

shows the difference in RMSEs compared to the linear model in each individual decile[11].

The most notable aspect of this graph is that the more complex models significantly outperform the linear model in the first (ages 16-24) and last (ages 76-98) deciles, and in the case of the oldest decile an RCS model with at least 7 knots is again the most accurate choice. Note that the magnitudes of the differences in this decile are around three times higher than the equivalent magnitudes in Figure 8, with the RCS-7 model outperforming the quadratic by 6.59E-04.

[11] Raw RMSEs differ significantly in each decile due to differences in variation (see Figure 6). Expressing performance as relative to the linear model allows for an easy comparison of RMSEs in each age group.

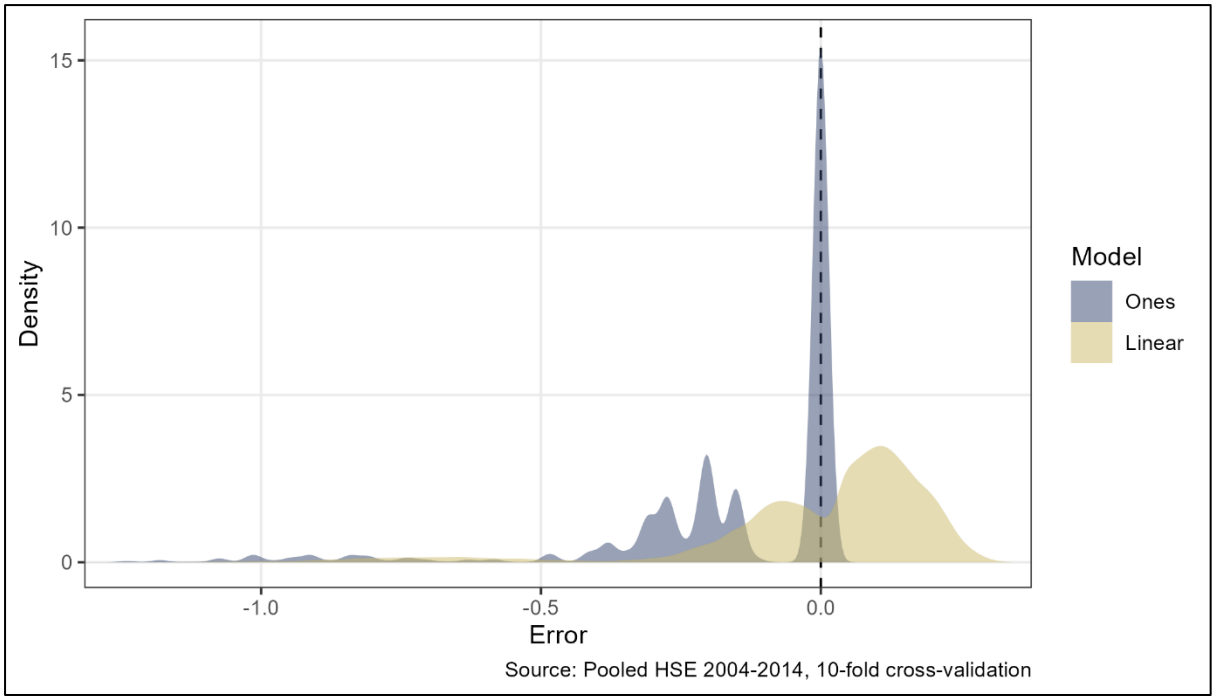

Figure 10: Kernel density plots of prediction errors across all folds for the 'ones' and 'linear' models

The RMSE metric deserves significant attention – it measures both a model's precision and accuracy, and it is closely related to the standard deviation, one of the most fundamental concepts in statistics. However, it is important to see whether the RMSE-based results generalise to similar measures of model performance. The mean absolute error (MAE) is one such metric, differing from the RMSE in that larger errors are not over-weighted.

Unfortunately, the MAE results are difficult to interpret. The linear model has the smallest MAE, firmly rejecting any degree of model complexity. In fact, an even lower MAE 'model' can be constructed by simply predicting perfect health (HSUV = 1) for every subject, no matter their age. This seems counterintuitive. To better understand this result, I created kernel-density plots for the errors of the linear model and the 'ones' model (Figure 10) which describe the error distributions that have resulted in the more complex model being preferred by RMSE but not MAE. From this graph, we can observe that the 'ones' model perfectly predicts the HSUVs for a large proportion

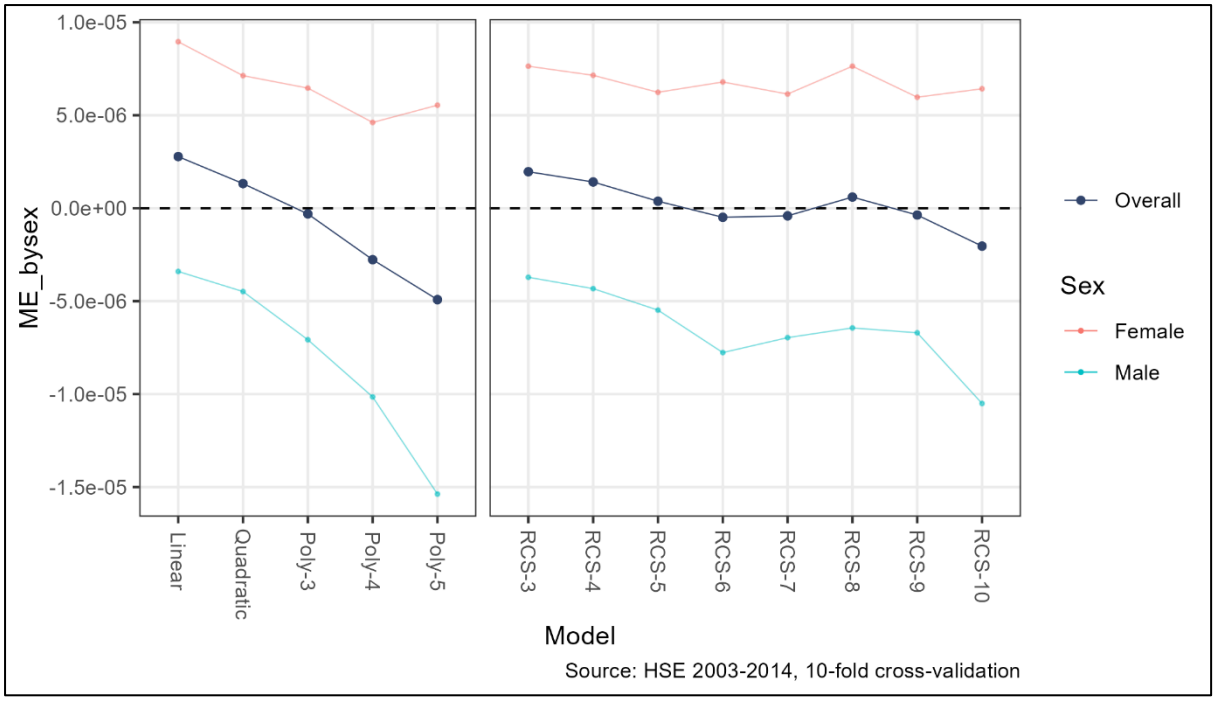

Figure 11: Mean MEs across all folds

of the population but performs poorly when the HSUV is not 1. In contrast, the linear model rarely perfectly predicts HSUVs but is far less likely to produce the larger errors of the 'ones' model. Avoiding larger errors seems the more equitable and generally desirable approach, highlighting the value of RMSE in penalising larger errors. This result suggests that MAE results in the EQ-5D prediction context should be treated with caution due to the large proportion of the population with a score of 1.

5.3 Mean Errors

The mean error (ME) statistic is of interest because it isolates the accuracy element of model performance rather than combining it with precision (like the RMSE/MAE). MEs give a clear indication of whether models tend to over- or under-estimate the outcome of interest, so results have very clear real-world implications. In this case, the mean errors of models are very small – with the exception of the ALDVMM (again highlighting problems with ALDVMM fitting), the MEs are of the order 10^{-6}. Figure 11

plots the overall MEs for all remaining models apart from the ALDVMM, as well as the MEs broken down by sex.

The overall ME results are difficult to interpret. Their magnitudes are qualitatively small, but there do seem to be trends with model complexity, especially for the polynomial models. However, the true nature of these trends can only be understood when also considering the results broken down by sex. These show that the apparent optimality of intermediate flexibility in the overall results is simply a product of taking the average of positive and negative MEs for women and men respectively. As a result, it is difficult to draw a clear conclusion about what these MEs mean.

Fortunately, the MEs are more informative when results are broken down by age. Following the same approach as the RMSE analysis, I conducted a subgroup analysis in age deciles (see Appendix B3), and the results are plotted in Figure 12. The magnitudes of MEs in some deciles were far larger than in the overall case, adding importance to these results. This graph highlights inequality in MEs between age deciles for simpler models, but in contrast to the RMSE results the inaccuracy is not confined to the oldest decile. Instead, simpler models perform worse for those in their 50s and 60s as well as over 75s. As model complexity increases the models perform better for these groups and there is a convergence – the MEs for each decile for the quadratic model have a range of 0.026 compared to 0.00739 for the RCS-7 model. As with the RMSE results, there seems to be little benefit to RCS models with over 7 knots.

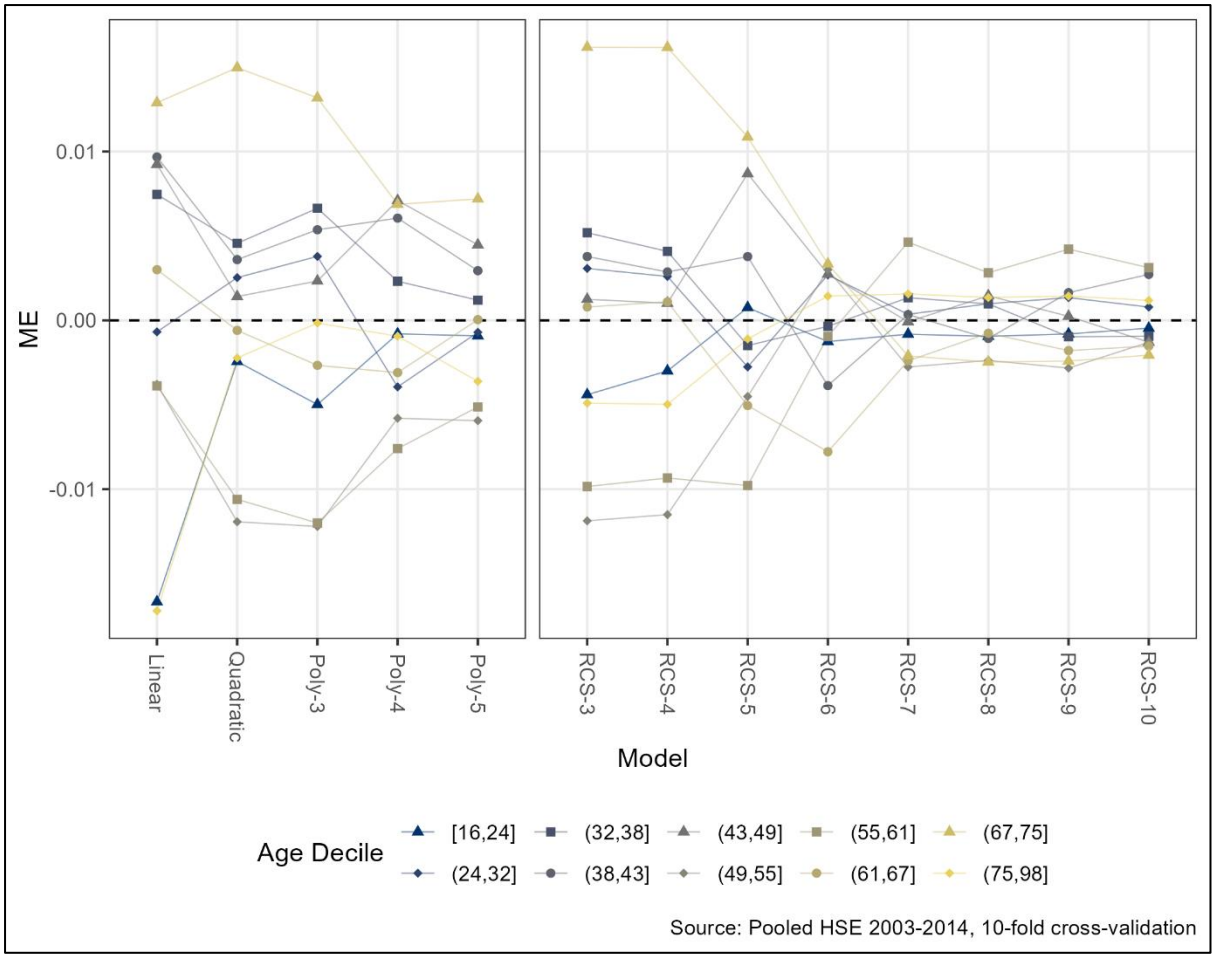

Figure 12: Mean MEs across all folds, separated by age decile, relative to the linear model within each decile

5.4 Qualitative Significance

QALE estimates, as discussed in section 4.3, generally revealed that the differences between the baselines produced by different models were unlikely to be meaningful in practice. Figure 13 illustrates the process of calculating QALEs for 30, 50, 70 and 90-year-old females using either the status quo quadratic model or the RCS-7 model preferred by this study's results, both trained on the pooled HSE dataset. It is clear from these plots that while there are noticeable differences in baselines, the sharp decline in survival observed later in life may render differences in QALE insignificant. When QALEs are calculated for these 4 ages, the magnitude of the differences between the results for the two models were 0.00817, 0.0793, 0.0533 and 0.170

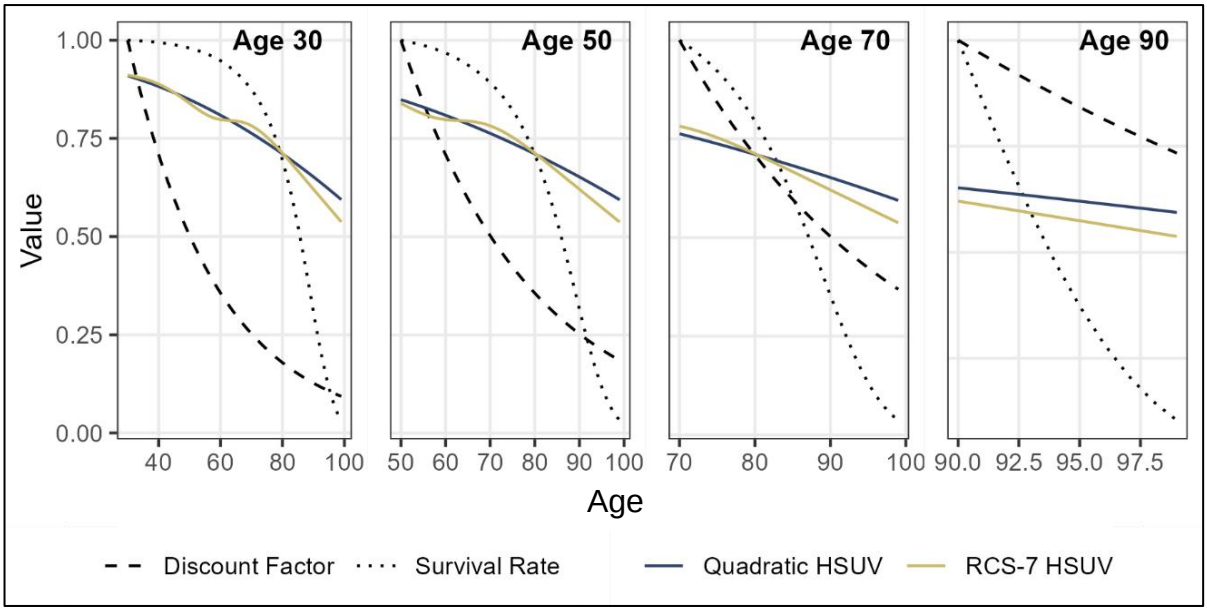

Figure 13: QALE Calculations for Quadratic and RCS-7 Models

respectively. If the discount rate is set to zero these figures become 0.0441, 0.101, 0.002 and 0.189. These are slightly different to the discounted results, but their magnitude is similar in general. These differences are also similar across a broader range of ages and sexes (see Appendix C1 for full results).

The fact that the difference in QALE between the RCS-7 and quadratic models is consistently below 0.2 QALYs means that the choice of baseline model is unlikely to be meaningful in practice. Such slight differences may be negligible when uncertainty is taken into account, and they are small relative to the 12 QALY absolute shortfall threshold required for severity modifiers (NICE, 2022b, p.164). If used for *proportional* shortfall calculations small differences in QALE may produce meaningful results, though this depends entirely on the utilities and survival of the untreated population. It is also worth noting that baseline utilities may be applied in a range of contexts which cannot be easily simulated. For example, they often form a part of utility calculations where multiple conditions are combined (Ara and Wailoo, 2013). However, given that work that does look at a broad range of real-world applications of baseline utilities has

failed to find that the choice of baseline affects results (Hernández Alava et al., 2022), it is difficult to argue that the choice of baseline model is qualitatively significant.

6 Discussion

6.1 Summary

I have analysed the results of the cross-validation in several different ways, each being useful for different reasons. Visual inspection reveals that higher order (6 and above) polynomial functions are unsuitable for prediction due to counter-intuitive results in the tails of the distribution, and that ALDVMM predictions are unstable and not always close to the predictions of the quadratic model. RMSEs reveal small differences between the models, supporting an increased degree of model flexibility and prescribing an RCS model with 7 knots. The RMSE results are particularly convincing when looking at the oldest age decile (ages 76-98). MAE results are counter-intuitive and suggest that this may not be a suitable error metric for this kind of data. Overall MEs are difficult to interpret, but the results stratified by age decile reveal that more complex models (RCS with at least 7 knots) improve the MEs in several deciles. In this discussion, I will explore the implications of these results, especially for the three key questions specified in section 4.4.

6.2 Key Questions

Are ALDVMMs a useful tool for capturing the relationship between utilities and age?

As I have repeatedly mentioned, it is difficult to know how much to draw from the ALDVMM results in this analysis. What is clear is that ALDVMM predictions may differ significantly across similar samples when the same model fitting settings are used, casting doubt on the reproducibility of results. It may be that if a more comprehensive approach to model fitting in each fold was taken (for example, running a series of ALDVMMs and choosing the one with the highest AIC in each case), predictions may

be more stable and perhaps more closely match the quadratic model as has been previously suggested, but this is the subject for another analysis.

While implementing it is not as straightforward as other models, it would be wrong to dismiss the ALDVMM approach as it is designed specifically to account for the quirks of the Dolan (1997) EQ-5D-3L value set. However, it is worth keeping in mind that it may become less applicable when NICE moves to the EQ-5D-5L. Though there is not currently an approved value set for the newer measure, some research has suggested that its increased sensitivity will lead to EQ-5D-5L scores having a less unusual distribution (Thompson and Turner, 2020), largely removing the need for ALDVMMs.

Do RCS functions outperform polynomial functions?

From visual inspection it is clear that both models allow for varying levels of model flexibility to be considered. As the degree of polynomial/number of knots increases, more 'wiggly' prediction curves are produced, as expected. However, the models perform very differently in the tails of the distributions, with higher 6^{th}-order and above polynomials being rejected due to strange predictions for the older population. RCS functions are linear in these groups by design, allowing for higher degrees of flexibility to be considered and ultimately resulting in a stronger performance in terms of RMSE and ME. As a result, model flexibility can be best understood by analysing RCS models exclusively.

What degree of model flexibility is optimal?

Model flexibility is perhaps the most interesting objective as previous work on HSUV baselines has not explored it, and the k-fold cross-validation procedure is well-suited to judge the overfitting that might occur in overly-flexible models. Looking at the various RCS models, though the differences between models are small, there seems

to be a clear pattern of improvement as the number of knots increases. This is particularly clear in the upper age decile for RMSEs, an area that previous models have noted as being difficult (e.g. Guthrie et al., 2023). MEs also favour a more flexible model, especially for those aged above 50. Overall, the results suggest an RCS model with around 7 knots would be optimal for model flexibility – more flexible models with more than 7 knots show little improvement in RMSE or ME. The predictions of the RCS-7 model when trained on the full dataset are reported in Appendix D1-2, and show that there are statistically significant differences between the quadratic and RCS-7 models for some age ranges. However, it is not clear that these differences are qualitatively significant. Analysis comparing QALE estimates for females aged 30, 50, 70 and 90 revealed very small differences between the two models, although there are many HTA contexts in which baselines might be implemented differently. The differences may also be more significant in terms of proportional shortfall where untreated QALE is low, which would occur particularly often in the older population.

6.3 Limitations

This study has several limitations. One that has not been discussed so far is that all models have been fitted under the assumption of random sampling. The HSE follows a complex survey design (Bridges et al., 2015) which means that this assumption does not hold in practice. Previous HSE analyses estimating baselines have differed in their approach to this feature of the dataset: some have adjusted for it (Hernández Alava et al., 2022), some have not (Ara and Brazier, 2010), and for others it is not clear. Not adjusting for the complex survey design may mean that there are biases in the results, making them unrepresentative of the true population. However, it is unclear whether these biases are significant, so the results may still be reasonably valid.

Another potential limitation is the use of internal validation. While k-fold cross-validation is a useful tool for detecting overfitted models, the optimal approach would be to conduct external validation in another dataset. This would correct for any trends in the HSE datasets which do not exist in the real world. Unfortunately, as has been previously noted, finding a dataset that has sufficient information and is collected from a plausibly related population is challenging.

A final, more fundamental, limitation is the quality of data for the older population. Even pooling nine HSE rounds together has failed to produce reasonable confidence intervals for the means in these age groups (see Figure 5). Finding data with sufficient sample sizes in these groups is very difficult, especially given that there is particularly high variation in their HSUV scores (see Figure 6). In addition, pooling smaller datasets is not always a valid solution as exact ages may be censored to protect anonymity. Another difficulty with producing baselines for older people is that the EQ-5D may not be an appropriate PROM. Penton et al.'s (2022) results suggest that older people may interpret certain EQ-5D terms like 'usual activities' differently from the rest of the population. They also found that scores were often conceptualised relative to the lower health expectations in old age, raising the issue of prospect theory (Kahneman and Tversky, 1979) and its application to health outcomes (Attema, Brouwer and I'Haridon, 2013). More specialised PROMs may therefore be useful for estimating baselines in older populations (see van Oppen et al., 2022 for a review in an emergency care context), though moving away from the EQ-5D might incur a trade-off in consistency. Worryingly, the older population may be where model choice is of the most qualitative significance, reinforcing the need for further study in this area.

7 Conclusion

In conclusion, the results of the stratified 10-fold cross-validation procedure suggest that more flexible models not previously considered in the literature outperform existing models in the estimation of age- and sex-adjusted baseline utilities. While the interpretation of the performance of such underpowered models is not straightforward, an RCS model with 7 knots seems to be the optimal functional form for a non-linear predictor of age, conditional on sex. In addition, the NICE's DSU's preferred model fails to outperform even relatively simple polynomial models, though these results should be treated with caution as ALDVMMs are not suited to cross-validation approaches. However, while the baselines produced by the preferred RCS-7 model are significantly different from the status quo quadratic model, it is unclear whether these differences would be meaningful if implemented in an HTA. But while this study fails to fully justify the use of more complex models for HSUV baselines in economic evaluations, its findings strongly suggest that future HSUV studies in which model selection is more important should consider more flexible approaches to modelling. In addition, when there is an approved value set for the EQ-5D-5L in the UK it may be beneficial to repeat this analysis and explore whether the increased sensitivity in the PROM allows for complex trends to emerge more clearly.

8 Bibliography

Ara, R. and Brazier, J. (2011a). Estimating health state utility values for comorbid health conditions using SF-6D data. *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*, 14 (5), pp.740–745.

Ara, R. and Brazier, J. E. (2010). Populating an economic model with health state utility values: moving toward better practice. *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*, 13 (5), pp.509–518.

Ara, R. and Brazier, J. E. (2011b). Using health state utility values from the general population to approximate baselines in decision analytic models when condition-specific data are not available. *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*, 14 (4), pp.539–545.

Ara, R., Brazier, J. and Zouraq, I. A. (2017). The Use of Health State Utility Values in Decision Models. *PharmacoEconomics*, 35 (Suppl 1), pp.77–88.

Ara, R. and Wailoo, A. J. (2011). *NICE DSU Technical Support Document 12: The Use of Health State Utility Values in Decision Models*.

Ara, R. and Wailoo, A. J. (2013). Estimating health state utility values for joint health conditions: a conceptual review and critique of the current evidence. *Medical decision making: an international journal of the Society for Medical Decision Making*, 33 (2), pp.139–153.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4, pp.40–79.

Attema, A. E., Brouwer, W. B. F. and I'Haridon, O. (2013). Prospect theory in the health domain: a quantitative assessment. *Journal of health economics*, 32 (6), pp.1057–1065.

Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research*, 5, pp.1089–1105.

Blanchflower, D. G. and Oswald, A. J. (2019). Do humans suffer a psychological low in midlife? Two approaches (with and without controls) in seven data sets. In: *The Economics of Happiness*. Cham: Springer International Publishing. pp.439–453.

Bridges, S. et al. (2015). *Health Survey for England 2014. Volume 2. Methods and Documentation*. Craig, R. et al. (Eds). Health and Social Care Information Centre. [Online]. Available at: https://files.digital.nhs.uk/publicationimport/pub19xxx/pub19295/hse2014-methods-and-docs.pdf.

Chang-Douglass, S., Bungey, G. and Dakin, H. (2022). Modelling of UK general population utility: The ALDVMM Approach in Practice. In: *ISPOR EU*. Poster MSR36. 2022. [Online]. Available at: https://www.ispor.org/docs/default-source/euro2022/modelling-of-uk-general-population-ispor-eu-2022-poster21oct22v1-0-pdf.pdf.

Chekroud, A. M. et al. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World psychiatry: official journal of the World Psychiatric Association* , 20 (2), pp.154–170.

Croxford, R. (2016). *Restricted Cubic Spline Regression - A Brief Introduction*. Institute for Clinical Evaluation Sciences. [Online]. Available at: https://support.sas.com/resources/papers/proceedings16/5621-2016.pdf.

Davison, N. J. et al. (2018). Generating EQ-5D-3L Utility Scores from the Dermatology Life Quality Index: A Mapping Study in Patients with Psoriasis. *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*, 21 (8), pp.1010–1018.

Dolan, P. (1997). Modeling valuations for EuroQol health states. *Medical care*, 35 (11), pp.1095–1108.

Gay, D. M. (1990). *Usage summary for selected optimization routines*. [Online]. Available at: https://ms.mcmaster.ca/~bolker/misc/port.pdf.

Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association*, 70 (350), pp.320–328.

Gibson, E. et al. (2017). Modelling the Survival Outcomes of Immuno-Oncology Drugs in Economic Evaluations: A Systematic Approach to Data Analysis and Extrapolation. *PharmacoEconomics*, 35 (12), pp.1257–1270.

Grambsch, P. M. and O'Brien, P. C. (1991). The effects of transformations and preliminary tests for non-linearity in regression. *Statistics in medicine*, 10 (5), pp.697–709.

Gray, L. A., Wailoo, A. J. and Hernández Alava, M. (2018). Mapping the FACT-B Instrument to EQ-5D-3L in Patients with Breast Cancer Using Adjusted Limited Dependent Variable Mixture Models versus Response Mapping. *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*, 21 (12), pp.1399–1405.

Guthrie, B. et al. (2023). *The implications of competing risks and direct treatment disutility in cardiovascular disease and osteoporotic fracture: risk prediction and cost effectiveness analysis*. [Online]. Available at: https://ssrn.com/abstract=4319295.

Harrell, F. E. (2015). *Regression Modeling Strategies*. 2nd ed. Cham: Springer.

Harrell, F. E., Jr. (2023). *rms: Regression Modeling Strategies*. [Online]. Available at: https://CRAN.R-project.org/package=rms.

Harrell, F. E., Jr, Lee, K. L. and Pollock, B. G. (1988). Regression models in clinical studies: determining relationships between predictors and response. *Journal of the National Cancer Institute*, 80 (15), pp.1198–1202.

Hernández Alava, M., Pudney, S. and Wailoo, A. (2022). *Estimating EQ-5D by Age and Sex for the UK*. NICE DSU.

Hernández Alava, M., Pudney, S. and Wailoo, A. (2023). Estimating the Relationship Between EQ-5D-5L and EQ-5D-3L: Results from a UK Population Study. *PharmacoEconomics*, 41 (2), pp.199–207.

Hernández Alava, M. and Wailoo, A. (2015). Fitting Adjusted Limited Dependent Variable Mixture Models to EQ-5D. *The Stata journal*, 15 (3), pp.737–750.

Hernández Alava, M., Wailoo, A. J. and Ara, R. (2012). Tails from the peak district: adjusted limited dependent variable mixture models of EQ-5D questionnaire health

state utility values. *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*, 15 (3), pp.550–561.

Hill, H. et al. (2019). *A review of the methods used to estimate and model utility values in NICE technology appraisals for paediatric populations*. NICE DSU.

James, G. et al. (2013). *An Introduction to Statistical Learning: with Applications in R*, Springer texts in statistics. New York: Springer Nature.

Jung, K. et al. (2020). Evaluation of Nitrate Load Estimations Using Neural Networks and Canonical Correlation Analysis with K-Fold Cross-Validation. *Sustainability: Science Practice and Policy*, 12 (1), p.400.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica: journal of the Econometric Society*, 47 (2), pp.263–291.

Kind, P., Hardman, G. and Macran, S. (1999). UK population norms for EQ-5D. *CHE Discussion Paper Series*, (172).

Lever, J., Krzywinski, M. and Altman, N. (2016). Points of Significance: Model selection and overfitting. *Nature methods*, 13 (9), pp.703–704.

Manca, A., Hawkins, N. and Sculpher, M. J. (2005). Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. *Health economics*, 14 (5), pp.487–496.

McNamara, S. et al. (2023). Quality-Adjusted Life Expectancy Norms for the English Population. *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*, 26 (2), pp.163–169.

Morton, F. and Nijjar, J. S. (2023). *eq5d: Methods for Analysing 'EQ-5D' Data and Calculating 'EQ-5D' Index Scores*. [Online]. Available at: https://CRAN.R-project.org/package=eq5d.

NatCen Social Research. (2023). *Health Survey for England*. UK Data Service. [Online]. Available at: doi:10.5255/UKDA-Series-2000021.

NICE. (2013a). *Aripiprazole for treating moderate to severe manic episodes in adolescents with bipolar I disorder*. NICE Technology Appraisal Guidance 292.

[Online]. Available at: https://www.nice.org.uk/guidance/ta292/resources/aripiprazole-for-treating-moderate-to-severe-manic-episodes-in-adolescents-with-bipolar-i-disorder-pdf-82600730031301.

NICE. (2013b). *Peginterferon alfa and ribavirin for treating chronic hepatitis C in children and young people*. [Online]. Available at: https://www.nice.org.uk/guidance/ta300/resources/peginterferon-alfa-and-ribavirin-for-treating-chronic-hepatitisc-in-children-and-young-people-pdf-82602359258821.

NICE. (2015). *Tolvaptan for treating autosomal dominant polycystic kidney disease*. [Online]. Available at: https://www.nice.org.uk/guidance/ta358/resources/tolvaptan-for-treating-autosomal-dominant-polycystic-kidney-disease-pdf-82602675026629.

NICE. (2019). *Position statement on use of the EQ-5D-5L value set for England (updated October 2019)*. [Online]. NICE. Available at: https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/technology-appraisal-guidance/eq-5d-5l.

NICE. (2021). *Fremanezumab for preventing migraine*. [Online]. Available at: https://www.nice.org.uk/guidance/ta764/resources/fremanezumab-for-preventing-migraine-pdf-82611435903685.

NICE. (2022a). *Abrocitinib, tralokinumab or upadacitinib for treating moderate to severe atopic dermatitis*. [Online]. Available at: https://www.nice.org.uk/guidance/ta814/resources/abrocitinib-tralokinumab-or-upadacitinib-for-treating-moderate-to-severe-atopic-dermatitis-pdf-82613310355141.

NICE. (2022b). *NICE Health Technology Evaluations: The Manual (PMG36)*. [Online]. Available at: https://www.nice.org.uk/process/pmg36/resources/nice-health-technology-evaluations-the-manual-pdf-72286779244741.

Norwegian Institute of Public Health. (2021). *Guidelines for the submission of documentation for single technology assessments (STAs) of medical devices and diagnostic interventions*.

Office for National Statistics. (2021). *National life tables: England*. [Online]. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesenglandreferencetables.

van Oppen, J. D. et al. (2022). A systematic review and recommendations for prom instruments for older people with frailty in emergency care. *Journal of patient-reported outcomes*, 6 (1), p.30.

Penton, H. et al. (2022). A Qualitative Investigation of Older Adults' Conceptualization of Quality of Life and a Think-Aloud Content Validation of the EQ-5D-5L, SF-12v2, Warwick Edinburgh Mental Well-Being Scale, and Office of National Statistics-4. *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*, 25 (12), pp.2017–2027.

Pletscher, M. (2023). *aldvmm: Adjusted Limited Dependent Variable Mixture Models*. [Online]. Available at: https://CRAN.R-project.org/package=aldvmm.

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. [Online]. Available at: https://www.R-project.org/.

Sainani, K. L. (2014). Explanatory versus predictive modeling. *PM & R: the journal of injury, function, and rehabilitation*, 6 (9), pp.841–844.

Sammut, C. et al. (2010). *Encyclopedia of machine learning [electronic resource]*. New York ; London: Springer.

Schneider, P. et al. (2023). *QALY Shortfall Calculator*. [Online]. Available at: https://shiny.york.ac.uk/shortfall/.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, 25 (3), pp.289–310.

Stavem, K. et al. (2018). General population norms for the EQ-5D-3 L in Norway: comparison of postal and web surveys. *Health and quality of life outcomes*, 16 (1), p.204.

Steyerberg, E. W. (2019). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer.

Szende, A., Janssen, B. and Cabases, J. (Eds). (2014). *Self-Reported Population Health: An International Perspective based on EQ-5D*. Dordrecht (NL): Springer.

Taagepera, R. (2008). *Making social sciences more scientific: the need for predictive models*. Oxford: Oxford University Press.

Thompson, A. (2019). *Joint Health State Utility values: an external validation study*. PhD Thesis, University of Manchester.

Thompson, A. J., Sutton, M. and Payne, K. (2019). Estimating Joint Health Condition Utility Values. *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*, 22 (4), pp.482–490.

Thompson, A. J. and Turner, A. J. (2020). A Comparison of the EQ-5D-3L and EQ-5D-5L. *PharmacoEconomics*, 38 (6), pp.575–591.

Walther, B. A. and Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, 28 (6), pp.815–829.

Yousef, W. A. (2021). Estimating the standard error of cross-validation-based estimators of classifier performance. *Pattern recognition letters*, 146, pp.115–125.

9 Appendix

A1: New data vs. old data OLS regression summary

	Male		Female	
	Old	**New**	**Old**	**New**
Constant	0.9512620 ***	0.9938977 ***	0.9573756 ***	0.9533967 ***
	(0.0179182)	(0.0185848)	(0.0165258)	(0.0172715)
Age	0.0001167	-0.0026846 ***	-0.0003596	-0.0012931
	(0.0007820)	(0.0007897)	(0.0007064)	(0.0007297)
Age^2	-0.0000318 ***	-0.0000041	-0.0000341 ***	-0.0000232 **
	(0.0000079)	(0.0000077)	(0.0000069)	(0.0000071)
N	6092	6496	7661	8353
R^2	0.0645159	0.0624424	0.0946870	0.0757222

*** p < 0.001; ** p < 0.01; * p < 0.05.

B1: Overall error statistic table

Model	**ME**	**RMSE**	**MAE**
Ones	-1.47E-01	0.275203	0.147958
Mean	6.89E-06	0.231726	0.165348
Linear	2.78E-06	0.222869	0.152815
ALDVMM	-1.42E-03	0.222929	0.153205
Quadratic	1.32E-06	0.222689	0.153198
Poly-3	-3.07E-07	0.222686	0.153207
Poly-4	-2.77E-06	0.222588	0.153084
Poly-5	-4.92E-06	0.222562	0.153091
Poly-6	-2.09E-06	0.222516	0.153039
Poly-7	1.50E-06	0.222507	0.153060
Poly-8	1.07E-06	0.222497	0.153040
Poly-9	-1.49E-06	0.222495	0.153044
Poly-10	2.53E-08	0.222492	0.153023
RCS-3	1.96E-06	0.222730	0.153169
RCS-4	1.41E-06	0.222722	0.153174
RCS-5	3.77E-07	0.222648	0.153077
RCS-6	-4.88E-07	0.222543	0.153068
RCS-7	-4.11E-07	0.222488	0.153062
RCS-8	5.99E-07	0.222467	0.153049

RCS-9	-3.66E-07	0.222459	0.153017
RCS-10	-2.04E-06	0.222466	0.153023

B2: Model RMSEs by age decile

Age	**[16,24]**	**(24,32]**	**(32,38]**	**(38,43]**	**(43,49]**
Mean	0.160976	0.176800	0.187941	0.200800	0.219404
Linear	0.138248	0.162386	0.179732	0.197265	0.218744
ALDVMM	0.137660	0.162483	0.179620	0.197187	0.218576
Quadratic	0.137192	0.162336	0.179646	0.197025	0.218537
Poly-3	0.137251	0.162373	0.179714	0.197066	0.218541
Poly-4	0.137462	0.162384	0.179587	0.197107	0.218662
Poly-5	0.137253	0.162304	0.179588	0.197065	0.218620
Poly-6	0.137174	0.162280	0.179614	0.197049	0.218692
Poly-7	0.137172	0.162316	0.179617	0.197052	0.218687
Poly-8	0.137195	0.162318	0.179623	0.197032	0.218684
Poly-9	0.137187	0.162318	0.179607	0.197018	0.218649
Poly-10	0.137181	0.162314	0.179602	0.197010	0.218639
RCS-3	0.137222	0.162352	0.179665	0.197027	0.218537
RCS-4	0.137201	0.162333	0.179643	0.197027	0.218543
RCS-5	0.137262	0.162324	0.179586	0.197083	0.218784
RCS-6	0.137181	0.162327	0.179623	0.197050	0.218705
RCS-7	0.137218	0.162326	0.179610	0.196993	0.218644
RCS-8	0.137208	0.162311	0.179603	0.197020	0.218645
RCS-9	0.137178	0.162298	0.179596	0.197044	0.218572
RCS-10	0.137164	0.162287	0.179585	0.197024	0.218640

	(49,55]	**(55,61]**	**(61,67]**	**(67,75]**	**(75,98]**
Mean	0.246712	0.264853	0.265859	0.264298	0.298762
Linear	0.246141	0.262637	0.261879	0.257203	0.269316
ALDVMM	0.246728	0.263053	0.261831	0.257485	0.269171
Quadratic	0.246464	0.262861	0.261844	0.257309	0.268191
Poly-3	0.246483	0.262921	0.261852	0.257218	0.268025
Poly-4	0.246253	0.262742	0.261790	0.256979	0.267676
Poly-5	0.246240	0.262706	0.261723	0.256984	0.267800
Poly-6	0.246136	0.262632	0.261686	0.256910	0.267665
Poly-7	0.246122	0.262646	0.261617	0.256900	0.267646
Poly-8	0.246079	0.262605	0.261610	0.256900	0.267644
Poly-9	0.246084	0.262612	0.261605	0.256915	0.267667
Poly-10	0.246068	0.262597	0.261600	0.256916	0.267699
RCS-3	0.246467	0.262837	0.261864	0.257383	0.268437
RCS-4	0.246450	0.262834	0.261841	0.257382	0.268451
RCS-5	0.246199	0.262804	0.261809	0.257122	0.268106

RCS-6	0.246080	0.262589	0.261762	0.256950	0.267844
RCS-7	0.246079	0.262691	0.261520	0.256916	0.267621
RCS-8	0.246078	0.262637	0.261428	0.256915	0.267571
RCS-9	0.246078	0.262603	0.261472	0.256920	0.267573
RCS-10	0.246051	0.262615	0.261514	0.256921	0.267579

B3: Model MEs by age decile

Age	**[16,24]**	**(24,32]**	**(32,38]**	**(38,43]**	**(43,49]**
Mean	0.083889	0.069987	0.054717	0.038530	0.019621
Linear	-0.016673	-0.000681	0.007461	0.009679	0.009237
ALDVMM	0.000246	0.002137	0.003491	0.002703	0.000636
Quadratic	-0.002437	0.002531	0.004564	0.003600	0.001410
Poly-3	-0.004986	0.003789	0.006640	0.005369	0.002333
Poly-4	-0.000799	-0.003941	0.002316	0.006056	0.007097
Poly-5	-0.000917	-0.000707	0.001198	0.002941	0.004473
Poly-6	0.000776	0.000131	-0.003134	0.000239	0.005887
Poly-7	-0.000182	0.001325	-0.001952	-0.000564	0.004287
Poly-8	-0.000644	0.002101	-0.002069	-0.001313	0.004080
Poly-9	-0.000228	0.000968	-0.000729	-0.000832	0.002872
Poly-10	-0.000214	0.000864	-0.000204	-0.001266	0.002414
RCS-3	-0.004403	0.003076	0.005191	0.003783	0.001248
RCS-4	-0.002994	0.002603	0.004087	0.002868	0.001016
RCS-5	0.000757	-0.002756	-0.001492	0.003780	0.008690
RCS-6	-0.001256	0.002695	-0.000352	-0.003863	0.002738
RCS-7	-0.000818	0.000359	0.001344	0.000342	-0.000084
RCS-8	-0.000946	0.000986	0.000979	-0.001071	0.001459
RCS-9	-0.000803	0.001347	-0.000970	0.001640	0.000241
RCS-10	-0.000463	0.000788	-0.000934	0.002725	-0.001292

	(49,55]	**(55,61]**	**(61,67]**	**(67,75]**	**(75,98]**
Mean	-0.013917	-0.034421	-0.047928	-0.061117	-0.125581
Linear	-0.003791	-0.003877	0.003004	0.012892	-0.017205
ALDVMM	-0.013170	-0.012754	-0.003658	0.011043	-0.006378
Quadratic	-0.011930	-0.010602	-0.000592	0.014969	-0.002227
Poly-3	-0.012209	-0.012003	-0.002664	0.013183	-0.000157
Poly-4	-0.005796	-0.007593	-0.003094	0.006873	-0.000934
Poly-5	-0.005941	-0.005130	0.000045	0.007196	-0.003609
Poly-6	-0.001685	-0.002381	-0.001759	0.002582	-0.001113
Poly-7	-0.002099	-0.001077	-0.000517	0.001529	-0.001185
Poly-8	-0.001396	-0.000485	-0.000966	0.000757	-0.000764
Poly-9	-0.002135	0.000618	-0.000095	-0.000549	-0.000330
Poly-10	-0.001636	0.001063	-0.000662	-0.000709	-0.000109

RCS-3	-0.011874	-0.009842	0.000791	0.016177	-0.004899
RCS-4	-0.011506	-0.009337	0.001102	0.016167	-0.004976
RCS-5	-0.004511	-0.009784	-0.005047	0.010864	-0.001098
RCS-6	0.002998	-0.000954	-0.007781	0.003342	0.001436
RCS-7	-0.002761	0.004632	-0.002394	-0.002105	0.001568
RCS-8	-0.002372	0.002817	-0.000762	-0.002465	0.001356
RCS-9	-0.002826	0.004217	-0.001789	-0.002408	0.001438
RCS-10	-0.001295	0.003122	-0.001524	-0.002059	0.001184

Appendix C1: Discounted QALE Estimates

Age	Female			Male		
	RCS-7	Quadratic	Difference	RCS-7	Quadratic	Difference
16	23.185	23.193	-0.0082	23.289	23.290	-0.0009
17	23.032	23.038	-0.0063	23.127	23.123	0.0035
18	22.877	22.881	-0.0043	22.962	22.954	0.0073
19	22.718	22.720	-0.0022	22.794	22.783	0.0105
20	22.554	22.555	-0.0001	22.623	22.610	0.0129
21	22.387	22.385	0.0020	22.448	22.433	0.0146
22	22.216	22.212	0.0040	22.267	22.252	0.0156
23	22.041	22.036	0.0058	22.082	22.066	0.0158
24	21.862	21.854	0.0073	21.892	21.876	0.0152
25	21.678	21.669	0.0086	21.697	21.683	0.0141
26	21.490	21.480	0.0095	21.498	21.486	0.0123
27	21.297	21.287	0.0100	21.295	21.285	0.0099
28	21.100	21.090	0.0100	21.086	21.079	0.0071
29	20.898	20.888	0.0094	20.874	20.870	0.0040
30	20.691	20.683	0.0082	20.657	20.656	0.0007
31	20.479	20.473	0.0062	20.436	20.439	-0.0026
32	20.263	20.259	0.0035	20.211	20.217	-0.0057
33	20.041	20.041	-0.0001	19.981	19.990	-0.0086
34	19.815	19.819	-0.0046	19.748	19.759	-0.0109
35	19.583	19.593	-0.0100	19.511	19.524	-0.0126
36	19.346	19.362	-0.0163	19.270	19.284	-0.0138
37	19.105	19.128	-0.0233	19.026	19.041	-0.0145
38	18.858	18.889	-0.0309	18.778	18.793	-0.0148
39	18.606	18.645	-0.0390	18.525	18.540	-0.0149
40	18.350	18.397	-0.0473	18.268	18.283	-0.0152
41	18.089	18.144	-0.0555	18.006	18.022	-0.0159
42	17.824	17.887	-0.0633	17.738	17.756	-0.0176

43	17.556	17.626	-0.0704	17.465	17.485	-0.0206
44	17.284	17.361	-0.0766	17.186	17.211	-0.0249
45	17.010	17.092	-0.0815	16.902	16.933	-0.0307
46	16.733	16.818	-0.0848	16.613	16.651	-0.0378
47	16.454	16.540	-0.0865	16.318	16.364	-0.0458
48	16.172	16.258	-0.0862	16.018	16.072	-0.0543
49	15.889	15.973	-0.0839	15.714	15.776	-0.0627
50	15.603	15.682	-0.0793	15.407	15.477	-0.0700
51	15.315	15.387	-0.0723	15.099	15.174	-0.0754
52	15.025	15.088	-0.0630	14.788	14.866	-0.0780
53	14.733	14.784	-0.0514	14.476	14.554	-0.0773
54	14.437	14.475	-0.0377	14.164	14.237	-0.0728
55	14.139	14.161	-0.0221	13.850	13.915	-0.0646
56	13.839	13.844	-0.0052	13.537	13.590	-0.0529
57	13.536	13.523	0.0126	13.223	13.261	-0.0380
58	13.230	13.199	0.0307	12.909	12.929	-0.0207
59	12.919	12.871	0.0483	12.593	12.595	-0.0020
60	12.604	12.540	0.0645	12.274	12.257	0.0167
61	12.284	12.205	0.0786	11.951	11.916	0.0345
62	11.958	11.869	0.0899	11.624	11.574	0.0502
63	11.628	11.530	0.0979	11.293	11.230	0.0632
64	11.289	11.186	0.1023	10.956	10.884	0.0727
65	10.943	10.840	0.1030	10.614	10.535	0.0784
66	10.590	10.490	0.0998	10.268	10.188	0.0801
67	10.230	10.137	0.0928	9.915	9.838	0.0778
68	9.865	9.783	0.0824	9.558	9.487	0.0716
69	9.495	9.426	0.0691	9.198	9.136	0.0620
70	9.121	9.068	0.0533	8.833	8.783	0.0496
71	8.743	8.707	0.0358	8.462	8.427	0.0350
72	8.362	8.345	0.0170	8.087	8.068	0.0188
73	7.982	7.984	-0.0025	7.714	7.713	0.0015
74	7.604	7.626	-0.0222	7.346	7.362	-0.0163
75	7.228	7.270	-0.0418	6.981	7.015	-0.0342
76	6.855	6.916	-0.0607	6.622	6.673	-0.0519
77	6.488	6.566	-0.0787	6.268	6.337	-0.0689
78	6.128	6.223	-0.0956	5.924	6.009	-0.0850
79	5.775	5.886	-0.1110	5.588	5.688	-0.1000
80	5.431	5.555	-0.1248	5.260	5.374	-0.1136
81	5.093	5.230	-0.1368	4.941	5.067	-0.1257
82	4.764	4.911	-0.1470	4.631	4.767	-0.1362
83	4.446	4.601	-0.1552	4.331	4.476	-0.1449

84	4.139	4.301	-0.1617	4.043	4.196	-0.1521
85	3.846	4.013	-0.1664	3.770	3.928	-0.1578
86	3.568	3.738	-0.1696	3.513	3.675	-0.1621
87	3.305	3.477	-0.1715	3.271	3.436	-0.1653
88	3.056	3.228	-0.1720	3.043	3.211	-0.1675
89	2.820	2.991	-0.1713	2.834	3.003	-0.1688
90	2.596	2.766	-0.1695	2.630	2.799	-0.1688
91	2.385	2.552	-0.1666	2.429	2.596	-0.1672
92	2.185	2.348	-0.1628	2.236	2.401	-0.1645
93	1.993	2.151	-0.1578	2.051	2.212	-0.1606
94	1.803	1.954	-0.1511	1.873	2.029	-0.1557
95	1.609	1.752	-0.1423	1.695	1.844	-0.1489
96	1.407	1.537	-0.1307	1.507	1.647	-0.1395
97	1.181	1.296	-0.1150	1.289	1.414	-0.1251
98	0.906	0.998	-0.0920	1.007	1.109	-0.1021
99	0.537	0.594	-0.0567	0.614	0.679	-0.0647

D1: RCS-7 predictions

Age	Female	Male	Age	Female	Male
16	0.9349	0.9500	59	0.7995	0.8214
17	0.9333	0.9487	60	0.7979	0.8198
18	0.9316	0.9475	61	0.7969	0.8191
19	0.9300	0.9462	62	0.7961	0.8191
20	0.9283	0.9449	63	0.7955	0.8195
21	0.9267	0.9436	64	0.7949	0.8201
22	0.9250	0.9422	65	0.7941	0.8208
23	0.9234	0.9407	66	0.7930	0.8212
24	0.9217	0.9391	67	0.7913	0.8212
25	0.9201	0.9374	68	0.7891	0.8206
26	0.9184	0.9356	69	0.7860	0.8192
27	0.9167	0.9337	70	0.7821	0.8170
28	0.9150	0.9315	71	0.7776	0.8140
29	0.9133	0.9292	72	0.7723	0.8104
30	0.9116	0.9267	73	0.7665	0.8061
31	0.9098	0.9240	74	0.7601	0.8012
32	0.9081	0.9211	75	0.7531	0.7957
33	0.9062	0.9180	76	0.7458	0.7898
34	0.9043	0.9149	77	0.7379	0.7835

35	0.9023	0.9117	78	0.7298	0.7768
36	0.9000	0.9085	79	0.7213	0.7698
37	0.8976	0.9055	80	0.7126	0.7625
38	0.8949	0.9027	81	0.7037	0.7550
39	0.8918	0.9001	82	0.6946	0.7473
40	0.8885	0.8978	83	0.6854	0.7395
41	0.8847	0.8959	84	0.6761	0.7317
42	0.8807	0.8943	85	0.6669	0.7239
43	0.8763	0.8927	86	0.6576	0.7160
44	0.8716	0.8910	87	0.6484	0.7082
45	0.8666	0.8892	88	0.6391	0.7004
46	0.8614	0.8869	89	0.6299	0.6925
47	0.8561	0.8841	90	0.6206	0.6847
48	0.8506	0.8806	91	0.6114	0.6769
49	0.8449	0.8763	92	0.6021	0.6690
50	0.8392	0.8711	93	0.5929	0.6612
51	0.8334	0.8651	94	0.5836	0.6534
52	0.8278	0.8586	95	0.5744	0.6455
53	0.8223	0.8519	96	0.5651	0.6377
54	0.8172	0.8453	97	0.5559	0.6298
55	0.8125	0.8389	98	0.5466	0.6220
56	0.8082	0.8331	99	0.5373	0.6142
57	0.8046	0.8281	100	0.5281	0.6063
58	0.8017	0.8241			

D2: RCS-7 model predictions, with quadratic predictions plotted for context and 95% confidence intervals for each