Severity Weights in NICE Technology Appraisals.

REPORT BY THE DECISION SUPPORT UNIT

5th September 2024

Allan Wailoo

Sheffield Centre for Health and Related Research (SCHARR), School of Medicine and Population Health, University of Sheffield.

Decision Support Unit, SCHARR, University of Sheffield, Regent Court, 30 Regent Street Sheffield, S1 4DA Tel (+44) (0)114 222 0801 E-mail dsuadmin@sheffield.ac.uk

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) is based at the University of Sheffield with members at the Universities of York, Bristol, Leicester, Warwick and the London School of Hygiene and Tropical Medicine. The DSU is commissioned by The National Institute for Health and Care Excellence (NICE) to provide a research and training resource to support the Institute's Centre for Health Technology Evaluation Programmes. Please see our website for further information: https://www.sheffield.ac.uk/nice-dsu

The production of this document was funded by the National Institute for Health and Care Excellence (NICE) through its Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the authors only. NICE may take account of part or all of this document if it considers it appropriate, but it is not bound to do so.

ACKNOWLEDGEMENTS:

Several people provided information and comments that contributed to this work. At NICE this includes Lorna Dunning, Koonal Shah, Juliet Kenny, Tuba Saygin Avsar, Emily Leckenby and others.

Matt Stevenson and Nick Latimer reviewed previous drafts of this report.

Mónica Hernández Alava checked code for these analyses.

Donna Davis provided expert support including formatting the report.

This report should be referenced as follows:

Wailoo, A. Severity Weights in NICE Technology Appraisals. A NICE DSU Report. 2024.

EXECUTIVE SUMMARY

NICE replaced the modifier for “End of Life” with “severity” in the latest version of its methods guidance in January 2022.

Committees are now instructed to consider the severity of a condition, defined as the future health lost by people living with the condition with standard care in the NHS. It is measured in terms of Absolute and Proportional Shortfall (AS and PS). Both measure the difference between the number of discounted Quality Adjusted Life Years (QALYs) patients would be expected to experience over the remainder of their lives under current care compared to the general population of the same age and sex.

Weights to be applied to incremental QALYs in the cost-effectiveness analysis fall into three categories: 1, 1.2 and 1.7 and depend on the degree of severity as measured by AS and PS. The definition that provides the highest weight for severity is the one that should be used. These weights, and the cutoffs defining the AS and PS categories, were designed to ensure that the change from using the End of Life modifier to severity modifiers was broadly “opportunity cost neutral”, by matching the average QALY weight per decision for the severity modifier to that under end of life. These estimates were based on analysis of previous, pre-2022 NICE appraisals.

This report:

a) Compares the use of severity modifiers in all appraisals conducted post 2022 (the “implementation” sample) with the pre 2022 appraisals used to devise the severity cutoffs and weights (the primary+ sample).
b) Examines the impact of making the methods for calculating AS and PS consistent across the samples.
c) Compares the samples in terms of easily identifiable characteristics and suggests other issues that could be important to consider in follow-up work.

The implementation sample comprises 68 appraisal decisions, up to 31st March 2024, in which AS and PS were calculated. The primary+ sample includes 464 decisions from appraisals between January 2009 and March 2021.

For a) we find that there is no difference in the distribution of decisions by overall severity category between the implementation sample and the primary+ sample ($p=0.321$). There is no statistically significant difference between the samples for the

distribution of PS severity (p=0.218). There is a lower percentage of appraisal decisions in the AS 1.2 category in the implementation sample compared to the primary+ sample (5.9% vs 24.4%) and a lower percentage in the 1.7 AS category (0 vs 1.9%). This difference in the categorical distribution is statistically significant (p=0.001).

For b) AS and PS have been calculated for implementation sample appraisals, using more up to date methods and data sources. Using the same approach in the primary+ sample does not make a material difference to the distribution of overall or PS severity weights. There is a slight reduction in the percentage of appraisals in AS category 1.7 (falling from 24.5% to 17.3%) but the difference between the distribution of AS in the primary+ and implementation samples remains a statistically significant one.

For c) There is a higher percentage of appraisal decisions deemed to meet the end of life criteria in the implementation sample. This does not seem to explain the observed differences in severity.
Analyses over time highlight the degree of variability across years in overall, AS and PS severity measures. This reinforces the finding that the results from the implementation subsample are not anomalous given the sample size. There is also some evidence of a decrease over time in the mean weight for overall severity, AS and PS from 2017 to 2020.

Only a relatively small number of appraisals (n=47 covering 68 decisions) have been conducted using the new manual and for which AS and PS have been calculated. The difference in the distribution of appraisal decisions across the three overall severity weight categories used in decision making between the implementation and primary+ samples is not statistically significant. This implies that it may be too early to draw conclusions about the use of severity weights.

There are statistically significant differences observed between the implementation and primary+ subsamples in terms of the distribution across AS categories. This difference remains statistically significant using the cumulative set of new calculation sources and methods. Whilst of interest, it should be noted that severity weights and cutoffs were designed to achieve a broadly similar overall weighting as under end of

life, not for the AS and PS components to remain similar across appraisal samples. Furthermore, because the majority of appraisal decisions in the implementation sample that were just below the AS boundary to achieve a weight of 1.2 (12 QALYs) already had an overall weight of 1.2 or higher (because they had high PS), changes to the AS boundary would have little impact on decision making.

All these analyses are limited by the underlying data. There is substantial missing data and the reasons for this may differ between the pre and post 2022 analyses. Data extracted often required judgement to identify estimates that most closely aligned to the committee preferred case. Categorisation of severity and end of life also differs between the samples because, in some periods, these were live considerations in appraisals and therefore subject to scrutiny by committees and assessment groups. In other periods the assessment has been taken from the evidence included in the cost effectiveness analysis, not from committee judgement. Different appraisal samples may therefore be incommensurable.

Nevertheless, consideration should be given to other characteristics of appraisals that may be relevant for future analyses. Age is one such characteristic that should be a prime candidate for consideration and evidence shows that differences in appraisals between the ages of 45-65 years is likely to be the key driver of differences in AS. Other candidate characteristics include oncological versus non-oncological topics and, within oncological topics, a classification of cancer stage may be informative.
The value of these comparisons is likely to be greater once a larger set of post 2022 appraisal decisions has been allowed to mature.

CONTENTS

TABLES

FIGURES

ABBREVIATIONS

AS	Absolute Shortfall
CUA	Cost Utility Analysis
EoL	End of Life
ERG	Evidence Review Group
HSE	Health Survey for England
MTA	Multiple Technology Appraisal
NICE	National Institute for Health and Care Excellence
ONS	Office for National Statistics
PS	Proportional Shortfall
QALE	Quality Adjusted Life Expectancy
QALY	Quality Adjusted Life Years
STA	Single Technology Appraisal
TSD	Technical Support Document

1. INTRODUCTION

NICE published the latest version of its guidance on the methods to be used in technology appraisals in January 2022[1]. One of the most substantial changes made in the manual was the introduction of "a decision modifier related to the severity of the condition under consideration", replacing the "End of Life modifier" that was itself first used in 2008.

Committees are instructed to consider the severity of a condition, defined as the future health lost by people living with the condition with standard care in the NHS. It is measured, for the purposes of implementing the decision modifier, in terms of Absolute and Proportional Shortfall (AS and PS). Both measure the difference between the number of discounted Quality Adjusted Life Years (QALYs) patients would be expected to experience over the remainder of their lives under current care (referred to as Quality Adjusted Life Expectancy (QALE)), compared to the general population of the same age and sex. AS expresses this in terms of the number of QALYs lost. PS is a ratio of the number of QALYs lost compared to the total number of QALYs expected.

Weights should be applied to incremental QALYs estimated in the cost-effectiveness assessment. The weight that is to apply is specified in Table 6.1 of the manual and reproduced below.

Table 1: QALY weightings for severity

QALY weight	Proportional QALY shortfall	Absolute QALY shortfall
1	Less than 0.85	Less than 12
x1.2	0.85 to 0.95	12 to 18
x1.7	At least 0.95	At least 18

The manual defines three categories of AS for each appraisal decision, attracting weights of 1, 1.2 and 1.7 respectively and three categories of PS attracting weights of 1, 1.2 and 1.7 respectively. The definition that provides the highest weight for severity is the one that should be used. There is therefore an "overall" categorisation that considers both AS and PS and is used for decision making. The cutoffs used to define the three categories of AS and PS, as well as the weights to be applied, were based

on analysis of previous NICE appraisals. Specifically, they were designed to ensure that the change from using the EoL modifier to severity modifiers was broadly "opportunity cost neutral"[2] by matching the average quality-adjusted life year (QALY) weight per decision for the severity modifier to that under end of life.

NICE collected data from their experience of a full year of appraisals conducted using the new manual and reported this at its board meeting held in December 2023. It covered 52 topics that had been to at least one committee meeting between October 2022 and September 2023. It was reported that use of the severity modifier had differed from what had been expected based on historic data. The purpose of this report is to validate this finding and examine potential reasons for any disparity between the expected and observed rate at which appraisals would use the severity modifier for QALY weights above 1, and to outline the extent to which these issues do (or do not) impact the observed differences.

First, we describe the samples that were used to calculate the severity weights and the sample of post 2022 appraisals where they have been applied. Then we examine the impact of amending the calculation methods in the calculation sample, so they are consistent with how the post 2022 assessments have been conducted. The report then compares the appraisal samples in terms of easily identifiable characteristics and suggests other issues that could be important to consider in follow-up work.

Note that this report includes no view about the appropriateness or otherwise of considering severity according to AS and PS, or the categories defined in the NICE manual.

2. DATA AND METHODS

2.1 Primary Calculation Sample

Initially, NICE calculated AS and PS cut-offs based on extraction of information from 364 decisions from 218 technology appraisals where the guidance was issued between April 2011 and November 2019 in which AS/PS could be calculated. A total of 570 decisions from 323 appraisals were in the dataset in this period meaning that 36% decisions could not have AS/PS calculated. 20 decisions (3.5%) were not

classified as cost-utility analyses. Where AS and PS were not calculated, this was often because we could not identify reporting of the number of QALYs that were estimated for current care, a necessary component for the AS/PS calculations. We attempted to record data that aligned to the stated preferred committee case but this was often not possible either because no such scenario was given by the committee or we could not identify baseline age or comparator QALYs for the relevant scenario. Appraisals that were not conducted using cost utility analyses would not have these calculations, for example. This is an important limitation to note when making comparisons between severity weights in appraisals undertaken in this period versus more recent appraisals where missing data for these reasons is much less frequent. Also note that the unit of analysis here is each decision, of which there may be multiple in any appraisal. Appraisals can comprise multiple decisions where there are different patient subgroups considered separately, including those subgroups defined by the comparator that would be considered relevant (e.g. those intolerant to the comparator in widespread use), or where there are multiple technologies under consideration including those cases where a technology is considered either in combination with other agents or as monotherapy.

We refer to this as the ***primary subsample (n=364).*** This refers to those appraisal decisions where severity has been calculated.

2.2 Additional Calculation Sample

Following consultation on the proposed changes to methods an additional 100 decisions were added from 51 appraisals that covered the period Jan 2009 to April 2011 and from Feb 2020 to March 2021 where AS and PS could be calculated. These were drawn from 175 decisions (83 appraisals) in total conducted during that time period. 43% of decisions did not have AS/PS calculated. 9 appraisal decisions (5.1%) were not classified as cost utility analyses. We refer to this as the ***+ subsample*.**

When these data are added to the primary subsample, this forms what we refer to as the ***primary+ subsample (n=464).***

2.3 Implementation sample

The manual, which includes NICE guidance on the use of the severity modifier, was published in January 2022[1]. The methods it describes were applied to all new evaluations that had not commenced prior to its publication. For the purposes of this analysis, NICE provided data from published technology appraisals which cover the period up to March 31st, 2024. 64 appraisals covering 85 decisions were included. 16 decisions (19%) were from appraisals conducted using cost-comparison methods.

There are 47 appraisals covering 68 decisions in this subsample where AS and PS have been calculated. One non-lifetime horizon model was excluded here because severity was not calculated for this appraisal.

This is referred to as the ***implementation subsample***.

Comparisons are made between the distribution of AS, PS and overall severity categories (as per Table 1 above) using simple descriptive statistics. We refer to these as Category 1, Category 1.2 and Category 1.7. Chi-squared tests are used to compare the distribution across severity categories in different subsamples. We also test differences in proportions for selected comparisons using the two-sample Z-test of proportions.

3. THE DISTRIBUTION OF SEVERITY CATEGORIES

Table 2: The distribution of AS/PS and combined severity categories in NICE appraisals

		Primary	+	Primary+	Primary+ excluding non-lifetime models	Implementation
	n	**364**	**100**	**464**	**423**	**68**
Absolute shortfall						
	Mean	9.39	9.00	9.31	9.04	7.79
Cat 1 (AS<12)		73.9	73.0	73.7	74.7	94.1
Cat 1.2 12≤AS<18)	**%**	24.5	24.0	24.4	24.4	5.9
Cat 1.7 (AS≥18)		1.7	3.0	1.9	1.0	0.0
Proportional shortfall						
	Mean	0.62	0.58	0.61	0.60	0.62
Cat 1 (PS<0.85)		71.7	71.0	71.6	71.6	72.1
Cat 1.2 (0.85≤PS<0.95)	**%**	21.4	22.0	21.6	21.3	19.1
Cat 1.7 (PS≥0.95)		6.9	7.0	6.9	7.1	8.8
Overall						
	Mean weight	1.119*	1.106	1.116*	1.111	1.103
Cat 1		61.3	66.7	62.5	63.4	70.6
Cat 1.2	**%**	30.5	25.3	29.3	29.1	20.6
Cat 1.7		8.2	8.1	8.2	7.6	8.8

* These figures (including sample sizes) precisely align with those reported by NICE in Table 2 of the document "Review of methods, processes and topic selection for health technology evaluation programmes: conclusions and final update Appendix: Further discussion and rationale for conclusions – methods"

Table 2 shows that there are no marked differences in the distribution of appraisal decisions to the three AS and PS categories between the primary and primary+ subsamples. This aligns with NICE's own conclusions in their response to consultation on the draft methods manual (NICE undated document). The primary subsample did contain a slightly higher proportion of appraisals where the severity weight of 1.2 would have applied (30.5% in the primary vs 25.3% in the + subsamples). In all of these subsamples, there are relatively few appraisals where the AS weight of 1.7 would apply (AS≥18) (1.9% in the primary+ subsample for example). A higher proportion of appraisals fall into the highest (1.7) category of severity based on PS (PS≥0.95) (6.9% in the primary+ subsample).

We undertook an additional amendment to the appraisal dataset that underpins these analyses. The original data extraction did not routinely categorise whether the decision models used adopted a lifetime horizon or not. This is potentially important because the calculation of both AS and PS assesses the severity shortfall by comparing QALEs for the otherwise healthy population (conditional on age and sex) to those under current NHS standard care. This is taken directly from the comparator arm of the economic analysis but where this does not adopt a lifetime horizon, the comparison would be potentially misleading. We re-examined all appraisals and categorised all appraisal decisions into those where a lifetime horizon was explicitly stated to have been taken, those where it was not, and those where this was unclear. Typically, this information was found in the Evidence Review Group (ERG) report for Single Technology Appraisals (STAs) but often we had to consult other documents, such as the manufacturer/company submission, to identify this information. Including only those decisions that were explicitly judged and stated to have adopted a lifetime horizon reduces the primary+ subsample from n=464 to n=423. Table 2 shows how the distribution of AS and PS severity categories changes in this reduced appraisal decision subsample. PS is unaffected. There are fewer appraisal decisions in the 1.7 category of AS in this reduced subsample (1.0% vs 1.9% in the primary+ sample). This occurs because the 1.7 category of AS severity is highly related to technologies for younger patients: the mean starting age is 10.1 years in the 9 appraisal decisions that were in this category. Other AS categories both had a mean age higher than 50 years. This relationship does not occur for PS: the mean age by PS severity category is 50, 59 and 53 years for categories 1, 1.2 and 1.7 respectively. Those appraisal decisions that stated explicitly that a time horizon less than lifetime was adopted, were in younger patient groups compared to those that stated a lifetime horizon was used (38.5 vs 53.9 years). Note that the implementation sample excludes the single appraisal decision that was conducted using an analysis that was less than a lifetime.

Table 2 also reports the distribution of decisions falling into each severity category in the post 2022 implementation period. There are 68 decisions for which the category of severity weight was recorded. 14 (20.6%) are in overall severity category 1.2 and 6 (8.8%) are in category 1.7. Of these 20 appraisal decisions that were deemed to meet the criteria for a severity weight above 1, 16 (80%) were on the basis of PS alone i.e.

they were in AS category 1. Of the remaining 4, 3 (75%) were in category 1.2 for both PS and AS.

A Chi-squared test fails to reject the null hypothesis that there is no difference in the distribution of decisions by severity categories between the implementation sample and the primary+ sample (p=0.321). There is no statistically significant difference between the samples for the distribution of PS severity (p=0.218) but the difference in AS severity is statistically significant (p=0.001).

There is no statistically significant difference (at the 5% level) between the proportion of appraisals meeting the criteria for a 1.2 weight in this implementation sample (14/68 = 20.6%) versus the 29.3% in the primary+ sample (p=0.135). There is no difference between the proportion of appraisals meeting the criteria for a 1.7 weight (6/68 = 8.8%) versus the 8.2% seen in the primary+ sample (p=0.862).

3.1 Alternative calculation methods

The original analysis for the calculation of AS and PS was initiated in 2016 and used the following calculation methods. Life expectancy by age and sex was based on national period life tables for the UK for the year 2014 published by ONS. Health utility by age and sex was based on a linear regression fitted to data from pooled 2003 and 2006 Health Survey for England (HSE) (n=26,679) responses reported by Ara and Brazier[3] (2010). Quality Adjusted Life Expectancy (QALE) for the general population was estimated as the life expectancy, conditional on starting age, with each year adjusted for quality and discounted at 3.5% per annum. The distribution of sex in the relevant patient group was not extracted from the appraisals. Instead, it was assumed that there was a 50:50 split between the two sexes for the purpose of calculating AS and PS.

This approach uses the expectation of life (e^x) for a given age (x) from life tables, that is, the mean additional number of years an individual of a given age and sex could expect to live.

The calculation methods for severity that are recommended in the recent DSU Technical Support Document 23[4] differ from the 2016 analyses. We have not conducted a formal review of the calculation methods that have been used in post 2022 appraisals, but we believe they are consistent with TSD23 based on discussions with NICE staff. Life expectancy by age and sex uses ONS period life tables for

England for 2017-2019. Health utility by age and sex is based on HSE data from 2014, the most up to date, large scale population source for EQ-5D-3L and consistent with the DSU report on this issue[5]. The calculation methods use a lifetable model approach. Given the starting age and sex of the patient population, the method calculates the proportion of people alive in the subsequent year and applies the relevant utility weight to this proportion. The process continues to the age of 100. A discount rate of 3.5% per year is applied.

This approach uses the mortality rate (q^x) between age *x* and *x+1* published in national Life Tables. A difference in the estimated discounted QALEs associated with using the mortality rate versus the expectation of life approach will be observed where there is a nonlinear relationship between life expectancy and discounted QALEs.

In summary, there are three sources of differences in how the AS/PS weights were calculated (for shorthand we refer to these as the **"2016"** calculations) and the post 2022 implementation of those weights (the **"2022**" calculations):

i) Life expectancy estimates by age and sex
ii) Health utility by age and sex
iii) The method for calculating QALE by age and sex

We illustrate the impact of each of these three differences, both individually and cumulatively, on the estimates of AS and PS and the proportions of sampled appraisals falling into each of the three categories of AS, PS and weight for severity overall. These results are reported in Table 3 for the primary sample and Table 4 for the + sample.

3.1.1 Calculation method (Option 1)

Table 3 shows that, in the primary subsample, there is a small impact of using the updated calculation method compared to the original approach on the distribution of the proportion of guidance decisions that would fall into each of the three overall severity weight categories. For AS, there is no change to the small proportion in the highest (1.7) severity category, but there is a substantial reduction of the proportion in the middle (1.2) category (from 24.5% to 12.6%) and a consequent increase in the proportion of decisions which did not meet the criteria for an additional AS QALY weight (73.9% vs 85.7%). Changes to the categorisation of appraisal decisions based

on PS are smaller, though there is some impact on the proportion of decisions that have a PS exceeding 0.95 (1.7 weight) (6.9% vs 5.8%). The overall categorisation for NICE decision making purposes is more closely aligned to PS than AS. Therefore, the impact of the different calculation method is relatively small. There is a reduction in the proportion of appraisals in the 1.7 weight category (8.2% vs 7.1%) and a reduction in those in category 1.2 (30.5% vs 25.3%).

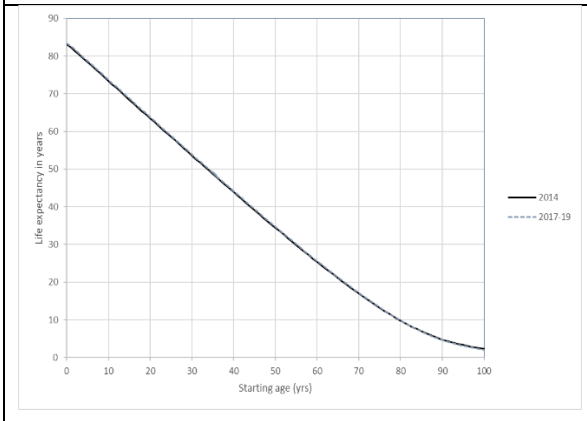
A similar pattern is shown in the +subsample (see Table 4). For AS, there is no impact on the proportion in the 1.7 category, and a reduction in the proportion in the 1.2 category (24.0% vs 12.0%). However, there is no impact on the distribution of PS categories and therefore only a marginal impact on the overall proportion of decisions where the severity modifier would be applied either at 1.2 or 1.7.

3.1.2 Life tables (Option 2 – new method, new life tables, old utility)

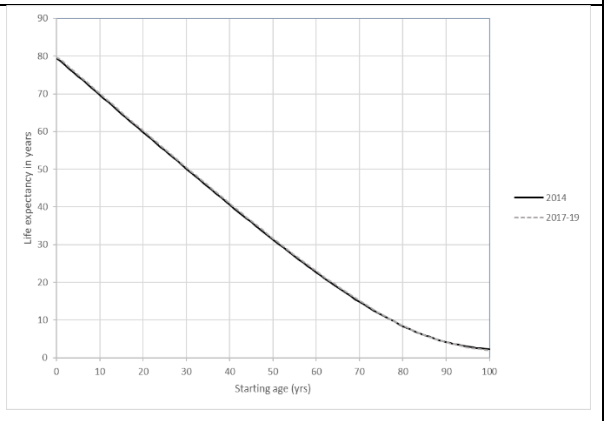
Figure 1 reports estimated life expectancy for a) females and b) males, comparing the 2014 and 2017-2019 ONS values. There are only very slight differences between the two different ONS dates, such that the plots are barely distinguishable from one another. For females the greatest difference in life expectancy is 0.37 years in favour of 2017-2019, which occurs at age 4 years. Interestingly, from the age of 84 years, life expectancy was higher using the 2014 data. For males, the maximum difference is 0.46 years, which also occurs at age 4 years.

Figure 1: Life expectancy by age (years) for a) females and b) males

a) Females



b) Males



The impact of this on the distribution of decisions to AS and PS categories is shown under the columns headed “Option 2” in Table 3 (for the primary sample) and Table 4 (for the + sample). Again, the same pattern is shown in both samples. Comparing option 1 and option 2, which differ only in terms of the life tables used, shows negligible impact. For AS, there is no impact on the proportion in category 1.7 and a small increase in the proportion in category 1.2 (12.6% vs 13.2% in the primary subsample and 12.0% vs 13.0% in the + subsample). PS changes are very small in the primary subsample and there is no impact in the + subsample. The same is true for the overall severity weight categorisation.

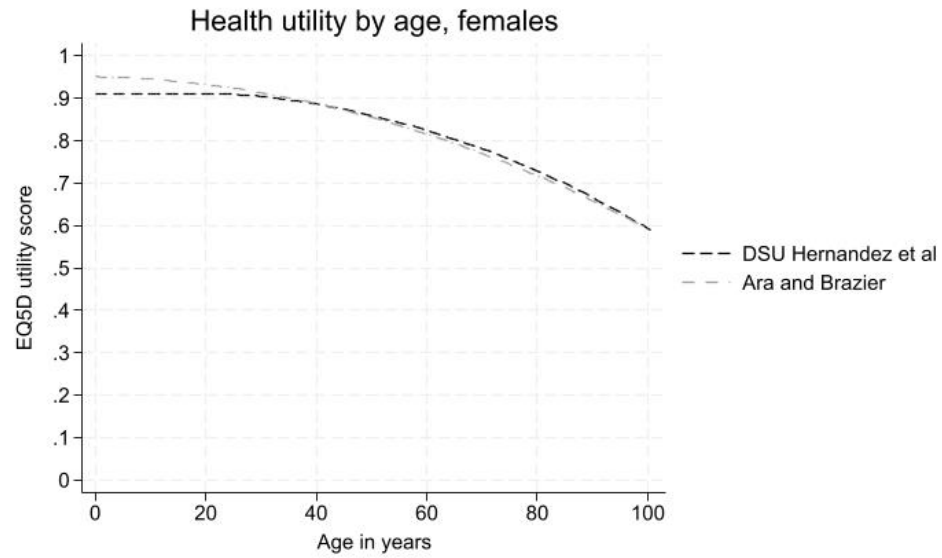
3.1.3 Health utility (Option 3 – new method, new utilities, old life tables)

Figure 2 compares the two sources of estimates for health utility by age. For females, the differences between the two sources are minimal above the age of 20 years. The Ara and Brazier estimates are higher than the DSU values at lower ages, though recall that, below the age of 16, the DSU recommendation is to use the same values as for 16-year-olds. This is because the Health Survey for England, which both studies base their analyses on though from different years, does not include those younger than 16 years in the sample. The DSU proposal is a pragmatic approach to the estimation of utilities for young age groups where extrapolation outside observed data is required.

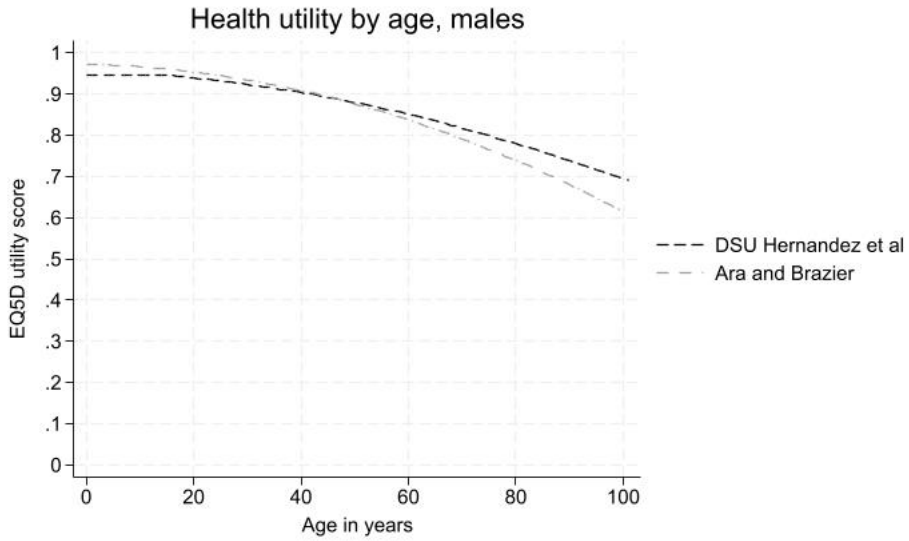
Differences are more marked between the estimates for males with the DSU model predicting higher utility values when age is greater than 40 years. The difference is greater as age increases.

Figure 2: Comparison of EQ-5D utility score estimates from the DSU and Ara studies, for a) females and b) males

a



b



The comparison with option 1 isolates the impact of the source of utility data alone. Table 3 shows that, in the primary subsample, this change increases the proportion of appraisals in AS category 1.2 by 2.2% with a corresponding reduction in the proportion of those in category 1. This pattern is seen for PS categories and the overall severity categories where a 1.6% increase in the proportion of decisions in category 1.2 is

offset by a reduction in category 1. These are relatively small changes. Table 4 shows a similar pattern for AS but there is no change to the distribution of overall severity categories (based on AS and PS combined).

3.1.4 Cumulative impact. New method, new data (Option 4)

Figure 3 displays QALE using the 2016 and 2022 set of different calculation methods. The 2016 estimates are consistently slightly higher than the 2022 estimates, at all ages up to 100 years. The mean difference is 0.87 QALYs (sd 0.216) and is relatively constant in absolute terms over all ages.

Figure 3: Discounted QALEs by age for 2016 and 2022 calculation methods and different data sources.

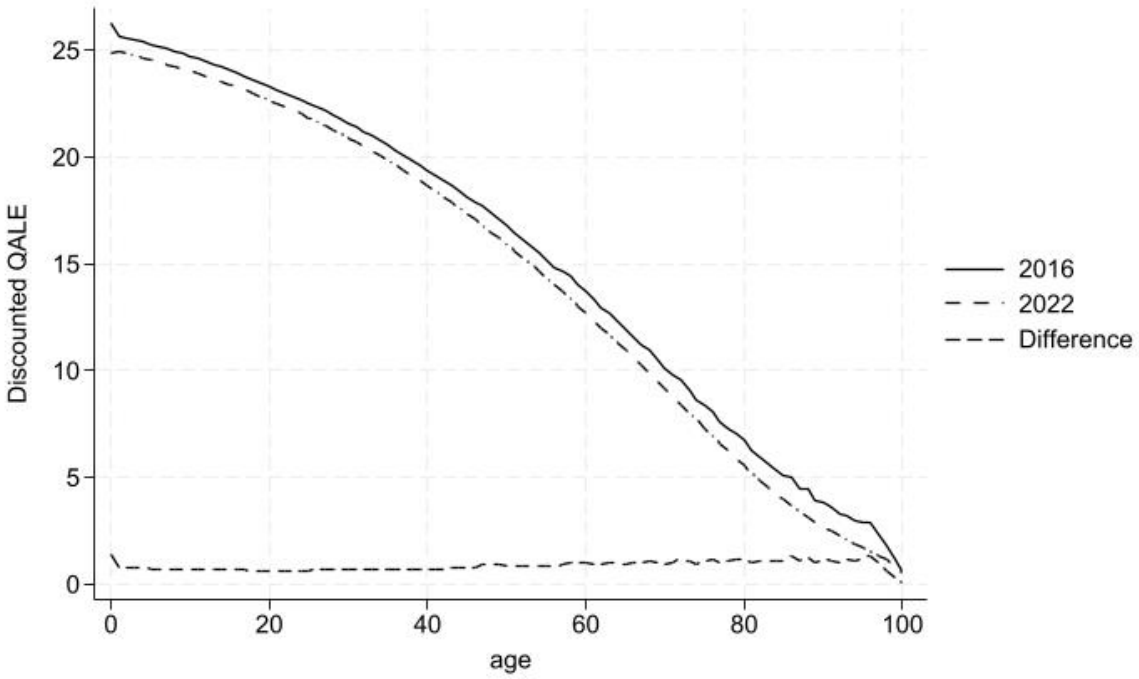


Table 3 shows that the cumulative effect of all changes remains very limited and, because the impacts of these different components work in opposite directions, the resultant distribution across severity weights categories is very similar to the original analysis. The most substantial differences occur in the categorisation of AS where the proportion of appraisal decisions in category 1.2 reduces from 24.5% to 17.3%.

This pattern is similar in the + subsample, where the reduction in the proportion of appraisals in the AS 1.2 category drops from 24% to 13%. The largest driver of this difference is the calculation method rather than the data sources.

Table 3: Impact of different AS/PS calculation methods on severity in the primary subsample

Primary subsample (n=364)						
		Original	**Option 1**	**Option 2**	**Option 3**	**Option 4**
Absolute shortfall						
	Mean	9.39	8.19	8.28	8.41	8.51
Cat 1 (AS<12)		73.9	85.7	85.2	83.0	81.0
Cat 1.2 12≤AS<18)	**%**	24.5	12.6	13.2	15.4	17.3
Cat 1.7 (AS≥18)		1.7	1.7	1.7	1.7	1.7
Proportional shortfall						
	Mean	0.62	0.59	0.59	0.59	0.59
Cat 1 (PS<0.85)		71.7	74.2	73.9	73.4	72.8
Cat 1.2 (0.85≤PS<0.95)	**%**	21.4	20.1	20.3	20.9	21.4
Cat 1.7 (PS≥0.95)		6.9	5.8	5.8	5.8	5.8
Overall						
	Mean weight	1.119*	1.101	1.101	1.104	1.108
Cat 1		61.3	67.6	67.3	65.7	63.7
Cat 1.2	**%**	30.5	25.3	25.6	27.2	29.1
Cat 1.7		8.2	7.1	7.1	7.1	7.1

Original = 2016 analysis method and data
Option 1 = 2022 analysis method, 2016 data sources
Option 2 = 2022 analysis method, 2022 life tables source, Ara and Brazier (2010) utilities
Option 3 = 2022 analysis method, 2016 life tables, New utilities.
Option 4 = 2022 analysis method, 2022 data sources

Table 4: Impact of different AS/PS calculation methods on severity in the + subsample

+ subsample (n=100)						
		Original	**Option 1**	**Option 2**	**Option 3**	**Option 4**
n		**100**				
Absolute shortfall						
	Mean	9.00	7.78	7.88	8.00	8.09
Cat 1 (AS<12)		73.0	85.0	84.0	84.0	84.0
Cat 1.2 12≤AS<18)	**%**	24.0	12.0	13.0	13.0	13.0
Cat 1.7 (AS≥18)		3.0	3.0	3.0	3.0	3.0
Proportional shortfall						
	Mean	0.58	0.54	0.54	0.55	0.55
Cat 1 (PS<0.85)		71.0	71.0	71.0	71.0	71.0
Cat 1.2 (0.85≤PS<0.95)	**%**	22.0	22.0	22.0	23.0	23.0
Cat 1.7 (PS≥0.95)		7.0	7.0	7.0	6.0	6.0
Overall						
	Mean weight	1.107	1.103	1.103	1.103	1.103
Cat 1		67.0	69.0	68.7	68.7	68.7
Cat 1.2	**%**	25.0	23.0	23.2	23.2	23.2
Cat 1.7		8.0	8.0	8.1	8.1	8.1

Original = 2016 analysis method and data
Option 1 = 2022 analysis method, 2016 data sources
Option 2 = 2022 analysis method, 2022 life tables source, Ara and Brazier (2010) utilities
Option 3 = 2022 analysis method, 2016 life tables, New utilities.
Option 4 = 2022 analysis method, 2022 data sources

3.1.5 Summary of findings

None of the differences in calculation methods between the original NICE analysis and how severity weights are (assumed to be) calculated in post 2022 appraisals have a substantial impact on the distribution of appraisal decisions across the overall severity weight categories.

Changes to the calculation method have a greater impact than those related to data sources for either life expectancy or age and sex dependent health utility. However, the impact is greater in terms of the categorisation of AS and much smaller for the final severity weights used in decision making (because PS categorisation is more closely linked to the overall severity weight than AS). The difference between the distribution in AS categories under Option 4 and the distribution in the implementation subsample remains a statistically significant one (p=0.028). The evidence suggests that these differences do not explain the experience of applying severity weights in real NICE appraisals post 2022. This remains the case when replicating these calculation methods and applying them only in the primary+ subsample where we can be certain the models were conducted using a lifetime horizon.

This suggests that observed, non-statistically significant differences in the distribution of severity weights pre and post 2022 are the result of differences between the appraisal decisions that are included in the subsamples. The issue of calculation methods may be something that should be revisited in subsequent analyses after examining differences in appraisal decision sub-sample characteristics, particularly in relation to AS.

3.2 Sampling issues

It is important to reiterate the degree of missingness in the primary and + samples. In the primary subsample AS and PS were not calculated in 36% of decisions. If these missing cases are systematically different to those included in the analysis in relation to the distribution of severity, then comparisons with the post 2022 implementation sample, where missing data is less of an issue, may be undermined.

A further, related consideration is the potential impact of cost-comparison analyses. In the primary sample, these would be part of the missing data for AS and PS. There are 16/83 (20%) such cases in the implementation subsample.

There are many other characteristics that could define appraisals and potentially impact the distribution of severity. However, to compare the primary / primary+ and implementation samples may require significant resource to extract relevant information from the large number of appraisals in each sample. For this reasons, detailed analyses of sample characteristics are not reported here. Comparisons are undertaken between samples in terms of appraisal decision categorised as meeting the end of life criteria. We also consider evidence on changes over time or by age within the primary+ sample.

3.2.1 End of Life (EoL)

463 appraisal decisions had information on both whether the topic was considered to meet the EoL criteria and had data available for the calculation of severity (see Table 5). 81 (17.5%) met the EoL criteria in the primary+ subsample. Of these 81 EoL cases, 43 (53%) had AS≥12, 80 (98.8%) had PS ≥0.85, and 74 (91.4%) attract a severity weight above 1. Only one EoL case would have met the 1.7 weight severity category on the basis of AS. All of those in the 1.7 category of severity overall, including this case, qualified on the basis of PS.

There were proportionally more EoL topics in the Implementation subsample (26.5% vs 17.5%, p=0.075). Since EoL was not formally assessed in these appraisals, this is based on the judgement of NICE staff retrospectively considering the available data and may therefore be an overestimate. These judgments were performed with the intention to avoid underestimation of the frequency with which EoL may have occurred. These judgements were also made on the basis of data within relevant clinical trials, and cost effectiveness modelling, and had decisions been based only on cost effectiveness modelling this would have substantially reduced the number of topics to have been considered to have met EoL. It is feasible that, had NICE committees considered EoL in these appraisals, they would have drawn on wider evidence and reached different conclusions. However, there are additional inconsistencies in this categorisation. EoL as originally defined by NICE in 2009 included the criteria that "The treatment is licensed or otherwise indicated, for small patient populations." There are appraisals included in the primary+ sample that were deemed not to meet the EoL criteria because of this criterion (e.g. TA242). Furthermore, there are examples of appraisals where the committee stated that the EoL criteria were met despite the economic model producing results contradictory to this (e.g. TA509, TA540).

When comparing the implementation subsample with the primary+ subsample, proportionately there were fewer of these EoL appraisal decisions in the 1.2 severity weight category (61.1% vs 71.6%) but more in the 1.7 category (33.3% vs 19.8%). The mean weight for severity in appraisals classed as meeting the EoL criteria is higher in the implementation subsample compared to the primary+ subsample (1.36 vs 1.28).

EoL is more strongly associated with higher PS (correlation 0.51) than higher AS (0.31) and this closer relationship to PS than AS categories is also apparent in Table 5.

Table 5: Severity category for appraisals meeting End of Life Criteria

	Primary+ (n=464*)				Implementation (n=68)			
	EoL Yes		EoL No		EoL Yes		EoL No	
n	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Total	81	17.5	382	82.3	18	26.5	50	73.5
Absolute shortfall								
Cat 1 (AS<12)	43	53.1	299	78.3	15	83.3	49	98.0
Cat 1.2 (12≤AS<18)	37	45.7	75	19.6	3	16.7	1	2.0
Cat 1.7 (AS≥18)	1	1.2	8	2.1	0	0.0	0	0.0
Proportional shortfall								
Cat 1 (PS<0.85)	11	13.6	321	84.0	1	5.6	48	96.0
Cat 1.2 (0.85≤PS<0.95)	54	66.7	45	11.8	11	61.1	2	4.0
Cat 1.7 (PS>0.95)	16	19.8	16	4.2	6	33.3	0	0.0
Overall								
Mean weight	1.281		1.081		1.356		1.012	
Cat 1	7	8.6	283	74.1	1	5.6	47	94.0
Cat 1.2	58	71.6	77	20.2	11	61.1	3	6.0
Cat 1.7	16	19.8	22	5.8	6	33.3	0	0.0

- For one decision, the appraisal committee did not consider EoL (TA192)

3.2.2 Changes over time

We looked for evidence of change in the distribution of weights for severity over the time period covered by the primary and + subsamples. Figures 4-6 show mean overall weight for severity, AS and PS by year, covering the period 2011 to 2020. We excluded other years from these plots due to small numbers of observations.

These analyses show that there is no clear increasing or decreasing trend over the whole period in these outcomes. However, there is some evidence of a decrease from 2017 to 2020 in all three measures. AS and PS show very similar patterns over time, which is not surprising since there is a high degree of correlation between them (0.837). Figure 7 plots the proportion of appraisals that were deemed to meet the EoL criteria. This plot includes all appraisals, whether AS and PS were calculated for them or not, but a separate analysis including only those where AS and PS were also available shows the same pattern. The relationship with EoL is interesting because it demonstrates a similar pattern to the measures of severity based on AS and PS, though a large increase in 2020 lends to an overall impression of a trend of growth over time.

The analyses also highlight the degree of variability across years, reinforcing the finding that the results from the implementation subsample are not anomalous given the sample size.

There are five years (from 10) where the 95% confidence interval spans the mean AS in the implementation subsample. Mean AS in 2010 (n=39) is below mean AS in the implementation subsample.

There are five years (from 10) when PS is below the mean observed in the implementation sample and nine years when it is within the 95% confidence interval.

There are three years (from 10) when the mean weight for severity is below the mean observed in the implementation sample and nine years where it lies within the 95% confidence interval.

Figure 4: Mean (95% Confidence intervals) QALY weight in appraisals by year

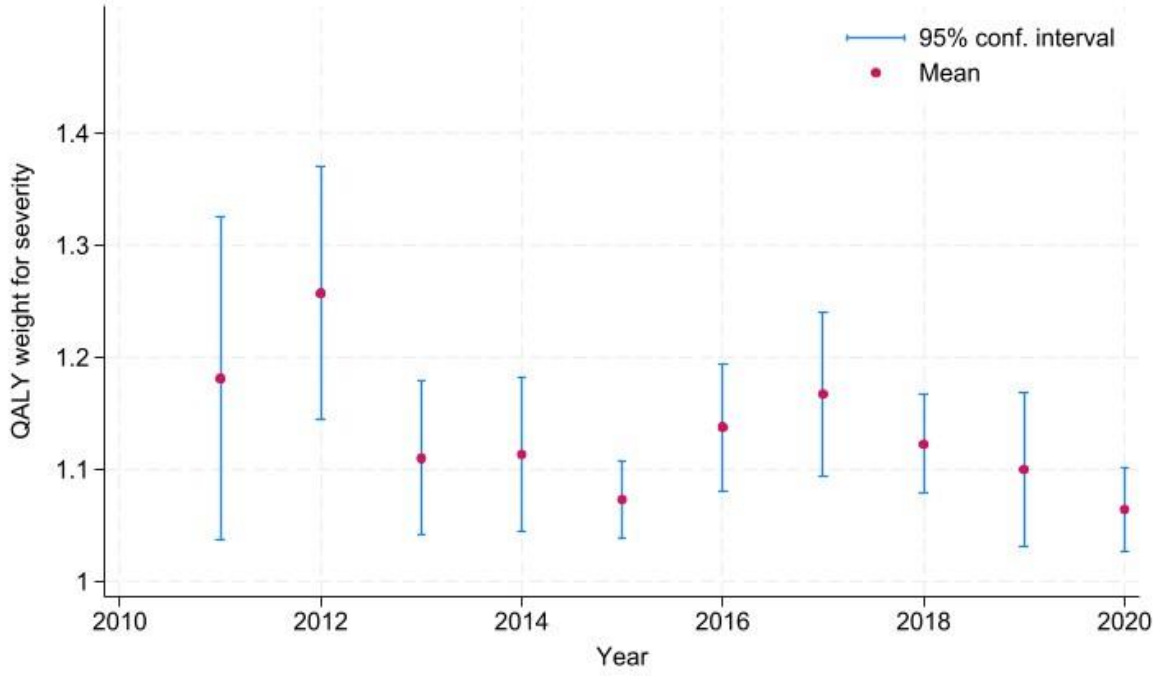


Figure 5: Mean (95% Confidence intervals) Absolute Shortfall in appraisals by year

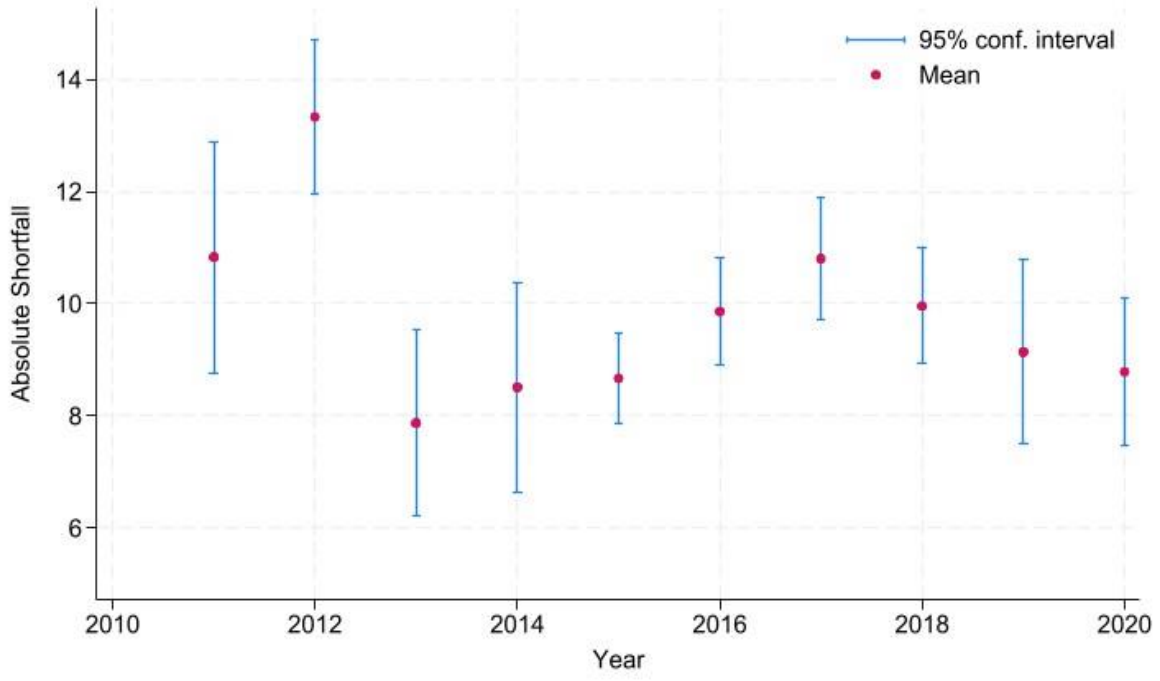


Figure 6: Mean (95% Confidence intervals) Proportional Shortfall in appraisals by year

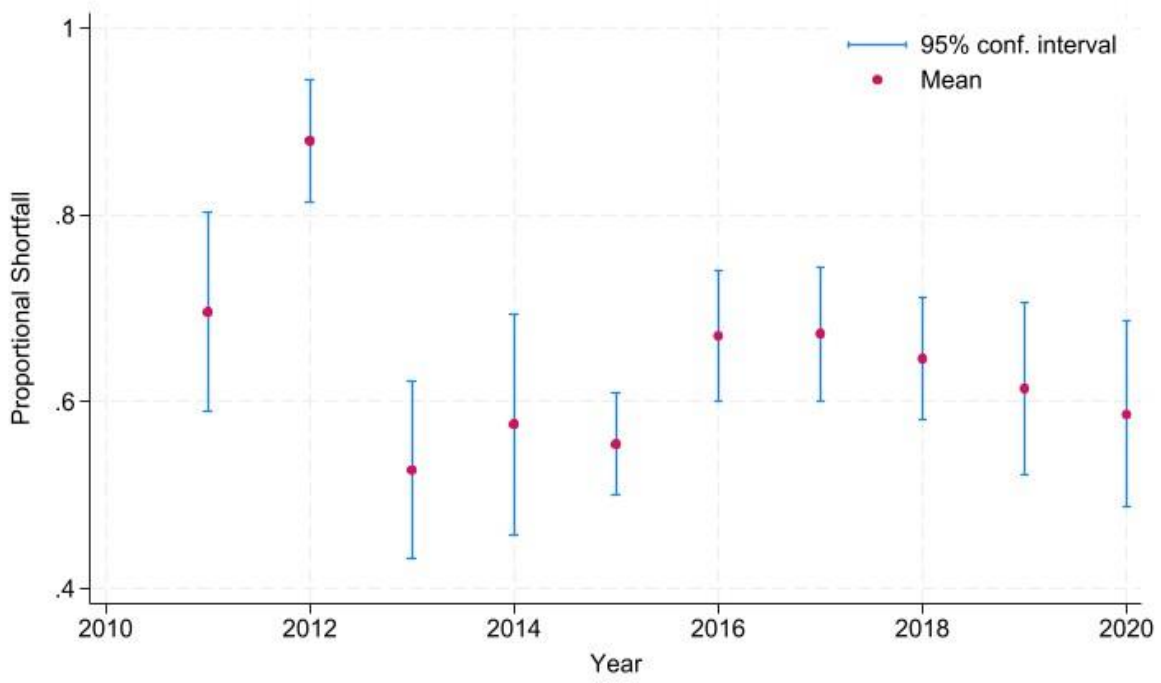
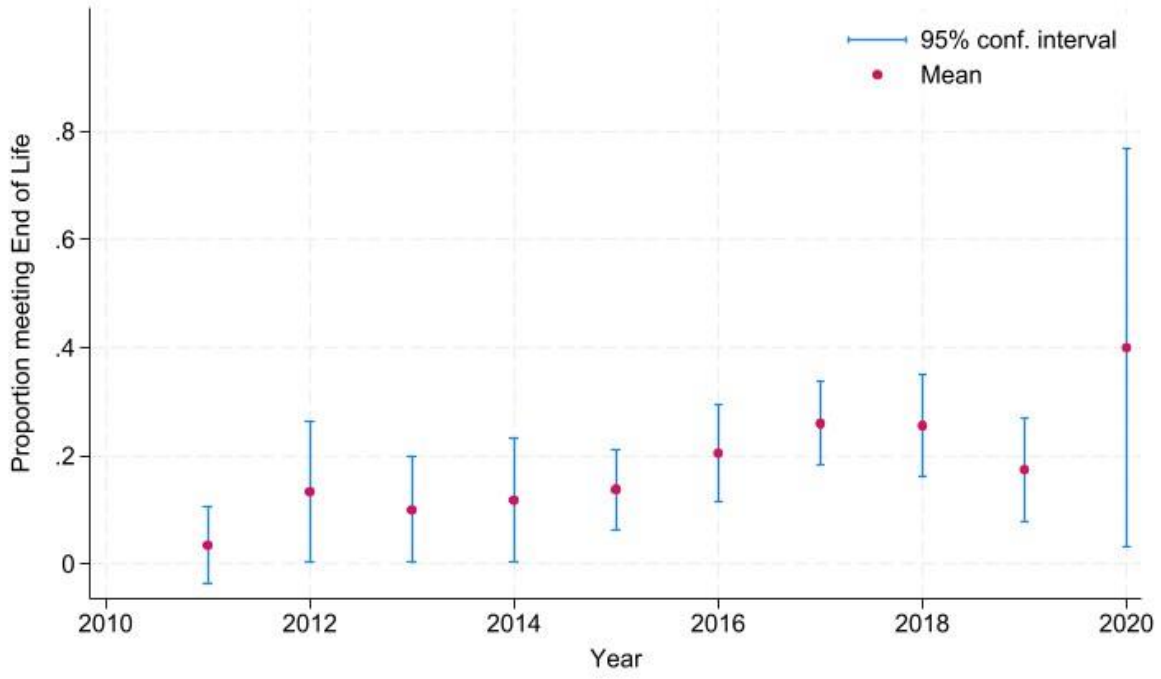


Figure 7: Mean (95% Confidence intervals) proportion of appraisals meeting EoL criteria by year



3.2.3 Age

At the time of writing, age has not been extracted for the implementation sample. Therefore, comparisons between the ages of patients in the appraisal samples cannot be undertaken. However, age is likely to be an important factor to consider. Figure 8 and Figure 9 illustrate how the AS and PS NICE categories are related to age. Figure 8 plots the maximum number of QALYs patients can obtain under current care, at each age, and generate AS=12 QALYs. Above the line indicates those combinations of age and comparator QALYs that would generate less than 12 QALYs for AS (i.e. be considered less severe using this definition alone). Below the line are combinations of age and comparator QALYs that generate more than 12 QALYs i.e. are more severe using this definition. Similarly, the line for PS indicates those combinations of age and comparator QALYs which generate PS=0.85. Above this line, PS is lower than 0.85 and below the line PS is higher. The frontier of the AS and PS functions therefore delineates the combinations of age and comparator QALYs required for a technology to meet the criteria for the overall severity modifier category 1.2 (or above).

There is a stronger relationship between age and comparator QALYs for AS than there is for PS. At young ages, the number of QALYs that patients can receive can be relatively high and still generate category 1.2 severity weight via AS. For example, at age 5 years, patients could still expect to receive over 12.5 discounted QALYs under current care and be assessed as category 1.2 severity because AS>12. Current care is required to far less effective for patients at these ages to have PS≥0.85. Up to the age of 55 years, any patient group assessed as meeting category 1.2 for PS would, by definition, also qualify for category 1.2 AS. Because there is a much weaker relationship between PS and age than AS and age, this switches at age 55 years.

Also note that at approximately 62 years, remaining discounted quality adjusted life expectancy falls below 12. At this point it is not feasible for AS to be above category 1 (there are some theoretical exceptions to this).

Figure 8 - The maximum number of comparator QALYs compatible with AS = 12 QALYs and PS = 0.85, by age (years)

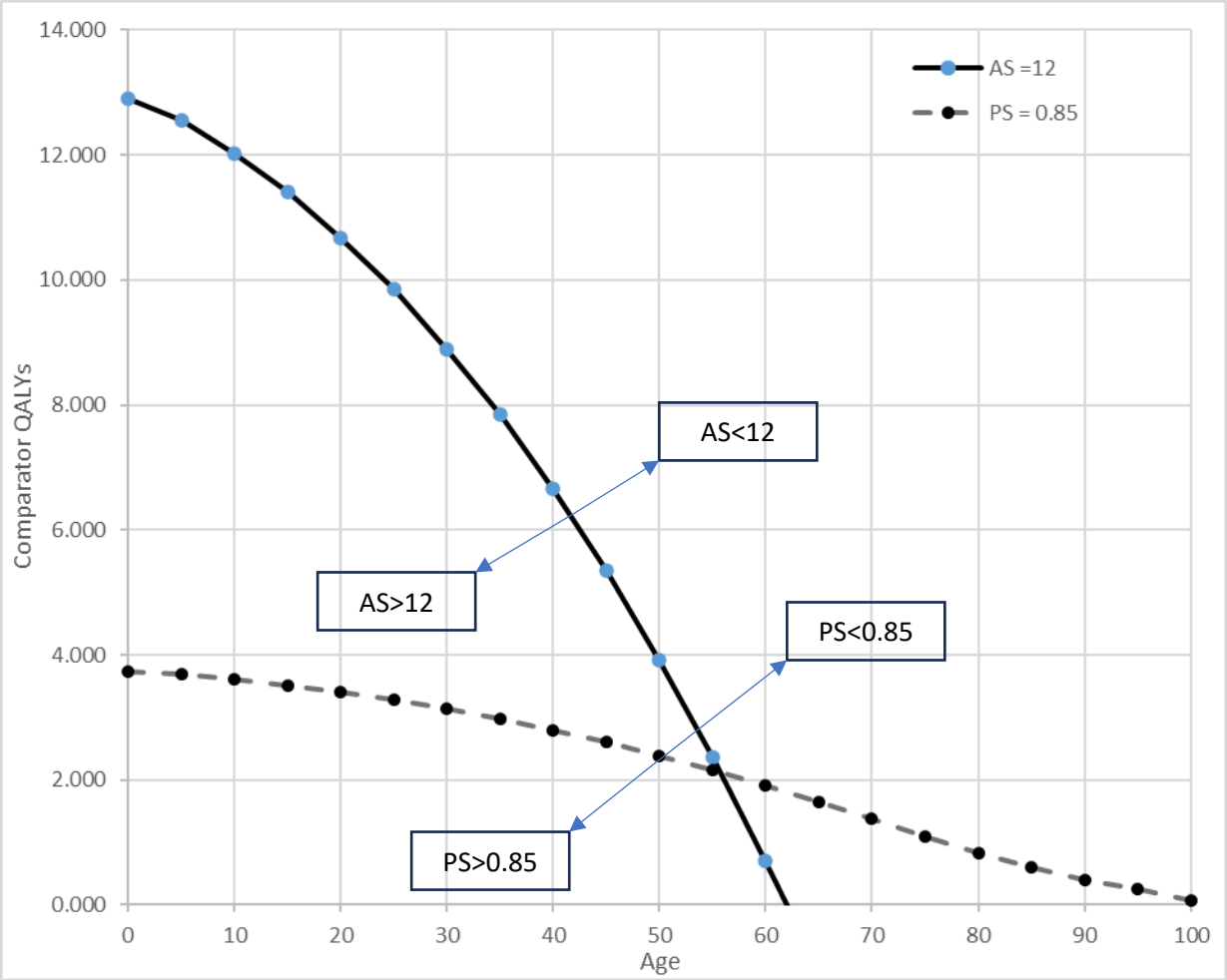
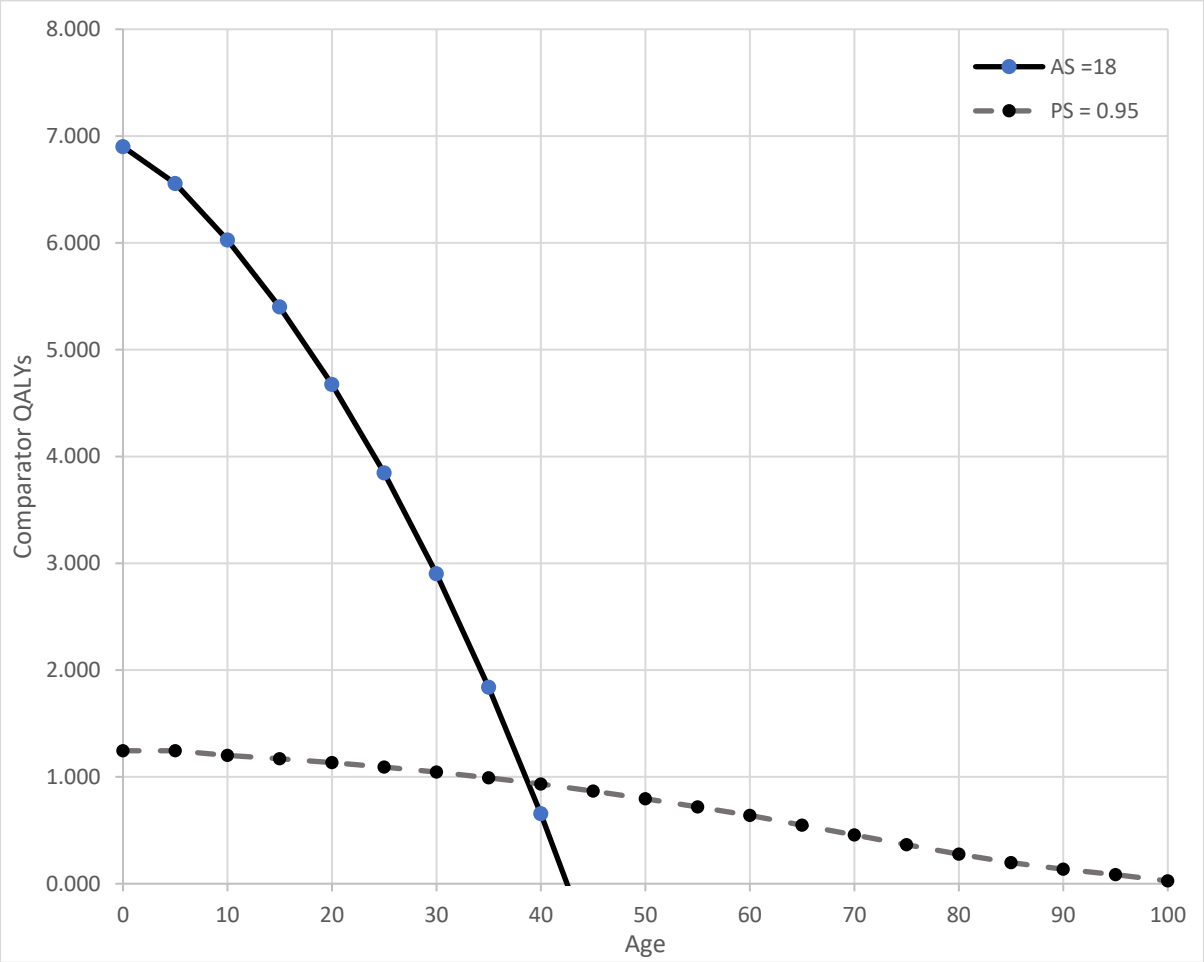


Figure 9 is similar but shows the combinations of age and comparator QALYs that generate AS =18 and PS =0.95. As age increases, the maximum number of QALYs expected for current care rapidly falls for AS but the relationship is more shallow for PS. The AS and PS lines intersect at age 48 years approximately.

Figure 9 - The maximum number of comparator QALYs compatible with AS =18 QALYs and PS = 0.95, by age (years)



These two plots demonstrate how AS and PS interact in their relationship with age. The complexity of this relationship is further demonstrated when looking at evidence from the primary+ subsamples, shown in Figure 10 and Figure 11 for AS and PS respectively. Figure 10 is a scatterplot of AS by age for the primary+ subsample, with a local smoother fitted. Values for AS=12 and 18 are also plotted. It demonstrates that there is a trend for AS to reduce with age, but that the relationship is flat, on average, between the ages of 40 and 60. All decisions where AS exceeds 18 are in those appraisals with a starting age below 20 years. For AS over 12, there are a large number of appraisals where the mean age lies between 40 and 60 years.

For PS, the relationship is more complex. On average, PS is broadly flat at around 0.55 where patient starting ages are below 40 years. There is then an increase in average PS that peaks soon after 60 years. PS above 0.85 and 0.95 occurs across the age distribution but the greatest frequency occurs between 40-65 years.

Figure 10: Absolute Shortfall by Age (years) in the primary+ subsample with LOESS smoother

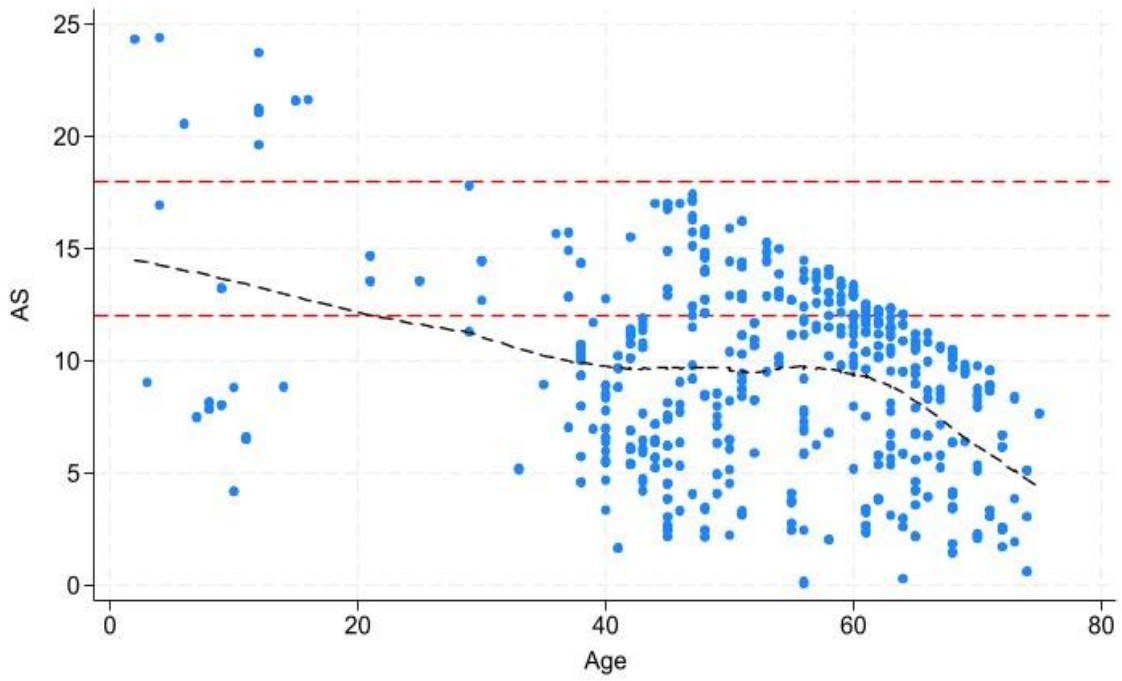
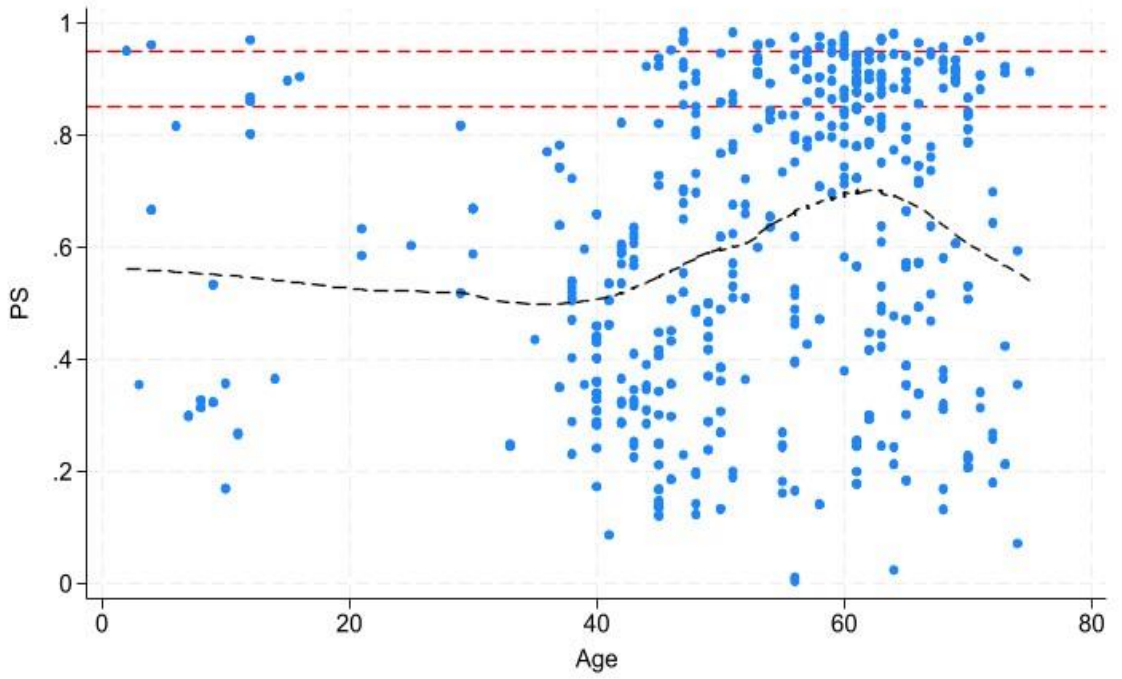


Figure 11: Proportional Shortfall by Age (years) in the primary+ subsample



This pattern is further emphasised in Figure 12 and Figure 13. Of particular note is the concentration of decisions with a starting age in the model between 45 and 65 years where AS falls into category 1.2.

Figure 12: Histogram of age distribution by AS category.

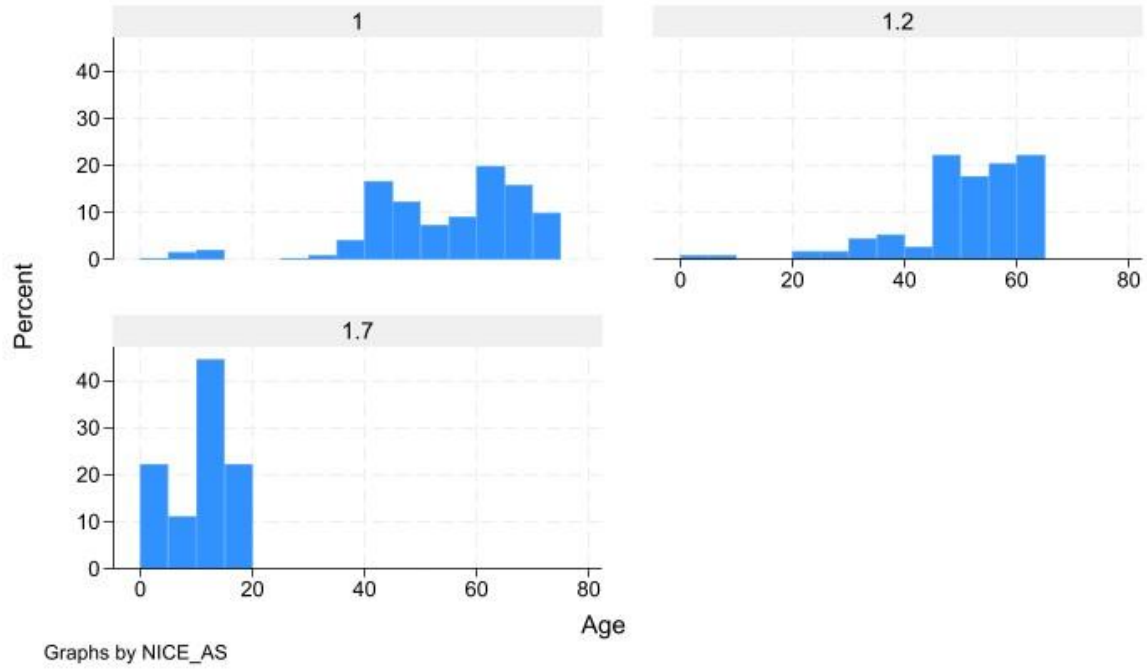
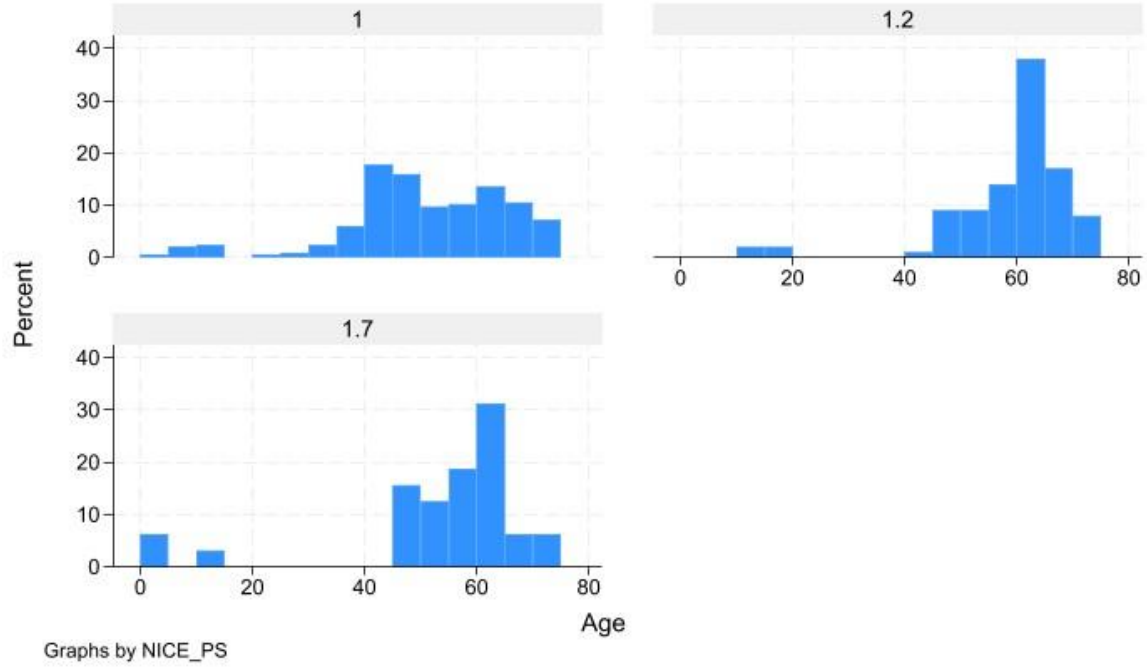


Figure 13: Histogram of age distribution by PS category.



4. DISCUSSION

There are concerns that the proportion of appraisal decisions that attract a severity weight, following the implementation of the changes in the 2022 NICE technology evaluations manual, differs from the corresponding proportion in the sample in which the weights and cut-offs were devised. However, only a small number of appraisals (n=47 covering 68 decisions) have been conducted using the new manual for which AS and PS have been calculated, and observed differences in proportions are therefore highly uncertain. The difference in the distribution of appraisal decisions across the three severity weight categories is not statistically significant. Comparisons with the primary+ subsample indicates that the largest difference is with the proportion of appraisal decisions that attract a severity modifier of 1.2. 20.6% of implementation sample decisions vs 29.3% in the primary+ sample is a relatively large difference, though not a statistically significant one. It is also notable that there is a higher proportion of appraisal decisions in the 1.7 severity modifier category (8.8% vs 8.2%), though again this is not a statistically significant difference. Overall, this implies that it may be too early to draw conclusions about differences in the distribution of severity weights in post 2022 appraisals from that observed in pre-2022 ones.

There are statistically significant differences observed between the implementation and primary+ subsamples in terms of the distribution across AS categories. This difference remains statistically significant using the cumulative set of new calculation sources and methods outlined in section 3.1 (p=0.028). Whilst of interest, it should be noted that severity weights and cutoffs were designed to achieve a broadly similar overall weighting as under end of life, not for the AS and PS components to remain similar across appraisal samples. Furthermore, in the implementation subsample, of the 10 appraisal decisions in the PS 1.2 category that were also in the AS 1 category, the mean AS was 10.5 QALYs (sd 0.31). 6 appraisals were in the PS 1.7 category and were also in the AS 1 category, but the mean AS was 11.7 QALYs (sd 0.30). Contrast this with the mean AS of 6.3 QALYs (sd 2.9) in the 48 appraisal decisions that were in both AS and PS = 1 categories. This suggests that the appraisal decisions that were close to the boundary of the AS 1.2 (12 QALYs) category were largely already assigned a weight of 1.2 or higher for the purposes of decision making because they qualified on the basis of PS.

If, for example, the threshold for the AS category 1.2 were as low as 10.2 QALYs an additional 20 appraisal decisions would then be in this category but only 5 of those would qualify for a higher overall weight as a result.

There is also some concern that these comparisons may be misleading for several reasons. Some of these were described in the original methods review board papers.

First, the extent of missing data should be noted. In the original primary and + samples, it was not possible to calculate severity shortfalls in approximately one third of cases, including cost-comparison cases that were introduced as an option by NICE in 2017. The implementation sample has 16/85 (19%) cases that are cost-comparisons where shortfall has not been calculated. There are different proportions of missingness between the two subsamples, and there are likely to be different reasons for this, that may impact the validity of making comparisons.

Second, both the primary+ and implementation subsamples extracted data that attempted to align to the preferred committee case, but this often required some judgement and was not always feasible. In numerous appraisals, the committee does not have an unambiguous preferred analysis or, where such a case does exist, the analysis does not appear in the set of evidence documents or only reports summary information (typically excluding comparator QALYs, a key component of the AS and PS calculations). Data extraction used judgement, and available analyses from assessment groups and manufacturers, to record the most suitable data.

Third, comparisons of data that categorise appraisal decisions to severity categories and EoL rely on data calculated from different processes. In the case of EoL, we reported how NICE staff classified appraisal decisions in the implementation sample according to the life expectancy reported in the appraisal and the addition to life expectancy from the technology in question. This is not the same as the primary+ sample, where this categorisation was undertaken as part of the appraisal process. The process ensured great scrutiny of these parameters from the Assessment Groups and the Appraisal Committee. The process also entailed the application of non-technical judgements by the committee which, on some occasions, resulted in great flexibility around the interpretation of the data and the stated criteria for EoL. Furthermore, for some of the period covered by the primary+ sample, EoL required the technology to be licensed for a small population. A similar caution applies to the

comparisons of severity weights across appraisals. There are obvious parallels with EoL because the categorisation of severity, both AS and PS, in the implementation sample will be subject to intense scrutiny given the impact on decision making. The importance of real world data sources versus clinical trials for estimates of comparator QALYs is just one example of this. In the historical primary+ samples, these categorisations have been made retrospectively and no such scrutiny applied. For these reasons, different appraisal samples may be incommensurable.

Fourth, the primary+ appraisal decision subsample spans back to 2009. There are likely to be many policy decisions taken within that period that we are unaware of, that may have impacted on the observed samples underpinning these data. This adds to the uncertainty of reported findings.

Despite these caveats, potential reasons for these observed differences were examined in two broad categories. First, that the calculation methods may differ and second that the samples do.

The report examines the impact of changing three aspects of the way in which AS and PS are calculated in the original primary and + subsamples compared to how we believe the implementation subsample weights have been. These are: the calculation method itself, the source of age and sex adjusted utilities, and the source of life expectancy estimates by age and sex. The conclusion is that these investigations rule out these changes, either alone or in combination, as the source of anything more than minor differences in the distribution of the overall severity categories. This conclusion is based on analyses that recalculated the proportions of appraisal decisions that would fall into each severity category (weights of 1/1.2/1.7), using the primary and + samples to illustrate the impact. There are larger differences when considering the distribution of AS.

Limited insight is provided by comparisons between the characteristics of the primary and implementation subsamples at this stage. We report summary statistics for whether appraisals met (or were judged that they would have met) the EoL guidance. The mean weight for severity in appraisals classed as meeting the EoL criteria is higher in the implementation subsample compared to the primary+ subsample (1.36 vs 1.28).

There is some limited evidence of a time trend of average AS and PS across appraisals. The observed reductions since around 2016/2017 perhaps indicate that further monitoring of these changes should be conducted once a larger number of appraisals has been conducted using the methods advised in the new manual.

Consideration should be given to other characteristics of appraisals that may be relevant for future analyses. Those factors that can be used to define an appraisal, pertaining to the characteristics of the patients affected or the nature of the technology under assessment, which are also plausibly related to AS and PS, should be considered. Age is one such characteristic that should be a prime candidate for consideration and we have provided evidence which shows that differences in ages between the ages of 45-65 years is likely to be the key driver of differences in AS. Other candidate characteristics include oncological versus non-oncological topics, and within oncological topics a classification of cancer stage may be informative.

Whether information on these or other characteristics can be easily extracted from both recent, post-2022 appraisals **and** the very large set of historical appraisals requires some judgement. It may be worthwhile for NICE to routinely collect this information for current appraisals, since the cost of collection is low, to allow limited, albeit non-comparable analyses to be conducted in future. The value of such a subsample could be realised once a larger set of appraisal decisions has been allowed to mature.

REFERENCES

[1] NICE (2022) “NICE health technology evaluations: the manual”, Last updated 31 October 2023. Available at: https://www.nice.org.uk/process/pmg36/resources/nice-health-technology-evaluations-the-manual-pdf-72286779244741 (last accessed 25/4/24)

[2] NICE (undated) Review of methods, processes and topic selection for health technology evaluation programmes: conclusions and final update Appendix: Further discussion and rationale for conclusions – methods

[3] Ara, R and Brazier, J. (2010) “Populating an Economic Model with Health State Utility Values: Moving toward Better Practice”, Value in Health, Vol.13: 509-18.

[4] Wailoo A (2024) Technical Support Document 23: A Guide to Calculating Severity Shortfall for Nice Evaluations. NICE DSU Technical Support Document 23. Available at: https://www.sheffield.ac.uk/nice-dsu/tsds/full-list (last accessed 25/4/24)

[5] Hernández Alava, M., Pudney, S., Wailoo, A. (2022) “Estimating EQ-5D by Age and Sex for the UK” NICE DSU report. Available at https://www.sheffield.ac.uk/nice-dsu/methods-development/estimating-eq-5d (last accessed 4/3/24)