TECHNICAL SUPPORT DOCUMENT 26: EXPERT ELICITATION FOR LONG-TERM SURVIVAL OUTCOMES

REPORT BY THE DECISION SUPPORT UNIT

March 2025

Jeremy E. Oakley[1*], Shijie Ren[2*], Jessica E. Forsyth[2],
John Paul Gosling[3], Kevin Wilson[4], Nick Latimer[2],
Mark J. Rutherford[5], Lesley Uttley[2], James Fotheringham[2,6]

[1] School of Mathematical and Physical Sciences, University of Sheffield
[2] School of Medicine and Population Health, University of Sheffield
[3] Department of Mathematical Sciences, Durham University
[4] School of Mathematics, Statistics & Physics, Newcastle University
[5] Department of Population Health Sciences, University of Leicester
[6] Sheffield Kidney Institute, Sheffield Teaching Hospitals NHS Trust
* Co-first authors

Decision Support Unit, SCHARR, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA

Tel (+44) (0)114 222 0734
E-mail dsuadmin@sheffield.ac.uk
Website www.nicedsu.org.uk
X @NICE_DSU

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) External Assessment Centre is based at the University of Sheffield with members at York, Bristol, Leicester 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 www.nicedsu.org.uk.

The production of this document was funded by 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.

ABOUT THE TECHNICAL SUPPORT DOCUMENT SERIES

NICE describes the methods it follows when carrying out health technology evaluations in its process and methods manual. This provides an overview of the key principles and methods of health technology assessment and appraisal for use in NICE appraisals. The manual does not provide detailed advice on how to implement and apply the methods it describes. The DSU series of Technical Support Documents (TSDs) is intended to complement the manual by providing detailed information on how to implement specific methods.

The TSDs provide a review of the current state of the art in selected topic areas. They make recommendations on the implementation of methods and reporting standards where it is appropriate to do so. They aim to provide assistance to all those involved in submitting or critiquing evidence as part of NICE technology evaluations, whether companies, assessment groups or any other stakeholder type.

We recognise that there are areas of uncertainty, controversy and rapid development. It is our intention that such areas are indicated in the TSDs. All TSDs are extensively peer reviewed prior to publication (the names of peer reviewers appear in the acknowledgements for each document). Nevertheless, the responsibility for each TSD

lies with the authors and we welcome any constructive feedback on the content or suggestions for further guides. The TSDs will be amended and updated whenever appropriate. Where minor updates or corrections are required, the TSD will retain its numbering with a note to indicate the date and content change of the last update. More substantial updates will be contained in new TSDs that entirely replace existing TSDs.

Please be aware that whilst the DSU is funded by NICE, these documents do not constitute formal NICE guidance or policy.

Professor Allan Wailoo
Director of DSU and TSD series editor

ACKNOWLEDGEMENTS

The authors would like to acknowledge Ruth Wong for her support in designing the search strategies for the broader literature. The authors would also like to thank the independent reviewers, Laura Bojke, Christopher Jackson, Hugo Pedder, Martine Barons, and Jacoline Bouvy, whose peer-review comments helped improve the clarity and quality of this document.

This report should be referenced as follows:
Oakley J. E., Ren S., Forsyth J. E., Gosling J. P., Wilson K., Latimer N., Rutherford M. J., Uttley L., Fotheringham J., NICE DSU Technical Support Document 26: Expert elicitation for long-term survival outcomes. 2025.
Available from http://www.nicedsu.org.uk

EXECUTIVE SUMMARY

This Technical Support Document (TSD) discusses how expert knowledge and uncertainty about long-term survival outcomes should be obtained and reported. This report will support health technology assessments in cases where appropriate data are lacking: where extrapolation is needed beyond the observed data. TSD 14 and TSD 21 have addressed model-based extrapolations, where it was noted that these can result in diverging long-term survival estimates, with significant implications for cost-effectiveness results. Both TSDs identified the role of expert judgement to support survival extrapolation. Here, we discuss how to obtain and use expert judgement.

The main methodological approach recommended is to elicit probability distributions from clinical experts, so that expert uncertainty is quantified, and experts are not merely asked to provide 'best estimates' or approve the clinical validity of pre-selected model-based survival extrapolations. There are various established protocols for eliciting probability distributions from experts in a structured manner, and such a protocol should form the basis of how the elicitation exercise is conducted. There are, however, aspects of the expert judgement task that are particular to survival extrapolation: the availability of data from which to extrapolate, and the way qualitative knowledge can be incorporated via the relationship between survivor and hazard functions. We discuss and illustrate how to modify a standard elicitation protocol accordingly. Software to support the elicitation process is also discussed.

We review the use of expert elicitation for long-term survival in health technology assessment as well as the broader literature and observe that the majority of NICE technology appraisals have not used structured expert elicitation. We set out recommendations for best practice and discuss future research directions for this methodology.

CONTENTS

FIGURES

TABLES

ABBREVIATIONS

DSU	Decision Support Unit
EPS	Equivalent prior sample
HTA	Health technology assessment
ICER	Incremental cost-effectiveness ratio
IDEA	Investigate, Discuss, Estimate, Aggregate
MRC	Medical Research Council
NICE	National Institute for Health and Care Excellence
RCT	Randomised controlled trial
RIO	Rational Impartial Observer
SEE	Structured expert elicitation
SHELF	Sheffield Elicitation Framework
TA	Technology appraisal
TSD	Technical Support Document

GLOSSARY OF TERMS

Definitions, where stated here, are not intended to be fully rigorous in the mathematical sense, rather the aim is that they are accessible to the reader. More precise definitions may be found within the referenced literature.

Behavioural aggregation/mathematical aggregation: These both refer to obtaining a single probability distribution from a group of experts.

- Behavioural aggregation refers to any process that involves interaction between the experts such that the experts are involved in deciding what the single probability distribution should be.
- Mathematical aggregation refers to any process in which each expert provides their own probability judgements, and then a formula is applied to obtain a single probability distribution from these.

Credible interval: An interval for an uncertain quantity, judged to contain that quantity with a specified probability. For example, a 95% credible interval for an uncertain quantity is an interval judged to have a 95% probability of containing that quantity. The word “credible” is used to distinguish it from a confidence interval, and credible intervals are typically used to refer to probability intervals computed from posterior distributions in Bayesian statistical inference.

Fixed interval method: An elicitation method in which an expert is provided with an interval, and asked for their probability of the uncertain quantity lying in that interval, e.g., “What is your probability that the uncertain quantity of interest would lie between 10 and 20?”

Hazard function: For time-to-event outcomes, informally, the hazard at time t is the risk ‘at that instant’ of the event occurring, given that event has not yet occurred by that time t. The hazard of death for a particular year would be the probability of an individual dying in that year, given that the individual is alive at the start of the year. The hazard function is related to the gradient of the survivor function: if the survivor function has a constant gradient (i.e., it is linear), the hazard is increasing with time. If the survivor function exhibits exponential decay, the hazard is constant over time.

Inverse cumulative distribution function: Also see **probability density function/cumulative distribution function**. The inverse refers to starting from a value on the y-axis on the plot of the cumulative distribution function (a probability) and then reading off the corresponding point on the x-axis.

Kaplan-Meier plot: An estimate of a survivor function, displayed graphically, constructed from time-to-event data observed in a sample of data.

Linear pooling: A technique for combining probability distributions elicited from individual experts into a single probability distribution. A linear pool is calculated as a weighted mean of each expert's probability density function. An example is shown below, with two experts and equal weights. Note that computing a linear pool will not produce a 'standard' probability distribution, as can be seen in the plot; for example, a linear pool of two normal distributions is not another normal distribution.

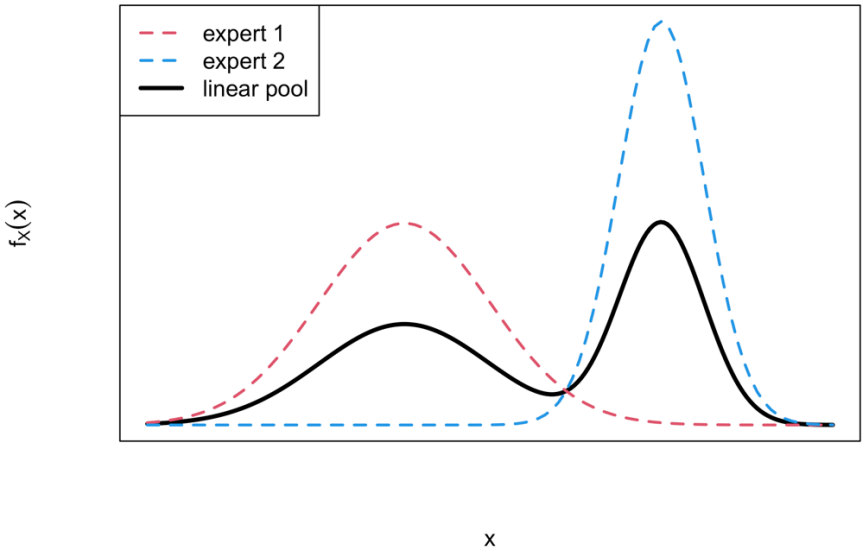

Figure I: An equal-weighted linear pool computed from two normal distributions. Note that this linear pool is not another normal distribution.

Probability density function/cumulative distribution function: Two equivalent ways of defining and visualising a probability distribution. For example, if an uncertain quantity X has a normal distribution function, then its probability density function is the

familiar bell-shaped curve, and the probability of X lying between two values is the corresponding area under the curve. The cumulative distribution function (distribution function for short) shows the probability that X will be less than or equal to value x, for all possible x. The density function gives a clearer graphical impression of what values of X are likely/plausible, but numerical probabilities are easier to read off from a plot of the distribution function.

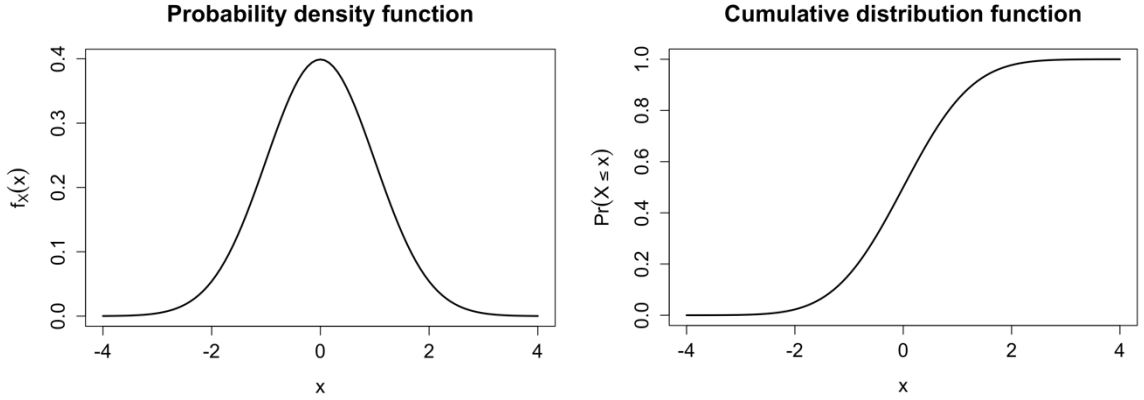

Figure II: The probability density function and corresponding cumulative distribution function for the same distribution (a standard normal distribution).

Prior and posterior distribution: In the Bayesian approach to statistical inference, subjective probability distributions are used to represent uncertain quantities of interest. Given some data relevant to an uncertain quantity, the prior distribution refers to an individual's uncertainty about that quantity *without* knowledge of the data. Bayes' theorem states how this prior distribution should be updated to incorporate the extra information from the data, and this updated distribution is referred to as the posterior distribution.

Quantile/percentile: Given a probability distribution for an uncertain quantity X, the α quantile of that distribution, where α is between 0 and 1, is the value denoted by x_α for which the probability that $\Pr(X \leq x_\alpha) = \alpha$. For example, if X has a $beta(4,8)$ distribution, the 0.9 quantile is 0.51: the probability that X is less than or equal to 0.51 is 0.9. This is illustrated below: in the density function plot, the shaded area is 90% of the total area under the curve; in the distribution function plot, the 0.9 quantile is read off directly. Percentiles are the same as quantiles, but are referred to as percentages rather than decimals: the 0.9 quantile would be referred to as the 90th percentile.

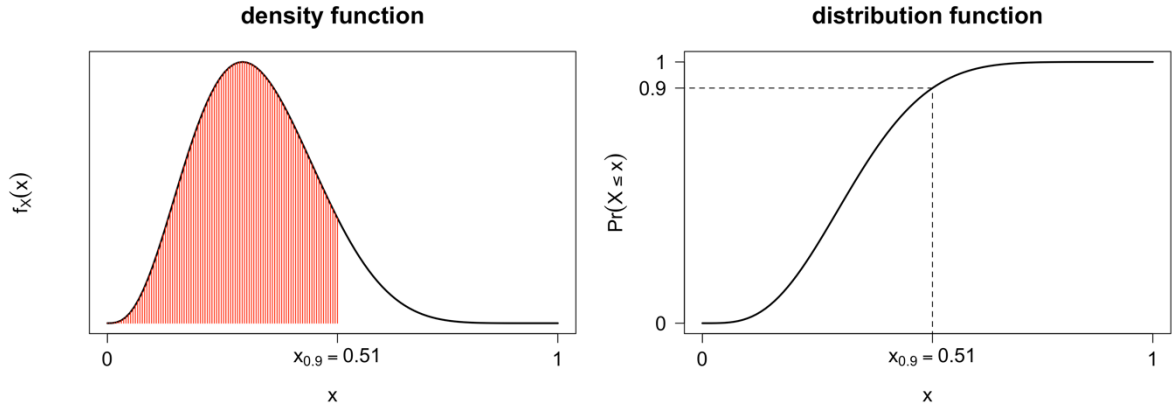

Figure III: Visualising a quantile. On a density function plot, the shaded area to the left of the 0.9 quantile is 90% of the total area under the curve. On a distribution function plot, we read off the 0.9 quantile directly from the curve.

Quartiles: The 0.25, 0.5 and 0.75 quantiles, also referred to as the lower quartile, median and upper quartile respectively.

Subjective probability distribution: A description of an individual's uncertainty about some unknown quantity, presented in the form of a probability distribution. The word "subjective" is used because for the same unknown quantity, different probability distributions would be used to describe the uncertainty of different individuals, if the individuals had different knowledge and opinions about that quantity.

Survivor function, survival curve, $S(t)$, $S(.)$: For time-to-event outcomes for a population, we use $S(t)$ to represent the proportion of the population for which the event occurs at time t or later. The survivor *function* is represented by $S(.)$: this function gives the value of $S(t)$ for all possible values of t. An example is shown below. If the event being observed is death, then we would read off that, starting with a population of patients at time 0, after 4 years, 20% of the population would still be alive. Informally, a survivor function is sometimes referred to as a "survival curve".

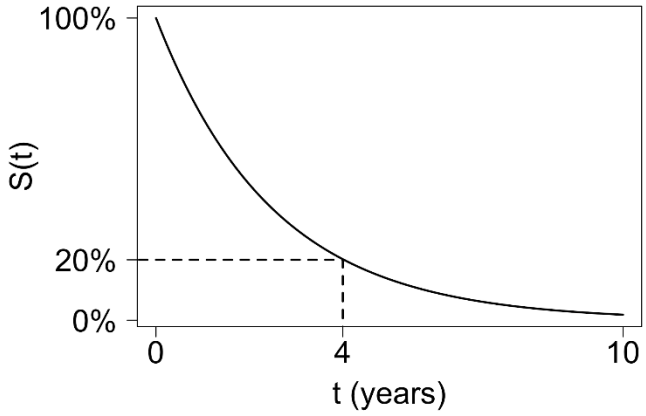

Figure IV: An example of a survivor function. If this was describing how many years patients survived after treatment, we would read off that 20% of the patients would still be alive after 4 years.

Tertile: The tertiles are the $\frac{1}{3}$ and $\frac{2}{3}$ quantiles, referred to as the lower and upper tertile respectively.

Variable interval method: An elicitation method in which an expert is asked to provide one or more quantiles, e.g., "What is your 0.5 quantile value (median) for the uncertain quantity of interest?" The term "interval" is used because intervals for the uncertain quantity can be constructed from the quantiles provided by the expert. For example, if an expert reports a value of 10 for their 0.5 quantile, and a value of 20 for their 0.75 quantile, we interpret this to mean that the expert's probability the uncertain quantity will lie in the interval [10, 20] is 0.25. The intervals are "variable" in the sense that different intervals can result from asking different experts the same questions.

1 INTRODUCTION

1.1 BACKGROUND

Healthcare decision-making often encounters significant challenges due to uncertainty, particularly when empirical evidence is sparse or incomplete. This is particularly evident in areas such as rare diseases, advanced therapy products, and precision medicine, where robust data from randomised controlled trials (RCTs) or observational studies may be unavailable. In such situations, decision-makers must rely on alternative sources of evidence to fill knowledge gaps; this typically involves using expert knowledge.

In NICE health technology evaluations: the manual (which we refer to from now on as 'the NICE Manual'), the distinction is made between expert opinion and "structured" expert elicitation, both of which play significant roles in healthcare evaluations.[1] Expert opinion typically involves qualitative or quantitative insights from clinical or patient experts, often used to supplement, validate, or interpret empirical data from RCTs or observational studies. For example, expert opinion may inform understanding of a technology's design, its application in clinical practice, or the context of its use. This kind of input can be particularly valuable in the evaluation of medical devices, diagnostics, or other interventions where operational, organisational, or experiential factors might influence outcomes.

Aspinall and Cooke (2013) use the term "structured expert elicitation" to mean that experts are asked questions such that their answers have clear operational meaning, and the process of arriving at a defined position is traceable and open to review.[2] In healthcare decision-making "structured" expert elicitation is taken to mean that a justified protocol is followed which lays out in advance the questions that the experts will be asked, and that the output of the process is a probability distribution to represent the uncertainty of the experts. Protocols for structured expert elicitation include Cooke's classical method;[3] ("modified") Delphi methods;[4] the Investigate, Discuss, Estimate, Aggregate (IDEA) protocol;[5] the Medical Research Council (MRC) reference protocol;[6] and the Sheffield Elicitation Framework (SHELF).[7] These protocols are described briefly in Appendix A.1.

The NICE Manual emphasises that structured elicitation methods are preferred because they minimise bias and (with appropriate questioning) provide some indication of the experts' uncertainty.[1] Careful quantification of uncertainty is important: without this, expert-elicited values may be perceived as mere 'guesstimates' which undermines confidence in using them.

1.2 MOTIVATION

Long-term survival estimates play a critical role in determining the cost-effectiveness of many new medicines, yet mature evidence is rarely available at the time of health technology assessment (HTA). NICE typically evaluates the clinical and cost-effectiveness of a medicine shortly before or just after marketing authorisation. For cancer medicines, this timing frequently means that mature evidence on overall survival is unavailable and there are often high levels of censoring. This can lead to significant uncertainty in the evidence base.

Estimating long-term survival in the absence of data is a persistent challenge in economic modelling for HTA and a common source of uncertainty within decision making, as estimates of long-term survival can heavily influence the resulting incremental cost-effectiveness ratios (ICERs). As per the NICE Manual, obtaining clinical experts' judgments has become standard practice for assisting in survival extrapolation, though this is typically *only* for validation of a statistical model, rather than in the conduct of structured elicitation exercises.

In the context of survival extrapolation, we would classify any method as "structured" that involves following a clearly specified elicitation protocol with justification for the methods chosen and that produces an assessment of the uncertainty associated with the extrapolated survival quantities elicited from the experts in the form of a probability distribution. Examples of 'unstructured' methods would be (solely) asking experts to comment on the plausibility of different extrapolated survival curves or only obtaining point estimates from experts for quantities of interest, e.g., the percentage of patients alive at ten years after treatment.

Time-to-event data differ substantially from other types of data due to the presence of censoring and the one-to-one relationship between the survivor function and the hazard function. This relationship introduces specific complexities in survival modelling. Two previous Technical Support Documents (TSDs) have focused on parametric survival modelling, providing guidance on the advantages, disadvantages and limitations associated with different models, and offering advice on how to select appropriate models on a case-by-case basis.[8,9]

The NICE Decision Support Unit (DSU) TSD 14: *Survival Analysis for Economic Evaluations alongside Clinical Trials - Extrapolation with Patient-Level Data* discusses how models fitted to trial data can appear almost indistinguishable within the trial period, but different models make different assumptions about hazard trends and extrapolations beyond trial follow-up periods can differ substantially as a consequence.[8] The NICE DSU TSD 21: *Flexible methods for survival analysis*, discussion is focussed on the shapes of hazards and it is recommended that these should always be considered when choosing which models to use.[9]

Both TSDs have advised that it is important to take external validity into account when selecting models. When survival models are fitted to data from a clinical trial with limited follow-up or high levels of censoring, it is important to consider the plausibility of the extrapolations associated with these models based on information outside of that provided by the trial. This information could be from other clinical trials, registry data, or clinical expert opinion. However, these previous TSDs have not provided guidance on how external information should be elicited to inform survival extrapolations.

The elicitation protocols referred to in Section 1.1 can all be used to obtain appropriate expert judgements for survival extrapolation; they are all applicable to generic uncertain quantities. However, just as there are specific complexities in modelling survival data, there are also specific complexities in elicitation for survival extrapolation. These arise from the nature of the available data: the observed survival data from which to extrapolate, and the relationship between survivor functions and hazard functions, which can be exploited in the elicitation methodology. Consequently, the general approach recommended in this TSD is to start with an established protocol

(such as one of those referred to in Section 1.1) as the basis for the elicitation but then make modifications that tailor the protocol to the extrapolation task.

1.3 SCOPE OF THIS REPORT

The scope of this TSD is limited to eliciting long-term survival outcomes using a structured approach. This document is written as a companion to TSD 14[8] and TSD 21[9]. The object of interest is the same here: to obtain an extrapolated survival curve (more formally, the survivor function).

To establish notation and definitions, we define $S(t)$ to be the proportion† of patients in the population who survive for at least time t. We suppose that $S(t)$ can be estimated for $t \leq t_0$ using available individual patient-level data. The time point t_0 may be the last time point in the data (either an event or an observation of censoring), but more generally, it would be the last time point at which we are willing to use the available data to report an estimate (with uncertainty) of $S(t)$. For economic modelling purposes, we also need to know $S(t)$ for $t_0 < t \leq T_H$ for where T_H is the time horizon in the economic model. We refer to $S(t)$ for all times t in the interval $[0, T_H]$ as the survivor function.

We suppose that model-based approaches to extrapolation have been attempted (as recommended in TSD 14[8] and TSD 21[9]), but that there remains significant uncertainty about the extrapolated survivor function (implying significant uncertainty about cost-effectiveness), such that it has been established that expert judgement will be necessary. It is assumed that this is in the context of an HTA. An alternative context could be an investment decision about future evidence collection, aimed at assessing the value of reducing uncertainty about a survivor function.

† Survivor functions are typically defined in terms of "probabilities" rather than "proportions". In this document, we reserve the term "probability" to mean an opinion of an expert; an expert may make probability judgements about the true value of a proportion.

In Figure 1 we give a hypothetical example of the data that are typically available: individual patient survival times, summarised as Kaplan-Meier plots. A 'non-expert' could probably provide a plausible-looking extrapolation, albeit with fairly substantial uncertainty, based solely on general familiarity with Kaplan-Meier plots and knowledge of the age of the patients. Therefore, it is important to consider, what knowledge would an expert have to distinguish them from the non-expert?

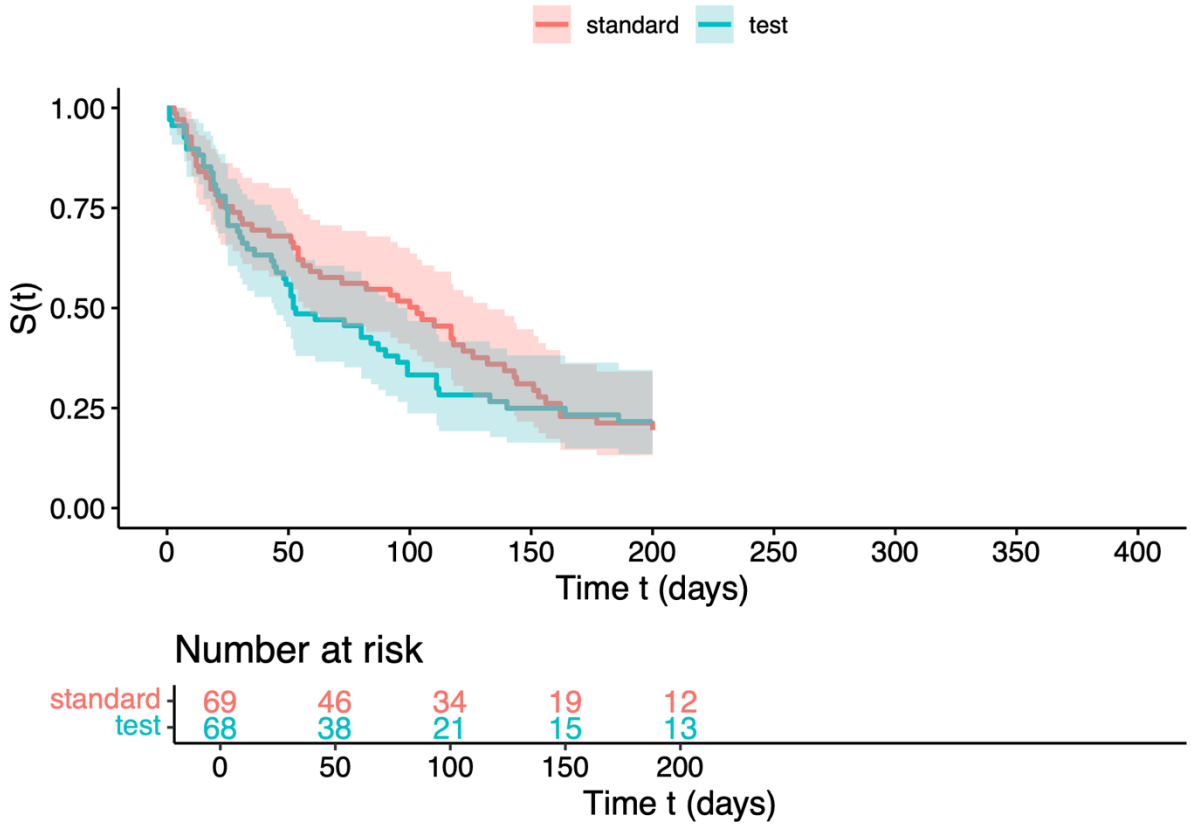

Figure 1: An example extrapolation task. We suppose data are available for the first 200 days, but we wish to extrapolate the survivor functions over the lifetimes of the patients. Data source: veterans dataset in the R package survival.[10]

We can distinguish between two general sources of expert knowledge. The following is not intended to be an exhaustive or unique categorisation of expert knowledge but is useful for considering elicitation methodology.

1. Understanding of long-term behaviours of the disease, patient population and mechanism of treatment action, covering the period of interest. These factors can all cause changes in the trend in the survivor function (more formally, the hazard),

beyond the observation period. Information about future changes in trend/hazard is not necessarily evident in the Kaplan-Meier plot and may not be reflected in statistical models fitted to the observed data only.

2. Experience of treating and observing patients over the period of interest. This is likely to apply to the control arm only, however, experience might be limited if the decision problem is restricted to a particular subgroup of the patient population.

The first source of knowledge is more particular to the survival extrapolation problem, as it relates to how relevant factors change over time. We therefore recommend incorporating additional steps within standard elicitation protocols to target this source of knowledge. The second source of knowledge is more generic to other elicitation problems in healthcare and can be targeted through the usual practice of collating and sharing evidence between experts.

We must also consider what is reasonable to expect from expert knowledge. In general, we do *not* expect that eliciting expert knowledge will necessarily produce a definitive survivor function, resolving the matter of which survivor function to use to populate an economic model. Experts are likely to have appreciable uncertainty about extrapolated survival proportions, perhaps only providing a clear steer regarding the order of magnitude (e.g., a 10% survival proportion versus a 1% survival proportion). As noted above, the experts may also be able to provide qualitative assessments relating to changes in hazard.

We therefore propose that an expert elicitation methodology for long-term survival should provide at least two outputs:

1. quantification of uncertainty about the survivor function at one time point, beyond any observed individual patient data, provided in the form of a subjective probability distribution;
2. qualitative assessments regarding plausible behaviour of the hazard function over the extrapolation period.

Note that point 1 states a minimum requirement regarding a single time point; we discuss the possibility of eliciting distributions at multiple time points in Section 3.5.3.

1.4 AIM AND OBJECTIVES OF THIS REPORT

The aim of this TSD is to show how, in HTAs involving survival extrapolation, we can incorporate expert knowledge and uncertainty formally through a structured expert elicitation process, improve upon the more common practice of consulting experts to validate survival models, and hence improve the way in which HTAs support decision-making. This TSD discusses the background theory and proposes a set of recommendations for using structured expert elicitation for survival extrapolation. An example elicitation protocol and supporting software are provided.

1.5 STRUCTURE OF THIS REPORT

In Section 2 we discuss some fundamental concepts for structured expert elicitation. Elicitation methodology is typically not application-specific; the content in Section 2 is all relevant to the application of survival extrapolation. In Section 3, we discuss aspects of structured expert elicitation that are more particular to specific problems and issues in survival extrapolation.

Section 4 provides an example protocol, bespoke for survival extrapolation, based on the task of eliciting a distribution for the proportion of a patient population who would still be alive at some specified time point. We describe each stage of the process in detail and provide free software for implementation, available at https://shelf.sites.sheffield.ac.uk/. This example is based on modification of the SHELF protocol. We also discuss modification of other standard protocols.

Section 5 reviews the use of expert elicitation for long-term survival in the broader literature, in addition to NICE oncology technical appraisal submissions. It serves to highlight potential deficiencies in methods of obtaining and reporting expert opinion/judgements and highlights areas for improvements which are addressed further within the recommendations. Section 6 provides recommendations, discussion and suggestions for further research.

2 ELICITING PROBABILITY DISTRIBUTIONS: GENERAL THEORY AND METHODS

The key methodological task is to obtain or "elicit" a probability distribution to represent expert uncertainty about an unknown quantity of interest. In this section we introduce and discuss various fundamental concepts for this methodology, but we do not attempt to give a comprehensive survey of the field. Various review textbooks and articles include Garthwaite et al. (2005),[11] O'Hagan et al. (2006),[12] Oakley (2010),[13] European Food Safety Authority (2014),[14] Morgan (2014),[15] Dias et al. (2018),[16] Bojke et al. (2021),[6] and Soares et al. (2024).[17]

This section outlines fundamental concepts related to probability, uncertainty, expert judgement, and elicitation. It highlights the distinction between two key interpretations of probability and their relationship to different types of uncertainty, discusses heuristics and biases in elicitation, and summarises the elicitation process.

2.1 FREQUENCY AND SUBJECTIVE PROBABILITY

Frequency probability represents an objective measure of probability based on the observed long-term relative frequency of an event occurring under repeated, identical conditions.[12] For example, the frequency probability of obtaining heads in a fair coin flip is defined as the proportion of times that head would occur if the coin were flipped an infinite number of times. Frequency probabilities can be estimated using empirical data and statistical reasoning, and are thought of as having unique, true values. Frequency probability is only applied in situations where an 'experiment' can be repeated under the same conditions, with the possibility of observing different outcomes in each case.

Subjective probability represents an individual's degree of belief that an event will occur or that a proposition is true.[12] For example, an individual may state, "I am 60% certain the incumbent government will be defeated at the next election." This value depends on personal judgement, experience, intuition, and potentially incomplete information. There is no unique, true value of a subjective probability; subjective probabilities may vary between individuals due to differences in knowledge,

perspective, or biases. Subjective probability can be applied to any uncertain quantity or outcome, but is typically used for one-off, non-repeatable events, e.g., whether the median survival time for the population of all lung cancer patients treated with a particular drug will exceed one year.

2.2 ALEATORY AND EPISTEMIC UNCERTAINTY

The distinction between frequency and subjective probabilities is closely linked to the distinction between aleatory and epistemic uncertainty.[12] Aleatory uncertainty refers to the inherent randomness or variability in a system or process. It is associated with natural variability that cannot be reduced, even with improved knowledge or more information, e.g., the outcome of tossing a coin. Aleatory uncertainty can only apply to events in the future, where the outcome has not yet been determined. Frequency probability applies exclusively to aleatory uncertainty, as it captures the long-term statistical behaviour of such inherently variable phenomena.

By contrast, epistemic uncertainty refers to uncertainty due to a lack of knowledge or information.[12] This type of uncertainty arises because we do not have complete understanding of the processes or parameters involved. If we had better information or more detailed data, epistemic uncertainty could be reduced or eliminated. Subjective probability can encompass both aleatory and epistemic uncertainty.

If we consider uncertainty about the survival time of a single patient in the future, randomly selected from a population, we would identify aleatory uncertainty in this context. Uncertainty about this survival time could still be appreciable, regardless of the data we obtain now; individual patient survival times may vary in ways that we cannot predict. Given sufficient data from the population, there may be acceptance of an (approximately) ‘correct’ probability distribution for this aleatory uncertainty, which could be verified against further repeated random sampling and observation of patients.

If we consider uncertainty about the proportion of *all* patients in a population who will survive for say, more than ten years, we would characterise this uncertainty as epistemic. The population proportion is a single number; collecting data may reveal,

to an arbitrary precision, what this number is, but cannot reveal a 'correct' probability *distribution* for this number.

2.3 HEURISTICS AND BIASES IN ELICITATION

There has been extensive study of the performance of individuals in making probability judgements. Perhaps the best-known research is the heuristics and biases programme pioneered by Kahneman and Tversky in the 1970s, summarised in Kahneman (2011).[18] Various heuristics (strategies or rules, typically quick to implement) were proposed for how individuals make probability judgements, and how the use of such heuristics might lead to biases, as demonstrated in a series of experiments. Two heuristics that may be relevant in the context of survival extrapolation are "availability" and "anchoring and adjustment".

The availability heuristic involves formulating a judgment about a probability based on the ease in which a particular outcome can be recalled. For example, when making judgements about the proportion of survivors at 2 years (i.e., $S(t = 2 \text{ years})$), an expert may attempt to recall instances where they have observed patients surviving for longer than 2 years. Cognitive bias may therefore occur from this heuristic if some outcomes are more memorable than others, e.g., if some observed patient outcomes were particularly distinctive.

The anchoring and adjustment heuristic involves starting with some initial value, the "anchor", and then adjusting from that to make an estimate. This can apply in any context of eliciting a distribution, for example, if an expert first provides a 'best estimate' of the unknown quantity of interest, and then "adjusts" from this anchor to give lower and upper limits of their (e.g., 95%) probability interval for the quantity. In the survival extrapolation context, anchoring and adjustment could occur if an expert is first shown an extrapolated survivor function from some parametric model, and then asked to provide their own estimate of the survivor function. Cognitive bias may occur from this heuristic if the expert gives undue weight to the anchor and does not adjust far enough.

The consequence of these potential biases arising from the use of these heuristics is typically 'overconfidence'. For example, if an expert is judging 95% probability intervals for a series of quantities, the proportion of intervals that contain the true values is (considerably) less than 95%: the expert is too confident in their knowledge and makes judgements with intervals that are too narrow. Overconfidence has been observed in expert judgement studies (Wilson (2017)).[19] See also the review in Chapter 4 of O'Hagan et al. (2006).[12]

For completeness, a third heuristic that was proposed and investigated in the heuristics and biases programme is "representativeness", which involves judging a conditional probability $Pr(A|B)$ by making assessments of how representative event A is of event B. It is less clear to us how this heuristic might apply in the survival extrapolation context; whilst elicitation methods may involve assessments of conditional probability, it is not clear how an expert may use notions of "representativeness" in this context to come to an assessment.

There has been extensive critique of the heuristics and biases programme, summarised in Kynn (2008),[20] who describes a "heuristics and biases *bias*" (our emphasis added). Kynn observes that the 1970s publications of Kahneman and Tversky have substantially more citations than their later publications, where they "soften their stance on inherent human bias". There is a considerable body of research, in some cases where different experimental results are found, that has not received the same attention in the statistical elicitation literature. One example is the development and testing of alternative cognitive models for probability judgement in Gigerenzer et al. (1991).[21]

In the survival extrapolation case, for overconfidence to occur, this would typically be due to some factor causing the survivor function to take a sharp change of gradient at some time in the future, in a manner unanticipated by any expert. Note that there is a possibility that experts may tend towards *under*confident judgements, or at least make judgements that are not very informative. Given the nature of the available data and that survivor functions are bounded between 0 and 1, it is not difficult to provide an interval for any $S(t)$ that would almost certainly contain the true value.

In any case, structured expert elicitation methods designed to mitigate potential biases would typically result in good practice. For example, whether or not experts attempt to use the availability heuristic, it is good practice to collate and share all available evidence between the experts, make the evidence easily accessible, and reduce the need for experts to rely on memory as much as possible.

Bojke et al. (2021) also highlight that motivational biases can arise from an expert's personal incentives or social dynamics within the elicitation process.[6] For example, experts with vested interests in the outcome of the elicitation may consciously or unconsciously skew their judgements to align with their goals. Motivational biases could influence experts' input in this context through financial conflicts of interest such as affiliations to the company for previous work, or non-financial researcher allegiances such as personal or collaborative relationships, study authorship or enthusiasm for the intervention.

2.4 ELICITING A DISTRIBUTION: THE MATHEMATICAL PROCESS

One technical issue is how to obtain a full probability distribution for an uncertain quantity, given the sorts of judgements an expert is typically able to make. For this discussion we consider a single expert only. For illustration we suppose the uncertain quantity is the survivor function at a single time point: $S(t)$. We discuss the choice of uncertain quantity for elicitation in Section 3.5, but it is helpful at this point to understand the detail of what eliciting a distribution involves.

Typically, we do *not* ask experts to propose probability distributions directly, i.e., to make statements such as, "My uncertainty about $S(t)$ can be represented by a beta distribution with parameters 2 and 10." In general, we expect experts to provide a small number of probability or quantile judgements only. For a probability judgement, the expert would provide the value of $Pr(S(t) \leq x)$‡ given some specified value of x, and

‡ $Pr\ (S(t) \leq x)$ represents an individual's probability that the uncertain quantity, $S(t)$, is less than or equal to some specified value x.

for a quantile judgement the expert would provide the value of x such that $Pr(S(t) \leq x) = p$, given some specified value of p.

The elicited probability judgements are listed in the form

$$Pr(S(t) \leq x_i) = p_i,$$

for $i = 1, \ldots, n$, where n is the number of judgements obtained from the experts relating to that quantity of interest. One possibility is to choose lower and upper limits for $S(t)$, and then choose a piecewise uniform distribution that interpolates between the specified $Pr(S(t) \leq x_i) = p_i$ judgements with straight lines. If the roulette method (discussed in Section 2.4.1) has been used, this uses the 'histogram-type' shape displayed to the expert as the expert's probability density function for $S(t)$.

If the piecewise uniform distribution does not appropriately represent the expert's uncertainty, an alternative is to select a parametric family of distributions with parameters θ, and then choose θ by minimising a sum of squares $R(\theta)$ defined as

$$R(\theta) = \sum_{i=1}^{n} \big(p_i - F(x_i; \theta)\big)^2, \quad (1)$$

where F is the cumulative distribution function from the chosen family of distributions. For example, if fitting a beta distribution to the expert's judgements, θ would be the two parameters of the beta distribution, and $F(x_i; \theta)$ would be the value of $Pr(S(t) \leq x_i)$ according to a beta distribution with these parameters; we are trying to find θ that makes this probability close to the probability given by the expert. We denote the value of θ that minimises $R(\theta)$ by $\hat{\theta}$. In most cases, there will not be a formula that gives the value of $\hat{\theta}$: we must find it using numerical optimisation methods. Code for doing this is available in the SHELF R package.[22]

Once $\hat{\theta}$ is obtained, the usual practice, as discussed in Section 5.4.3 in O'Hagan et al. (2006),[12] is to provide some feedback in the form of some quantiles from the fitted distribution, e.g., the 0.05 and 0.95 quantiles (i.e., $F^{-1}\big(0.05; \hat{\theta}\big)$ and $F^{-1}\big(0.95; \hat{\theta}\big)$, where $F^{-1}\big(.; \hat{\theta}\big)$ is the inverse cumulative distribution function from the parametric family of distributions with parameter values $\hat{\theta}$). Given this feedback, the expert may

accept the fitted distribution as an acceptable representation of their beliefs, or they may propose modifications until an acceptable distribution can be found.

2.4.1 Choices of judgement

As discussed above, experts might be asked to make probability or quantile judgements. These are generally referred to as fixed interval and variable interval methods respectively. In the fixed interval method, an interval is specified, and an expert is asked to provide their probability of the quantity lying in that interval. In the variable interval method, a probability is specified, and the expert is asked to provide an interval such that they think the uncertain quantity will lie in that interval with the specified probability; the size of the interval will vary depending on the opinions of the expert. Common variable interval methods involve asking the expert to make quartile or tertile judgements. We provide a detailed description of the quartile method in Section 4.1. Note that standard elicitation protocols involve training of the experts, which would include training in making the required judgements.

To the best of our knowledge, there is no conclusive evidence that favours one method over the other, but individual experts may find some probability judgement tasks easier than others. However, if the intention is to fit a probability distribution using Equation (1) above, then caution is needed with fixed interval methods. This is because fixed interval methods do not guarantee suitable probability judgements for distribution fitting with this approach. For example, if using the SHELF R package for distribution fitting, for distributions we may wish to use for survival extrapolation, there is a requirement for at least one elicited probability greater than 0, but less than 0.4, and at least one probability less than 1, but greater than 0.6. This is to ensure there is some information regarding both tails of an appropriate distribution, and to enable robustness of the numerical minimisation required. This requirement can only be guaranteed using variable interval methods, e.g., by eliciting lower and upper quartiles.

There is a related issue with one particular fixed interval method: the roulette method. In the roulette method, an expert is asked to allocate a number of chips to bins, with the proportion of chips allocated to a particular bin representing the expert's probability of $S(t)$, the quantity of interest, lying in that bin. This is equivalent to the expert

providing a set of judgements $P\left(x_j < S(t) \leq x_{j+1}\right)$, where $x_1, \ldots, x_n$ correspond to the endpoints of the bins. This method can be appealing because of its apparent simplicity, and because of the immediate graphical feedback that the expert gets as they allocate the chips to the bins.

However, there are difficulties in implementing the roulette method that are not immediately apparent. The roulette method can only produce useful results if the bins have been chosen appropriately. In an extreme case, an inappropriate choice would result in an expert allocating all the chips to a single bin. In some situations, the context may determine the bins of interest. In other situations, which may include survival extrapolation, it may be necessary to know something about the probability judgements the expert will make *before* specifying the bins. This is to ensure the bins have the appropriate location and granularity for the expert's probability distribution. Whilst the expert can be asked to choose the bins, it is not guaranteed that the expert will make a choice that results in probability judgements suitable for distribution fitting.

In summary, our recommendation is to use variable interval methods, to ensure that distribution fitting is viable. Fixed interval methods can still be used, but it may be necessary to have some prior knowledge of what an expert is likely to think, such that appropriate fixed intervals can be given to the expert. Note that in the SHELF protocol, variable interval methods are used initially, and then fixed interval methods can be used at a later stage in the protocol; suitable fixed intervals are identified using the responses to the variable interval method questions.

2.4.2 Eliciting distributions for proportions

The process described in the previous section can be applied to any uncertain scalar quantity. There are also methods designed specifically for eliciting a distribution about an uncertain proportion (see Section 6.3 in O'Hagan et al. (2006)[12]). As the quantity $S(t)$ is a proportion: the proportion of a population surviving until at least time t, we discuss one such method here.

In the "equivalent prior sample" (EPS) method proposed in Winkler (1967),[23] the expert is asked to provide an estimate of the proportion, and to estimate a sample size

'equivalent' to their knowledge. For example, an expert estimates a proportion ($S(t)$) to be 0.1 and reports that, in some sense, their knowledge is equivalent to one patient surviving for at least time t out of a sample of ten patients. The point estimate and the equivalent sample size can then be used to select the parameters of a beta distribution to represent the expert's judgements.

The rationale for this is as follows. Suppose we have an observation x of a binomial random variable X, with n trials and an unknown probability φ of 'success' on each trial. Given an improper prior distribution $beta(0,0)$ for φ, the posterior distribution of φ would be $beta(x, n-x)$. Hence, working backwards, a judgement of a $beta(a,b)$ distribution for φ can interpreted as a posterior distribution following an observation of a successes out of $a+b$ trials (for integer a and b), and no other information.

Rather than eliciting the EPS directly, we can elicit a distribution using the general procedure described above, choosing a beta distribution as the parametric family of distributions, and then use the parameters of the fitted beta distribution to infer an EPS. This can have *some* value in providing feedback to the experts and validating the elicited distribution, though care is needed with the interpretation. For example, suppose an expert judges a median value of 0.05 for $S(t)$, and claims to be 99% certain that $S(t)$ lies between 0.03 and 0.07. Fitting a beta distribution to these judgements using Equation (1) results in approximately a $beta(48,900)$ distribution. This would imply an EPS of 948 patients, with 48 survivors at time t. Depending on the context, this could suggest overconfidence if the actual number of patients observed by the expert is considerably fewer. The cause is the narrow 99% probability interval specified around the expert's median value. Note that we are not obliged to use a beta distribution; a different family of distributions could be used, but the concern of possible overconfidence would be the same. The use of a beta distribution and EPS simply gives a way to articulate the suspected overconfidence.

However, there is no simple correspondence between expert knowledge and an equivalent sample size. We have discussed in Section 1.3 what knowledge an expert might draw on when extrapolating survival outcomes; we expect there to be knowledge that is not easily equated to equivalent sample data. Whilst experts may draw on

experiences of treating patients, there may be patient and treatment characteristics that make the extrapolation task distinct from such experience. The equivalent prior sample from a fitted beta distribution could be reported as part of the feedback process, but experts should be advised to interpret it with caution.

2.5 MULTIPLE EXPERTS AND ELICITATION PROTOCOLS

It is always desirable to elicit the judgements of more than one expert, but it is usually necessary to report a single conclusion of the exercise, in the form a single probability distribution for the uncertain quantity of interest. The process of obtaining a single distribution from multiple experts is referred to as aggregation of opinions, and this is typically performed either mathematically or behaviourally.

Mathematical aggregation involves fitting a distribution to each individual expert's judgements, and then computing a pointwise average (typically either an arithmetic mean: the "linear pool", or geometric mean: the "logarithmic pool") of the distributions. Experts may be given different weights when computing averages. Different aggregation methods are discussed in Chapter 10 in O'Hagan et al. (2006),[12] and software for implementing linear pooling is included in the SHELF R package.[22]

Note that there is a technical difficulty if using linear pooling and attempting to use Bayesian methods to update an expert elicited distribution with data. When updating the data, there is a choice to either update the pooled distribution or update the individual distributions first and then pool.[24] However, these two choices can give different results, with no obvious justification for one over the other. Though Bayesian updating is a possibility, as we discuss in Section 3.6, the methods recommended in this TSD do not involve Bayesian updating.

Behavioural aggregation involves discussion between the experts, with the experts asked to agree on a single set of probability judgements. This does not imply the experts are asked to come to a consensus regarding what the uncertain quantity is likely to be, rather, they may be asked to agree on a single set of judgements that appropriately represent the diversity of opinions with the group. A distribution is then fitted to this single set of probability judgements.

Closely related to the choice of aggregation method is the elicitation protocol, which specifies how the elicitation exercise is carried out, including the aggregation of expert's judgements. A summary of some commonly used elicitation protocols (involving different aggregation methods) is given in Appendix A.1 (Cooke's classical method; Delphi; IDEA; MRC reference protocol; SHELF).

We recommend using any one of the protocols above, but with additional steps incorporated to address issues specific to survival extrapolation. We illustrate this with a modified SHELF protocol in Section 4; we also discuss modification of the other protocols. We do not make a recommendation regarding which of the above protocols to adapt, but we will comment further on issues regarding elicitation conducted via surveys versus face-to-face elicitation (see Section 2.8).

2.6 ROLES IN AN EXPERT ELICITATION EXERCISE

There will normally be a team of individuals involved in the design and conduct of an expert elicitation exercise. Protocols involving group interaction require a "facilitator" to chair the discussion. Here, we use the term facilitator in a more general sense, to mean the individual with overall responsibilities for

- choosing the uncertain quantities for which distributions will be elicited;
- implementing the elicitation protocol including leading any interaction with experts that involves making a probability judgement;
- proposing full probability distributions to represent an expert's (or group of experts') uncertainty, given the probability judgements provided, with any distribution fitting implemented as needed;

hence the facilitator will need expertise in implementing the chosen protocol.

Group discussions should be documented, and in SHELF, the term "recorder" is used to mean the individual who takes notes and works with the facilitator and any other team members on writing up the report of the elicitation. Where software is being used, for example, for distribution fitting and feedback, the recorder may operate the software. Other team members would work on the preparation of an evidence dossier to assist the experts; this requires skills in systematic literature review.

2.7 RECRUITMENT OF EXPERTS

2.7.1 How many experts to recruit?

When using multiple experts, an obvious question is how many experts to recruit, but this is not an easy question to answer. The MRC protocol recommends a minimum of five experts,[6] but it would be difficult to provide a rigorous justification for the exact number.

To the best of our knowledge, there has not been a study that has directly compared the performance of expert panels of different sizes, using the same elicitation protocols, on the same set of uncertain quantities of interest. Some investigation was reported in Mannes et al. (2014)[25] and Budescu and Chen (2015),[26] but in a different context. These studies showed when aggregating judgements from a "crowd" of forecasters (examples included starting with 90 forecasters) improved performance could be achieved by identifying the 'best' forecasters, e.g., the top five, and excluding the judgements of the others. Note that the forecasters were providing best estimates only for uncertain quantities, rather than full probability distributions.

We should not think of a 'sample of experts' as having the same characteristics as a sample of data. If we wanted to know the value of $S(t)$ for some population, and we drew a random sample of patients from that population and observed their survival times, we would reduce uncertainty about $S(t)$ as the sample size increases, to the point where we could, in effect, consider $S(t)$ to be known. The same is not true with 'sampling experts': we should not suppose the true value of $S(t)$ would be revealed simply by asking a sufficiently large number of experts to estimate it.

We need to consider how increasing the number of experts in a structured expert elicitation exercise would increase the pool of *knowledge*. Two experts may share the same knowledge, though there can still be value in them both participating in the elicitation, as they may interpret it differently, and think different probability judgements based on the same knowledge to be appropriate: an unavoidable aspect of subject probability.

Regarding the minimum number of experts to recruit, two relevant factors are:

1. the number of experts needed to ensure that all the appropriate domains of expertise are represented;
2. the number of experts needed such that stakeholders have confidence in the use of structured expert elicitation.

The first factor is domain specific, and we will not suggest a minimum. For the second factor, we would tentatively suggest a minimum of three. The rationale is that this would be analogous with asking one expert for an opinion, and then asking two other experts to independently assess the opinions of the first. However, using more experts would give more confidence to stakeholders, and we recommend aiming to recruit more than three. Diversity of opinion is also important. For example, a panel of three 'independent' experts may be preferable to a panel of four experts who have all collaborated on the same study.

2.7.2 Identifying and selecting experts

An expert is defined as an individual with significant knowledge in a specific domain and the competence to apply this knowledge practically.[6,12,27] The aim of expert selection is to assemble a panel with a suitable breadth of experience and expertise. Additional consideration should be given to potential conflicts of interest and considering the representativeness of equity characteristics among the invited experts.

Elicitation protocols are designed on the assumption that experts do *not* have training in probability theory and making probability judgements. Therefore, we do not recommend considering this as a factor when recruiting experts; experts should be recruited based on their subject-matter expertise. Elicitation protocols include a component for training the experts; this is an important part of the process to ensure that all participating experts have the same level of understanding of subjective probability and making probability judgements. In the context of survival extrapolation, some experts may have had statistical training regarding the presentation and analysis of survival data, but this is likely to be based on the use of frequency probability for assessing aleatory uncertainty only. For such experts, training will need to emphasise differences when using subjective probability to quantify epistemic uncertainty.

In HTA, consulting experts is common practice, and we generally do not anticipate difficulty in identifying individuals with extensive knowledge of the subject domain. However, we would highlight some common pitfalls in identifying and recruiting experts:

- **Relying solely on experts from pre-existing lists**. A 'pre-existing' list in this context could be a pool of experts that the company collaborates with in the disease area. This approach may narrow the pool of candidates and introduce selection bias.
- **Experts with experience of using the new intervention under appraisal**. Such experts will need to be included to give the expert panel credibility, but risks of motivational bias will need to be managed. Such experts may have enthusiasm about the intervention, and conflicts of interest if affiliated with the company or involvement in the pivotal trial(s) design/conduct. Including experts who are more distant from the new intervention is also desirable.
- **Prioritising only the most senior experts**. Seniority does not necessarily equate to being the most appropriate expert for the elicitation. Consideration should be given to the experts' experience and whether this is in a population representative of the target population.
- **Limited geographical diversity**. Including experts only from the same geographic location (exact definition is context dependent) and/or organisation. Including experts from limited areas or organisations could result in the lack of representation of the full range of expert beliefs.
- **Leaving recruitment to the last minute**. To maximise the likelihood of successfully assembling a panel of experts for an elicitation exercise, advance planning is essential. Based on our experience, around three months of preparation is typically required to identify and recruit experts and to coordinate dates and times that suits all participants.

2.8 FACE-TO-FACE EXPERT INTERACTION VERSUS ELICITATION USING SURVEYS

Related to the choice of elicitation protocol, there are four general options regarding how a structured expert elicitation can be conducted:

1. an in-person facilitated workshop, held at a single location;
2. an online facilitated workshop using videoconferencing;

3. a survey in which each expert has a one-to-one interview with a facilitator;
4. a survey which the experts respond to in their own time, without any interaction with a facilitator.

We refer to options 1 and 2 as both involving 'face-to-face' interaction between experts (whether in-person or online). Options 3 and 4 can involve sharing of opinions between experts if there are multiple rounds of the survey. Note that options 2-4 could all be described as 'remote elicitation'.

Option 4 may seem attractive, for the reason that it may require less financial and time commitments. There are, however, some challenges to be aware of.

- **Expert engagement**. Participation in an elicitation workshop (whether in-person or online) ensures a certain level of commitment to the exercise from the experts, and the time the experts spend on their deliberations can be recorded. In our experience, experts tend to spend a relatively small fraction of the time making probability judgements, compared with the time spent in discussion with each other. In answering a survey, it is more difficult to assess the level of engagement from each expert. It is feasible that, on occasion, an expert might answer survey questions in a rush, without deliberating for the time one would hope for, if dealing with other pressures simultaneously. Morgan (2014) comments that, "It is an open question whether experts working on their own will devote the same degree of serious consideration in responding to an automated elicitation system that they clearly do when responding to a well-developed protocol during a face-to-face interview with attentive and technically knowledgeable interviewers sitting with them in their office."[15] If using a survey, the interview setting of option 3 can mitigate this.
- **Expert understanding**. It is generally appreciated that experts are likely to be unfamiliar with making probability judgements to quantify uncertainty, and training will be needed. Training is typically easier to conduct face-to-face (either in-person or online), rather than through self-directed learning, particularly if the experts need to ask questions. In our experience, even after training, some experts can still find the process difficult and make judgements that they would be quick to change, once they have seen the judgements of other experts.

- **Scrutiny of expert judgements**. To give stakeholders confidence in the use of structured expert elicitation, there needs to be careful scrutiny and challenge of the experts' judgements. This would need to be done by peer experts with the same understanding of probability judgements. This is more difficult to achieve in a survey. Though a survey can be implemented in multiple stages, with sharing of responses between experts, written justifications for probability judgements can be somewhat terse, and there may little or no opportunity for proper debate within a survey framework.

Issues with methods involving interactions between experts include: the practicalities of assembling expert panels; the risk of bias if one or more experts are unduly influential; undue influence of a facilitator if one is used to manage discussions.

Regarding the first issue, we observe that it is already common practice for a pharmaceutical company to assemble expert advisory boards and that these typically meet face-to-face; the time commitment is similar for an expert elicitation exercise and thus unlikely to be any more difficult to organise. Additionally, we have found that online elicitation meetings can work as effectively as in-person meetings, which can make scheduling of elicitation workshops easier. We have also found that splitting the schedule can help: eliciting individual judgements from each expert organised separately, followed by a shorter meeting of all experts for a facilitated discussion.

Measures can be taken to mitigate the risk of bias resulting from a single expert attempting to exert undue influence.

1. It is good practice to first obtain (and document) judgements independently from each expert before any interactions between the experts. This step is included in both the SHELF and IDEA protocols and is recommended in more general contexts in Kahneman (2011).[18] This can also guard against overconfidence in the aggregated output distribution, in that it typically establishes a wider plausible range of values than that proposed by any single expert.
2. The risk of bias from an unduly influential expert can be discussed with the experts as part of their training, just as it is recommended to discuss other biases with the experts, e.g., anchoring effects. Facilitators should anticipate that due to interpersonal relationships between experts (such as perceived professional

standing, reputation or seniority) some individuals can dominate discussion and in some instances the group dynamic can shift toward more dogmatic viewpoints. Cultural differences and implicit biases (such as gender bias) can also prevent some individuals, particularly those from minoritised backgrounds, from feeling comfortable to contribute to the discussion.

3. The facilitator should set expectations at the start of the meeting to promote parity of contributions between experts. Throughout the meeting, the facilitator can help to manage the discussion to ensure all experts are contributing their views. This behaviour would be expected of the chair of any meeting in which individuals are sharing information and deliberating a course of action.
4. Following Step 1, the final distribution selected as the output of the exercise can be compared against the initial individual judgements, and the rationale in selecting this distribution given the starting point of the individual judgements can be checked.

Regarding potential facilitator bias, we first note that any elicitation methodology will involve choices about how the exercise is conducted that will have some effect on the outcome. This includes, but is not limited to, decisions on how to aggregate expert responses, what judgements to elicit, and if questionnaires are used, how they are worded. Some influence is unavoidable. We recommend, however, that experts are first asked to make judgements independently of each other (and without discussion with a facilitator), using variable interval methods such as the quartile or tertile methods. These approaches help avoid prompting the experts with any specific values of the uncertain quantity of interest.

There is a separate issue regarding how to manage structured expert elicitation with a large number of uncertain quantities, and whether surveys are appropriate in this context, though this may not be relevant for survival extrapolation. Guidance of what to elicit is given in Section 3.5, including an option to elicit a single distribution per treatment group. More generally, this is an open question for research; attempting to elicit a large number of probability distributions via a survey could exacerbate the difficulties described previously.

2.9 VALIDATION OF AN EXPERT ELICITATION EXERCISE

Stakeholders will need confidence in the outputs from an expert elicitation, and it is natural to ask questions such as, “How can the expert elicitation be validated?”. Validation is difficult: the true value of an uncertain quantity of interest is a single, fixed number, but the output of an elicitation exercise is a probability distribution representing a group of experts’ uncertainty. There are extreme cases of distributions that we might judge to be indefensible, and hence evidence of an invalid elicitation exercise. For example, a uniform distribution for a particular $S(t)$ between 0 and 1, or a distribution that gives probability one of $S(t)$ taking some (non-zero) single value. However, it is effectively impossible to say what the correct distribution *should* be, given the knowledge of the experts.

We should, nevertheless, do all we can to give confidence in the elicitation *process*. Stakeholders might reasonably expect the following (with evidence provided in the reporting of the elicitation exercise).

1. There has been a thorough attempt to collate all relevant evidence, with the evidence shared and discussed as necessary by the experts.
2. The experts have the appropriate breadth of expertise and experience to interpret and weight the evidence appropriately.
3. Conflicts of interest are minimised as far as possible when involving experts in the elicitation, and any relevant commitments or potential conflicts are stated clearly.
4. The experts have clearly understood the probability judgement tasks required of them, with training provided as necessary.
5. A clear attempt has been made, through appropriate training, to avoid any biases in the experts’ judgements, that might lead to undue overconfidence or underconfidence.
6. The experts have spent appropriate time formulating and reviewing their judgements; they have engaged properly with the task of making probability judgements.
7. There has been adequate scrutiny and challenge of the experts’ judgements.
8. If behavioural aggregation has been used, that all experts have participated fully in the exercise, that no subgroup of experts have exerted undue influence, and

that the experts accept the chosen distribution as an appropriate representation of the knowledge and uncertainty of the group.

9. If mathematical aggregation has been used, that any outlying judgements that may exert undue influence on the result have been investigated and either justified or modified as needed. (If using Cooke's classical method, alternative assurance can be given here via the method of the weight given to each expert in the aggregation).

Regarding point 7, this might be achieved in different ways, depending on the choice of protocol. In Cooke's classical method, there is scrutiny of each expert's ability to make good probability judgements in the general subject area, through their performance on separate "seed" questions.[3] In SHELF, each expert would first make judgements independently of the others, with the results then shared amongst all the experts for discussion and debate.[7]

The results of any elicitation exercise can be shared externally for additional validation. However, this has limitations, as external experts may not have had the same training in making probability assessments as the participating experts and may not give the same consideration to uncertainty. More generally, following the discussion in the previous section, validation will be most difficult if the structured expert elicitation has been conducted via a survey, with no supporting of the experts by a facilitator.

3 STRUCTURED EXPERT ELICITATION FOR SURVIVAL EXTRAPOLATION

We now discuss aspects of structured expert elicitation that are more particular to specific problems and issues in survival extrapolation. This includes some further commentary on themes from Section 2, and additional methodological topics. Where appropriate, we highlight additional processes that we recommend adding to existing protocols for expert elicitation.

3.1 EXPERT RECRUITMENT AND KNOWLEDGE OF THE AVAILABLE DATA

An additional consideration for the recruitment of experts is whether they will have knowledge of the available survival data, from which we wish to extrapolate. There are some technical advantages if the experts have *not* seen the data, for example, if we wish to use Bayesian methods to synthesise the available data and expert judgement (see Section 3.6). However, this approach would have implications for what can be presented to the experts during the elicitation.

Excluding any expert with any awareness of the available survival data (even if in summary form) may restrict the pool of suitable experts too severely, to the extent that the credibility of the elicitation exercise is undermined. Hence, we recommend that knowledge of the data is not used as an exclusion criterion, and that presentation of the data is integrated into the elicitation protocol. If feasible, we would exclude experts who have already viewed statistical model-based extrapolations, as this may act as too strong an anchor, as discussed in Section 2.3.

3.2 TARGET AND TRIAL POPULATIONS

A general consideration in any HTA is the specification of the target population (e.g., Section 2 of the NICE Manual[1]), which may differ from the clinical trial population in the evidence base. For survival extrapolation, we assume that the choice has been made regarding what data will inform survival outcomes in the health economic model. The choice of data will imply a particular population, and we consider the use of expert judgement to extrapolate survival outcomes for the same population. The definition of

the quantity/quantities of interest should clearly reflect this same population. Experts should therefore see a Kaplan-Meier estimate of the survivor function *intended for use in the economic model* and be asked to consider uncertainty about the *same* survivor function at later times.

Issues of adjusting from a trial population to a different target population would be present regardless of whether expert judgement is used to extrapolate survival outcomes. Expert judgement can, however, be used for this purpose, and a framework for this has been set out in Turner et al. (2009),[28] but we do not consider this further here.

3.3 PREPARATION OF THE EVIDENCE DOSSIER

In structured expert elicitation, an evidence dossier, prepared by the elicitation team, is a compilation of information and evidence provided to experts to help inform their judgements. (All five protocols referred to previously recommend this step, but may not use the term “evidence dossier”). Its purpose is to ensure that experts base their assessments on the same evidence, minimising variability caused by differing levels of background information. In this section, we discuss some general principles, and additional requirements for survival extrapolation.

The advantages of using an evidence dossier include: (i) ensuring experts are well-informed before making judgements; (ii) demonstrating rigour and transparency in the elicitation process, thereby enhancing credibility; (iii) helping to mitigate biases arising from variations in experts’ knowledge.

An evidence dossier should organise information in a way that is easily accessible and relevant to the elicitation process. Accessibility is important because the experts will need to refer to the dossier at the point of making their probability judgements.

The evidence should be presented in a neutral manner and not overwhelm the experts with too much detail. However, supporting contextual information may also be appropriate to ensure sufficient detail of the decision problem and quantities of interest. The dossier also should be reviewed by the participating experts ahead of the

elicitation workshop, providing them with an opportunity to raise questions about any missing or misinterpreted evidence.

A systematic literature review (SLR) of clinical evidence, including RCTs and real-world evidence, is typically conducted as part of the HTA process. It is envisaged that most of the information required for the dossier will be covered by the SLR of clinical evidence, thereby negating the need for additional searches or reviews.

An evidence dossier template is included in Appendix B.1. Recommended components are listed below.

1. **Overview**: The purpose of the dossier, background of the problem, target population (as the elicitation may be focussed on a particular subgroup of the wider population), and objectives of the elicitation.
2. **Quantities of interest**: Clear definitions of the uncertain quantities for which elicited probability distributions are required.
3. **Evidence summary**: A summary of the evidence included within the dossier.
4. **Survival data**: The main evidence, usually presented as a Kaplan-Meier plot. Uncertainty in Kaplan-Meier estimates of the survivor function must be reported, e.g., using 95% confidence intervals as well as the numbers of patients at risk. We recommend including trends in the empirical hazard, e.g., presented within discrete time intervals.
5. **Prognostic patient characteristics:** If available, it is helpful to report summary characteristics of patients at the start of the trial, and the same summaries for patients surviving at the end of the trial.
6. **General population mortality data**: This can be constructed from the Office for National Statistics life tables (or other country-specific life tables appropriate for decision problems outside the UK). We suggest tabulating a survivor function for population with age and sex matching the target population. This will be particularly relevant for older trial populations.
7. **Supporting evidence**: Key external studies that provide additional information to help experts formulate their judgements.
8. **Appendices**: Supplementary information, such as study characteristics, trial inclusion/exclusion criteria, time to treatment discontinuation, the use of subsequent treatments and where necessary cross-over handling.

To assist the experts when reviewing the evidence dossier, it may also be useful to include a checklist outlining key areas for review and expert assessment.

3.4 TRAINING

It is standard practice to provide training for the experts, as they typically will not have participated in expert elicitation exercises before and will not be used to making the sorts of probability judgements required. We recommend that the training includes some standard content relevant to generic elicitation exercises, some additional bespoke content for survival extrapolation, and a practice survival extrapolation exercise.

Standard training content should include the following:

- understanding subjective probability as a degree of belief;
- probability density functions as representations of subjective uncertainty about fixed, unknown values;
- awareness of potential biases in making subjective probability judgements resulting from anchoring; availability; overconfidence (see Section 2.3);
- interpretation of specific probability judgements as used in the elicitation protocol, e.g., quartile judgements.

Regarding additional training, we recommend the following. We suggest first reviewing the definition of a survivor function and presenting an example Kaplan-Meier plot that includes estimated point-wise confidence intervals (we would expect such data/plots to be included in the evidence dossier). The experts may be familiar with these concepts, but it is important to emphasise that the Kaplan-Meier curve is an *estimate* of the true survivor function, in particular, there is uncertainty in the survival proportion for the last reported time-point in the study.

The concept of hazards should be explained, and how hazard functions relate to survivor functions. Some examples for discussion are plotted in Figure 2.

- The dashed black line shows a linear decline in the survivor function. The linearity could give an impression of a ‘constant rate’, but it is important that the experts understand that this implies an increasing hazard; the risk of death is

increasing over time. This is easily understood from the plot by noting that, starting with a cohort of 100 patients, we expect 50 survivors out of 100 after 1 year, and 0 survivors out of 50 from the end of year 1 to the end of year 2.

- The solid red line shows exponential decay in the survivor function, corresponding to constant hazard. Starting with a cohort of 100 patients, we expect 50 survivors out of 100 after 1 year, and 25 survivors out of 50 from the end of year 1 to the end of year 2; for patients alive at the start of a year, there is constant probability of 0.5 that they survive until at least the end of that year.
- The dot-dashed blue survivor function gives an example of decreasing hazard. Starting with a cohort of 100 patients, we expect 50 survivors out of 100 after 1 year, and *more* than 25 survivors out of 50 from the end of year 1 to the end of year 2; for patients alive at the start of a year, probability of surviving until at least the end of that year *increases* year on year.

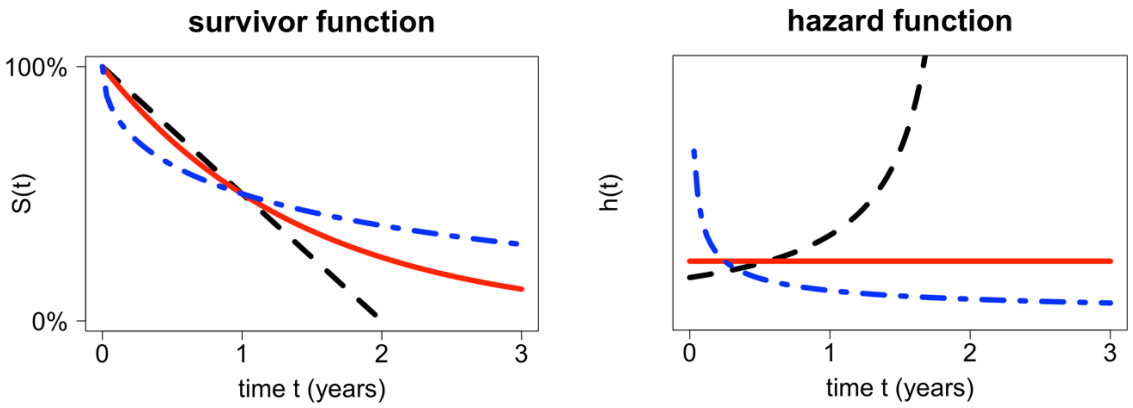

Figure 2: Examples of survivor and hazard functions. These can be discussed as part of the expert training, to help experts interpret trends in the survivor function.

Finally, we suggest discussing the relationship between sample sizes and uncertainty in the survivor function $S(t)$, based on the equivalent prior sample method discussed in Section 2.4.2.

3.4.1 Practice exercise

Once the training presentation is complete, the experts should be given a practice exercise as a part of the training, to give them a dry run through the full elicitation protocol and experience in expressing their uncertainty. Practice exercises sometimes involve a general knowledge question (e.g., a population of a country), assuming the answer is not known by the experts at the time. Here, we recommend using a practice exercise involving survival extrapolation; the data would be censored at a suitable point, so that, after the exercise, the full data can be revealed to the experts. The experts would then be given feedback on how their judgements relate to this revealed data.

3.4.2 Training resources

Editable training slides and a practice exercise for eliciting long-term survival outcomes are available at https://shelf.sites.sheffield.ac.uk/survival-extrapolation. General training materials are also available at

- https://shelf.sites.sheffield.ac.uk/e-learning-course
- https://www.york.ac.uk/che/economic-evaluation/steer/.

3.5 CHOOSING WHAT TO ELICIT

The health economic model may require the full survivor function, for example, if a partitioned survival model has been used. This raises the question of how expert judgement might be used to obtain the full survivor function, which we denote by $S(.)$, as distinct from $S(t)$: the value of the survivor function at a single value of t.

Some possible approaches, which we discuss in turn, are

1. eliciting a distribution for a single $S(t)$ and using it in combination with qualitative judgements about the hazard function to choose a parametric survival model for $S(.)$;
2. eliciting a parametric survival model for $S(.)$;
3. eliciting the survivor function at multiple time points $S(t_1), S(t_2), \ldots, S(t_n)$, and using interpolation.

It would also be possible to use these approaches in combination, but we will not discuss this further. We will briefly discuss eliciting hazard ratios and treatment effects in Section 3.8.

3.5.1 Eliciting a distribution for a single $S(t)$ and using it to choose a parametric survival model

We first need to choose an appropriate time point, which we denote by T, and we elicit a distribution for $S(T)$. We recommend choosing T to consider the following five factors:

1. the latest available time point at which individual patient data are available (e.g., as seen in a Kaplan-Meier plot). If T is too close to this time point, the task may reduce to one of statistically extrapolating the trend seen in the available data, noting the discussion about statistical versus expert extrapolation (see Section 1.3);
2. how parametric models diverge in their extrapolations. We should avoid choosing T at a time point at which all statistically plausible models (models with acceptable fit to the available data as discussed in NICE DSU TSD 14[8] and TSD 21[9]) give similar extrapolations. An example of checking this is shown in Figure 3. Similar model extrapolations at time T could be an indication that this time point T is too close to the latest available time point in the data, as discussed in point 1;
3. time points at which the proportion of survivors is likely to be negligibly small. If T is too large, with $S(T)$ expected to be close to 0, expert opinion about $S(T)$ is not likely to add any new information;
4. limitations of expert knowledge. Experts may be unwilling to make judgements about $S(T)$ if T is too large;
5. the time horizon of the health economic model, and how the choice of T is likely to reduce uncertainty about the whole survivor function, up to the time horizon.

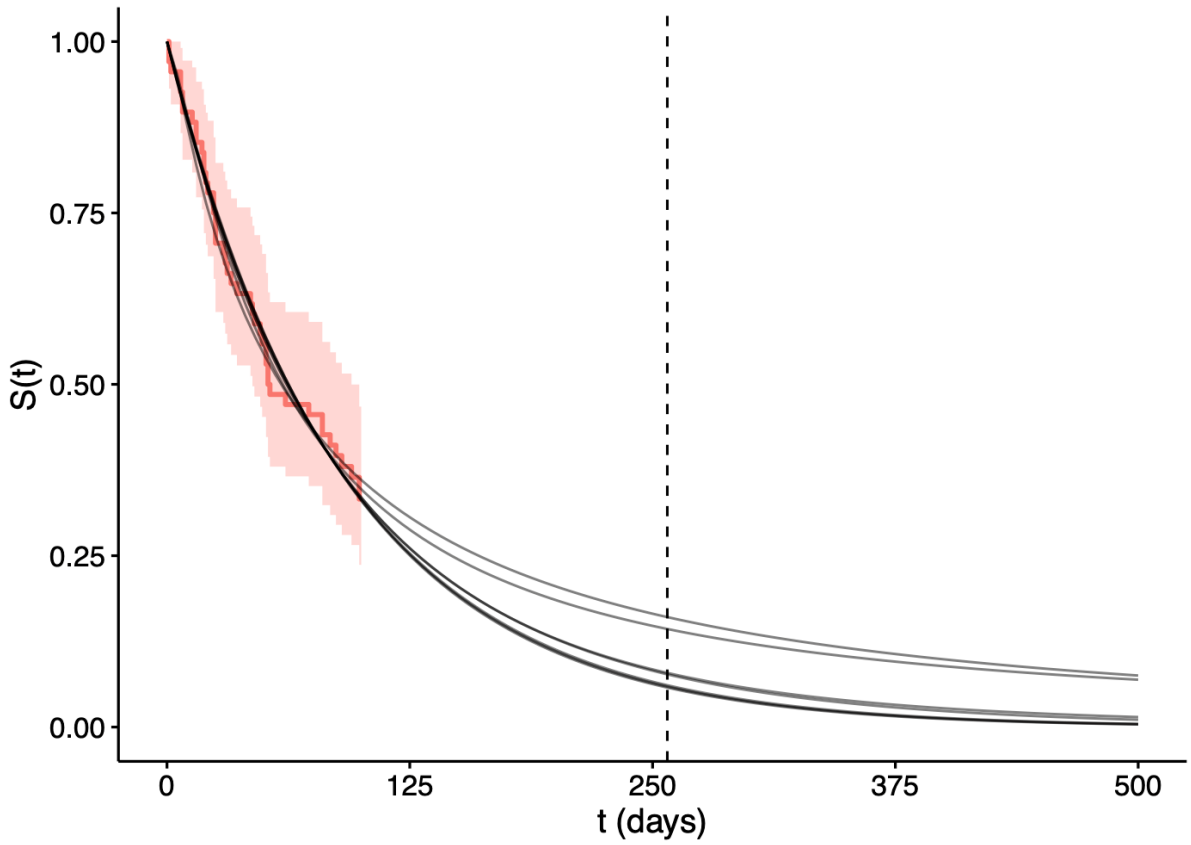

Figure 3: An example of checking models divergence in their extrapolation using the survivalModelExtrapolations() function in R package SHELF, with the model fitting implemented using the R package flexsurv.[29] The data are constructed from the veterans dataset in the R package survival.[10] The fitted models are exponential, Weibull, gamma, Gompertz, log-logistic, log normal and generalised gamma (some model fits are indistinguishable). The dashed line indicates that the greatest difference between the extrapolations is seen just after 250 days. The code to produce this example is in Appendix C.1.

3.5.1.1 Using the distribution for $S(T)$ to choose a parametric survival model for $S(.)$

As part of the elicitation procedure for $S(T)$, we recommend eliciting qualitative judgements about the hazard function: we discuss this in more detail in Section 3.7. Note that these would typically be qualitative judgements about the hazard over the extrapolation period only. These judgements, together with the elicited distribution for $S(T)$ could be used to select a parametric survival model as follows.

1. Identify a collection of candidate parametric models, fitted to the available data, and excluding any with poor fit, following the recommendations in NICE DSU TSD 14 and TSD 21.[8,9]
2. For each model remaining from Step 1, exclude any models with hazard functions in conflict with the qualitative expert judgements obtained about the hazard function.
3. For each model remaining after Step 2, compare quantiles for $S(T)$ from the candidate parametric models with those from the elicited distribution for $S(T)$. For example, we suggest a visual comparison of the median and 90% credible interval for each parametric model with the median and 90% credible interval for the elicited distribution. We do not suggest a formal criterion here, but a model may be excluded if there is a little or no overlap; a judgement would be made that, even allowing for expert uncertainty and model prediction uncertainty about $S(T)$, a model extrapolation is inconsistent with expert judgement.

If there are multiple models remaining after Step 3, then we recommend reporting cost-effectiveness estimates for each model, and acknowledging that neither the data nor expert judgement can provide a strong case for selecting one over another.

3.5.2 Eliciting a parametric survivor model for $S(.)$

Another option is to choose a parametric form for $S(.)$, and then construct a probability distribution for the parameters based on expert judgement. For example, we might assume a Weibull survivor function $S(t) = \exp(-(t/\lambda)^{\kappa})$ for $t > 0$, and then elicit probability judgements from which a joint probability distribution for λ, κ can be constructed (e.g., following the methodology in Ren and Oakley (2014)[30]). The experts would not normally propose the parametric form themselves; this would be selected by a facilitator, perhaps based on qualitative expert opinion about the hazard.

The difficulty with this approach is in choosing the parametric family of survivor functions and the circular chain of reasoning: we are using expert judgement because of the difficulty in selecting one parametric family of survival distributions, and yet we would now require a choice of parametric family to represent expert opinion.

3.5.3 Eliciting the survivor function at multiple time points

In this approach, we elicit judgements about the survivor function at multiple time points $S(T_1), \ldots, S(T_n)$, and then interpolate between these time points. This would be similar to the elicitation framework in Garthwaite et al. (2013).[31] One issue to consider in this case is dependence: if an expert were to discover the true value of $S(T_1)$ for some time T_1, this may change their opinions about the value of $S(T_2)$ for another time T_2; the elicitation method needs to produce a joint probability distribution that accounts for any such dependence.

Our view is that the eliciting judgements about the survivor function at a single time point is likely to draw out most of the substantive expert *knowledge* (the first source of knowledge discussed in Section 1.3). It is possible that, conditional on $S(T_1)$ for a particular T_1, consideration of $S(T_2)$ for another time point T_2 may reduce to an exercise in judging what a survivor function, in general, 'should' look like: what $S(T_2)$ should be relative to $S(T_1)$ to give a smooth-looking curve. Making judgements about $S(T_2)$, conditional on $S(T_1)$ could reduce to the 'non-expert' approach to the extrapolation we discussed in Section 1.3.

We do not recommend eliciting the survivor function at multiple time points *by default*. Rather, given elicitation for one $S(T)$, careful thought should be given to whether elicitation for additional time points will draw out additional expert knowledge. Rather than making additional quantitative judgements, the experts may be able to provide additional qualitative judgements about the shape of the survivor or hazard function.

3.6 BAYESIAN UPDATING

Having elicited a distribution for $S(T)$ (or $S(T_1)$, $S(T_2)$, … , $S(T_n)$ in the case using multiple points), we could attempt to derive a posterior distribution for $S(.)$ given both the elicited distribution and any available data. One problem here is that it is likely to be difficult (and perhaps undesirable) to withhold the available data from the experts at the point of the elicitation exercise; we discussed this issue and expert recruitment in Section 2.7.

Subjective probability judgements are conditional on whatever knowledge an individual has at the time. If an expert has provided judgements about $S(T_i)$ and has knowledge of data D, then we have, in effect, 'elicited a posterior' $P(S(T_i)|D)$. If we then attempt a Bayesian update for $S(.)$ using D and these elicited judgements, we would need to ensure that this does *not* change the distribution for $S(T_i)$ from what has already been elicited. If there is any change to this distribution then we have double-counted the data: the expert has used the data D to specify $P(S(T_i)|D)$, and then the data D has been used a second time to 'update' this distribution.

One possibility to avoid the double-counting problem is to instead elicit a distribution for the proportion of survivors at a time point T_i, out of those known to have survived for at least time t_0: the end of the trial period. We denote this by $S(T_i)|T > t_0$. This would be an attempt to construct a quantity that is independent of the trial data, that would still enable extrapolation. It is not clear if independence could be justified, hence there may still be a risk of double counting. The observed trend in the hazard in the period leading up to time t_0 might be expected to continue in the short term. Other ways the observed data might be informative for $S(T_i)|T > t_0$ would be if the trial data suggested a delayed treatment effect, or if there was a rapid decline in the survivor function early in the observation period, followed by a levelling-off, indicating a possibility of healthier longer-term survivors.

Nevertheless, there have been various methodological developments in combining data and expert judgement for survival extrapolation, and this is currently an active area of research.[32-34]

3.7 INCORPORATING QUALITATIVE OPINION ABOUT THE HAZARD FUNCTION

In addition to eliciting a distribution for one point on the survivor function $S(T)$, we recommend obtaining qualitative opinions about the hazard function over the extrapolation period. We suggest two ways in this can be done: working through a 'hazard checklist', and 'scenario testing'. We recommend the first approach is incorporated in any structured expert elicitation of long-term survival; we suggest the second approach as an option. We illustrate incorporating both these methods within an elicitation protocol in Section 4.2.

3.7.1 A hazard checklist

The experts can be asked to discuss factors that may cause the hazard to increase or decrease over the extrapolation period. We suggest working through a checklist with the experts, that groups potential factors by patient characteristics, disease progression, and mechanism of treatment action.

Although we suggest only asking for qualitative judgements, experts should still be asked to reflect on their uncertainty about such factors, as this may help them to assess uncertainty about survivor function values.

A discussion between the experts might be structured as follows. Here, we suppose there are two treatment groups and therefore two hazard functions; some factors should be considered jointly for the two hazard functions.

1. The experts are asked to consider the hazard function over the extrapolation period, considering the available data. The experts would be shown a plot of an empirical hazard and/or given a table of estimated hazards over contiguous time intervals for the observed period to assist with their judgements over the extrapolation period.
2. Depending on the age profile of the target population, the dominant effect on the hazard may be the increasing age of the patients; this should be checked with reference to life tables. In this case, this should be discussed with the experts, who should then be asked to consider *residual* effects on hazard from other factors.
3. The experts are first asked to consider factors that may increase the hazard.
    - It should be emphasised that the experts are not being asked to *predict* what the hazard will do: they are being asked to consider *possible* factors that would have an increasing effect on the hazard, even if the net effect were downwards.
    - They should consider patient characteristics, disease progression, and mechanism of treatment action.
    - One source of uncertainty may be the characteristics of the patients who have survived to the end of the trial period.

- The experts should comment on whether any of the factors would apply to both treatment groups, or to one only.
- The experts should comment (qualitatively) on how treatment effects might change over time, e.g., if sustained benefits of treatment would be expected, or whether the effect of the treatment might wane over time.

4. The experts are then asked to consider factors that may decrease the hazard
   - The same considerations apply as in the increasing hazard case.
5. The experts are asked to comment on whether, considering the discussion from points 3 and 4, they would expect a net increase or decrease in hazard (or residual hazard, if age is the dominant factor), with the option to say that they are uncertain, and that either is possible.
6. The experts are asked if they wish to provide any additional qualitative or quantitative opinions, for example, about timings of changes in hazard.

If there is disagreement between the experts, this should be recorded, but we do not suggest attempting to achieve a consensus view.

3.7.2 Scenario testing

The hazard checklist will help each expert to formulate their opinions about how the hazard might change over time. It may then be possible to link such opinions to quantitative judgements about the survivor function.

We consider eliciting a distribution for the survivor function at a single time point: $S(T)$. In the scenario testing method, we make an assumption about the hazard function, the "scenario", and present the implications for this regarding uncertainty about $S(T)$, for example, a probability distribution[§] for $S(T)$ given the available data and the assumption (but no other information otherwise). There are two conditions if using this approach:

[§] more specifically, an approximate posterior distribution given the observed data and a noninformative prior, technical details are given in Appendix A.2.

1. The experts should be asked to make their own probability judgements about $S(T)$ before being presented with any scenarios, to avoid their judgements being anchored on the scenario.
2. It must be emphasised to the experts that the scenario is not a claim about the true behaviour of the hazard function; the choice of the word "scenario" is an attempt to reinforce this.

The experts can then compare their own judgements with the probability distribution for $S(T)$ under the stated scenario. This may reveal inconsistencies between an expert's judgements and their opinions about the hazard and so provide useful feedback.

The choice of scenario will depend on the feasibility of deriving the corresponding probability distribution for $S(T)$, and the interpretability of the results: the usefulness of any feedback that can be given to the experts. We suggest considering a constant hazard scenario.

Specifically, we suggest choosing some time point t^*, where $t^* < t_0$, where t_0 is the last observed time point in the individual patient-level data (e.g., the final quarter of the observation period in the clinical trial). We then consider a scenario of constant hazard from time t^*. Based on this assumption, an approximate 95% interval for $S(T)$ can computed, denoted by $\left(S_{0.025}(T), S_{0.975}(T)\right)$, with details of the computation given in Appendix A.2.

The interpretation of $S(T)$ exceeding $S_{0.975}(T)$ is that it is highly likely that the hazard must have *decreased* between times t^* and T. If an expert has given non-negligible probability to the event $S(T) > S_{0.975}(T)$ then the expert should confirm that they think a decrease in hazard is plausible; otherwise, they should modify their judgements with the effect of reducing their probability of $S(T)$ exceeding $S_{0.975}(T)$. Elicitation of individual expert judgements is discussed in Section 4.1.

Similarly, the interpretation of $S(T)$ less than $S_{0.025}(T)$ is that it is highly likely that the hazard must have *increased* between times t^* and T. If an expert has given non-

negligible probability to the event $S(T) < S_{0.025}(T)$ then they should confirm that they think an increase in hazard is plausible; otherwise, they should modify their judgements with the effect of reducing their probability of $S(T)$ being less than $S_{0.025}(T)$.

Note that no conclusions can be drawn regarding changes to the hazard for $S(T)$ lying within the interval $\left(S_{0.025}(T), S_{0.975}(T)\right)$. This is because both the constant hazard assumption *and* alternative assumptions with changing hazards can result in $S(T)$ lying inside the interval. We give a more technical discussion of this in Appendix A.2.

To summarise:

- if an expert has judged significant probability of $S(T)$ ***above*** $S_{0.975}(T)$, they can be given feedback that this would suggest a decrease in hazard at some point after time t^*;
    - if an expert has judged significant probability of $S(T)$ ***below*** $S_{0.975}(T)$ then the expert should simply be told that there is no feedback that can be reported from this.
- if an expert has judged significant probability of $S(T)$ ***below*** $S_{0.025}(T)$, they can be given feedback that this would suggest an increase in hazard at some point after time t^*;
    - if an expert has judged significant probability of $S(T)$ ***above*** $S_{0.025}(T)$ then the expert should simply be told that there is no feedback that can be reported from this.

A function for implementing the scenario test for the constant hazard assumption is available in the SHELF R package,[22] and an illustration is shown in Figure 4.

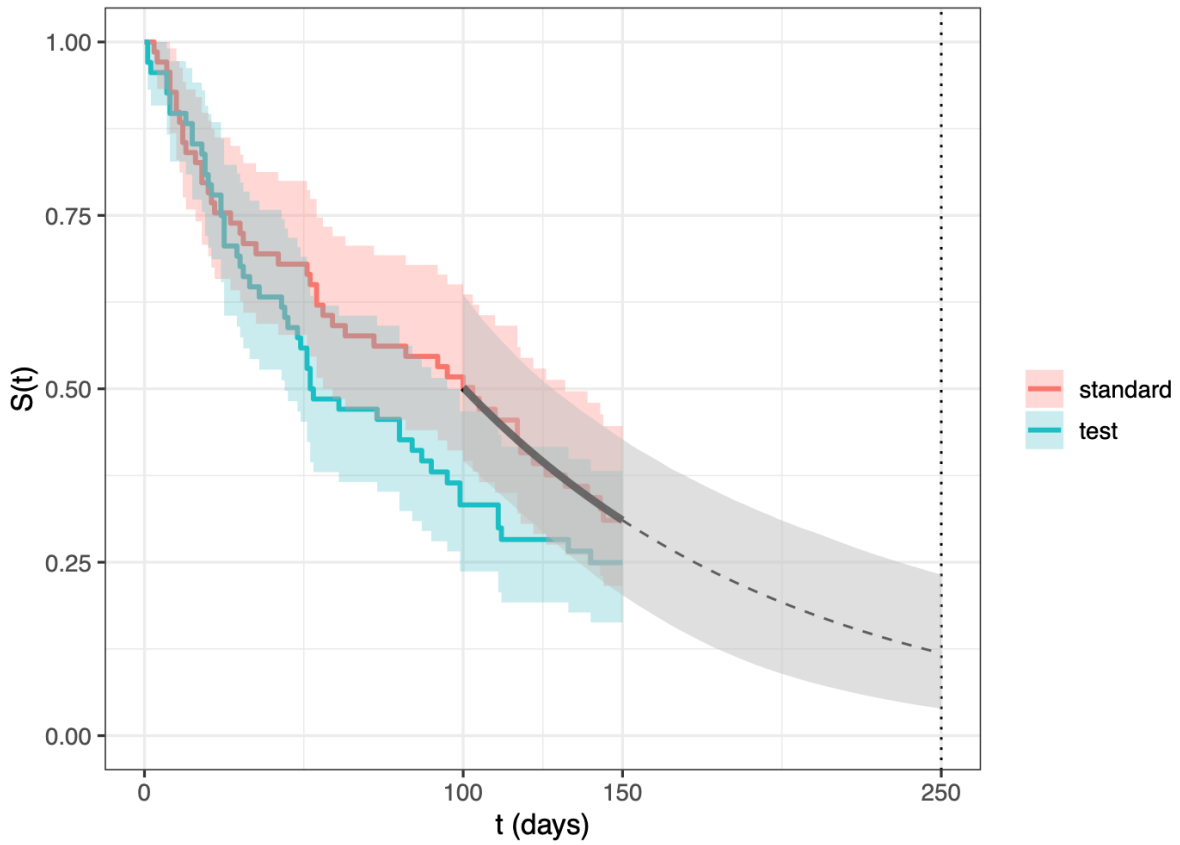

Figure 4: Implementing the scenario test for the constant hazard assumption, using the veterans data in the R package survival.[10] This is implemented using the R package SHELF.[22] For the "standard" group, a constant hazard is assumed for times exceeding 100 days, and an interval $\left(S_{0.025}(T = 250), S_{0.975}(T = 250)\right)$ is obtained as (4%, 23%). Hence, based on the available data, for $S(T = 250)$ to exceed 23% (by some non-trivial amount), we would expect a decrease in hazard at some point after 100 days. The code to produce this example is in Appendix C.1.

Note that the choice of t^* will need careful investigation, in advance of the presentation of the analysis to the experts. The assumption of constant hazard within the interval (t^*, t_0) needs to look plausible, as assessed visually by the fit to the observed data within this interval. The credible interval $\left(S_{0.025}(T), S_{0.975}(T)\right)$ should not be too sensitive to the choice of t^*; small changes may be acceptable, given that experts are likely to acknowledge some imprecision in their judgements. If it is difficult to find a suitable t^* that can lead to useful feedback via the interval $\left(S_{0.025}(T), S_{0.975}(T)\right)$, then we would not recommend using this method.

Regarding other choices of scenario to present, we commented previously on the need to be able to derive the credible interval, and that the assumptions of the scenario need to be meaningful to the expert, such that they might adjust their judgements if there is conflict with the credible interval. For example, if an expert has given substantial probability to $S(T)$ exceeding $S_{0.975}(T)$ in the constant hazard scenario, the expert can reflect on whether they think a reduction in hazard is likely over the interval $[t^*, T]$.

Whilst it may be possible to construct additional scenarios that add value to the elicitation process, here we make no recommendations beyond presentation of the constant hazard scenario. For example, we would *not* present a scenario in which survival times are assumed to have a Weibull distribution over some interval. If an expert has given substantial probability to $S(T)$ exceeding $S_{0.975}(T)$ under a Weibull distribution scenario, we would not expect an expert to be able to reflect meaningfully on whether their distribution of survival times should lie outside the family of Weibull distributions, and whether they should adjust their probability.

3.8 HAZARD RATIOS AND RELATIVE TREATMENT EFFECTS

In other contexts involving elicitation and survival analysis, it is sometimes useful to elicit a probability distribution for a hazard ratio, even though it is not a directly observable quantity. For example, Salsbury et al. (2024) elicit distributions for hazard ratios in the context of planning clinical trials where delayed treatment effects are expected.[35] The rationale is that experts may think of relative treatment effects in terms of hazard ratios, making the hazard ratio an appropriate target quantity for elicitation. A critical assumption here is proportional hazards; we may be less willing to make such an assumption for extrapolation, compared with trial planning. Non-proportional hazards may already be apparent in the observed data. We do not recommend eliciting a distribution for a hazard ratio for extrapolation purposes.

More generally, there is a question of whether to elicit quantitative judgements about extrapolated relative treatment effects. Note that consideration of the hazard checklist in Section 3.7.1 can involve qualitative judgements about relative treatment effects,

when considering factors that may affect the hazard in both two treatment groups, or one group only.

The approach we have suggested in Section 3.5.1.1, if applied to two treatment groups, would involve choosing survival models for each treatment group based on fit to the observed data, and consistency of the extrapolations of these models with expert judgements about $S(T)$ for a suitable time point T. The fit to the observed data in each treatment group implies a relative treatment effect, which is then extrapolated forwards in time. The assumption in 3.5.1.1 is that if the model extrapolations are consistent with the elicited distributions for $S(T)$ in each group, we then suppose the extrapolated relative treatment effect is also appropriate.

It would be possible to elicit further judgements regarding treatment effects at time T, for example, about differences between survivor proportions between two treatment groups at this time. This could result in an elicitation process that is more difficult for the experts: there would be various technical challenges to work through to ensure consistency between probability judgements and/or to avoid asymmetries where uncertainty about one survivor proportion greater than uncertainty about the other. It is not clear if this additional elicitation would add value to the process set out in section 3.5.1.1 and we do not consider it further here.

4 AN EXAMPLE PROTOCOL

We now give an example protocol, bespoke for survival extrapolation, based on the task of eliciting a distribution for a single $S(T)$. This is to illustrate how an existing protocol (SHELF in this example) can be adapted to incorporate the additional aspects particular to survival extrapolation, as discussed in Section 3.

A flow diagram of the process is presented in Figure 5. Templates and further advice for this protocol are available at https://shelf.sites.sheffield.ac.uk/, and supporting software is included in the SHELF R package.[22] An online app is available via the website. The same app is available for offline use in the R package; this may be preferable to some users as the app requires uploading of individual patient-level data. Instructions for installing the SHELF R package, along with the code used for the examples in this report, can be found in Appendix C.1.

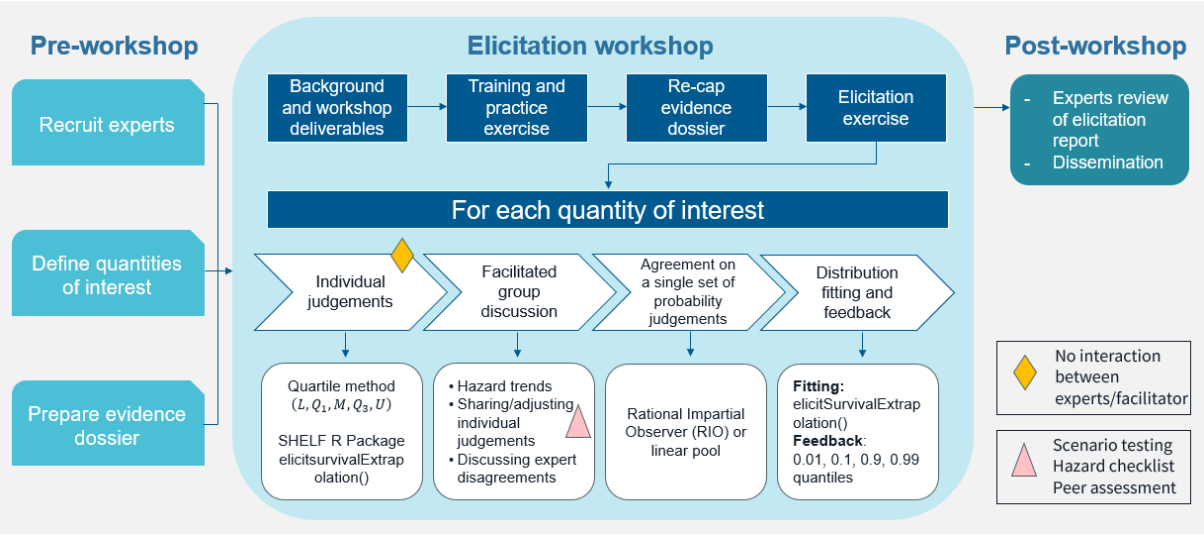

Figure 5: A flow diagram of the example protocol procedure.

We assume the expert panel has been recruited, an evidence dossier has been prepared and circulated, and the experts have received training, as described in Sections 3.1 to 3.4. A time point T has been selected following the guidance in Section 3.5, and the aim is to elicit a probability distribution for $S(T)$. The experts are participating in a joint meeting (either in-person or online), conducted by a facilitator.

4.1 INDIVIDUAL ELICITATION

Each expert is asked to provide an initial set of probability judgements about $S(T)$, without conferring with other experts or with the facilitator. This is to establish what each expert thinks and should be documented in the report of the elicitation exercise. The quartile method is used, to avoid the facilitator proposing any numerical values of $S(T)$ to the experts. Each expert is asked for the following five quantities, and makes all five judgements before sharing them with the other experts or the facilitator. The five judgements are:

- **a lower plausible limit** L. An expert should be confident in ruling out any values of $S(T)$ below their choice of L as implausible. The smallest possible value of $S(T)$ would be 0; experts are asked to consider if they can propose L greater than 0;
- **an upper plausible limit** U. An expert should be confident in ruling out any values of $S(T)$ above their choice of U as implausible. Again, experts should consider choosing U below a maximum *possible* value of $S(T)$. If t_0 is the last time point at which a Kaplan-Meier estimate of the survivor function is available, a point of reference here would be a suitable upper confidence limit for $S(t_0)$ (e.g., a reported 95% confidence interval for $S(t_0)$), combined with an assumption of no further deaths between times t_0 and T;
- **a median value** M. This splits the interval $[L, U]$ into two intervals $[L, M]$ and $[M, U]$ such that the expert would judge that the two intervals have equal probability (0.5) of containing $S(T)$. Experts are likely to find it difficult to propose a unique value of M, but a *precise* value for M is not important at this stage. The important point is that, having chosen M, the expert *cannot* easily identify $[L, M]$ as more likely to contain $S(T)$ than $[M, U]$, or vice-versa;
- **a lower quartile** Q_1. This splits the interval $[L, M]$ into two intervals $[L, Q_1]$ and $[Q_1, M]$ such that the expert would judge that the two intervals have equal probability (0.25) of containing $S(T)$. Similar advice applies regarding difficulties in choosing a precise value of Q_1 as with the median. A further prompt for the experts is to consider, if M was to be used as an estimate for $S(T)$, how close they would expect $S(T)$ to be to M; the more uncertain they are, the further they should place Q_1 from M. It should, however, be explained to the experts that we

would usually expect Q_1 to be closer to M than to L; their median value is M; but L is at the extreme of what they consider to be plausible;

- **an upper quartile** Q_3. This splits the interval $[M, U]$ into two intervals $[M, Q_3]$ and $[Q_3, U]$ such that the expert would judge that the two intervals have equal probability (0.25) of containing $S(T)$. The same considerations apply as in the case of specifying the lower quartile Q_1.

Following the specification of these five values, each expert should check that they have no clear preference for selecting one of the four intervals $[L, Q_1]$, $[Q_1, M]$, $[M, Q_3]$, $[Q_3, U]$ as having a significantly higher probability than any other of containing $S(T)$, as the implication is that the expert would judge a 0.25 probability for each interval containing $S(T)$. The experts can be asked to consider a bet, in which they choose one of the four intervals, and receive a reward if $S(T)$ lies in the selected interval. They should *not* have a clear preference for any one interval out of the four.

The approach of eliciting quartiles combines both variable interval and fixed interval methods. An expert is asked to make variable interval judgements when providing quartiles, but fixed interval judgements when asked, for example, to confirm that they would judge a 0.5 probability of $S(T)$ lying in the interval $[L, M]$.

The SHELF resources (https://shelf.sites.sheffield.ac.uk/) contain slides designed to guide the experts through the process of making these judgements. These slides are presented to the experts as they make each judgement.

4.2 GROUP DISCUSSION

The next step in the process is to have a facilitated discussion between the experts, where differences of opinion can be debated. The experts should be alerted to the considerations regarding group interaction discussed in Section 2.5, specifically the importance of ensuring all group members contribute fully, and the role that each expert must play in providing scrutiny of other opinions put forward.

The group discussion can be managed in six stages. We outline these first, before describing each in more detail.

1. Qualitative discussion of hazard over the extrapolation period, including discussion of a hazard checklist. The conclusions of this discussion should be reported alongside elicited probability distributions.
2. Presentation of the individual judgements. If the scenario testing method is used, this is discussed and presented at the same time. The facilitator provides brief commentary and interpretation but does not yet invite debate between experts regarding quantitative judgements.
3. Optional adjustment of any individual probability judgements.
4. Identification and discussion of significant disagreements between experts.
5. Agreement on a single set of probability judgements.
6. Distribution fitting and feedback.

4.2.1 Qualitative discussion of hazard

The experts are invited to discuss potential changes to the hazard over the extrapolation period. As a prompt, the points discussed in Section 3.7.1 are presented. The experts should consider, separately, factors that may increase the hazard, and factors that may decrease the hazard. The experts' opinions are recorded, specifically, qualitative opinions regarding how the hazard might change after the observed period and what could cause this. Disagreements should be noted.

Expert comment on the hazard does not need to be collected individually; this can be collected via group discussions, but it is the facilitator's role to ensure that all experts contribute to the discussion. Experts are not required to reach a consensus on the leading factors contributing to hazard changes; disagreements would be reported.

4.2.2 Sharing of individual judgements and scenario testing

The experts now share their judgements from the individual elicitation stage, and these are displayed graphically. If appropriate to do so, the facilitator now implements the scenario testing method for discussion (i.e., if there is suitable scenario that has the potential to generate useful feedback, as discussed in Section 3.7.2).

Table 1 shows a hypothetical example of three experts' individual judgements of $S(T)$. Figure 6 presents this hypothetical example and a 95% credible interval $\left(S_{0.025}(T), S_{0.975}(T)\right)$ based on a constant hazard scenario. Each expert's individual judgements with the four equally probable intervals $[L, Q_1]$, $[Q_1, M]$, $[M, Q_3]$, $[Q_3, U]$ are indicated as four coloured vertical bars in Figure 6. We suppose the constant hazard scenario has a resulted in a 95% credible interval (6%, 21%), indicated as the dashed lines in Figure 6.

Table 1: A hypothetical example of three experts' individual judgements.

	Expert A	Expert B	Expert C
$\boldsymbol{U}$	20%	15%	30%
$\boldsymbol{Q_3}$	12%	12%	20%
$\boldsymbol{M}$	10%	11%	12%
$\boldsymbol{Q_1}$	8%	10%	10%
$\boldsymbol{L}$	0%	8%	7%

$\boldsymbol{U}$, upper plausible limit; $\boldsymbol{Q_3}$, upper quartile; $\boldsymbol{M}$, median; $\boldsymbol{Q_1}$, lower quartile; $\boldsymbol{L}$, lower plausible limit

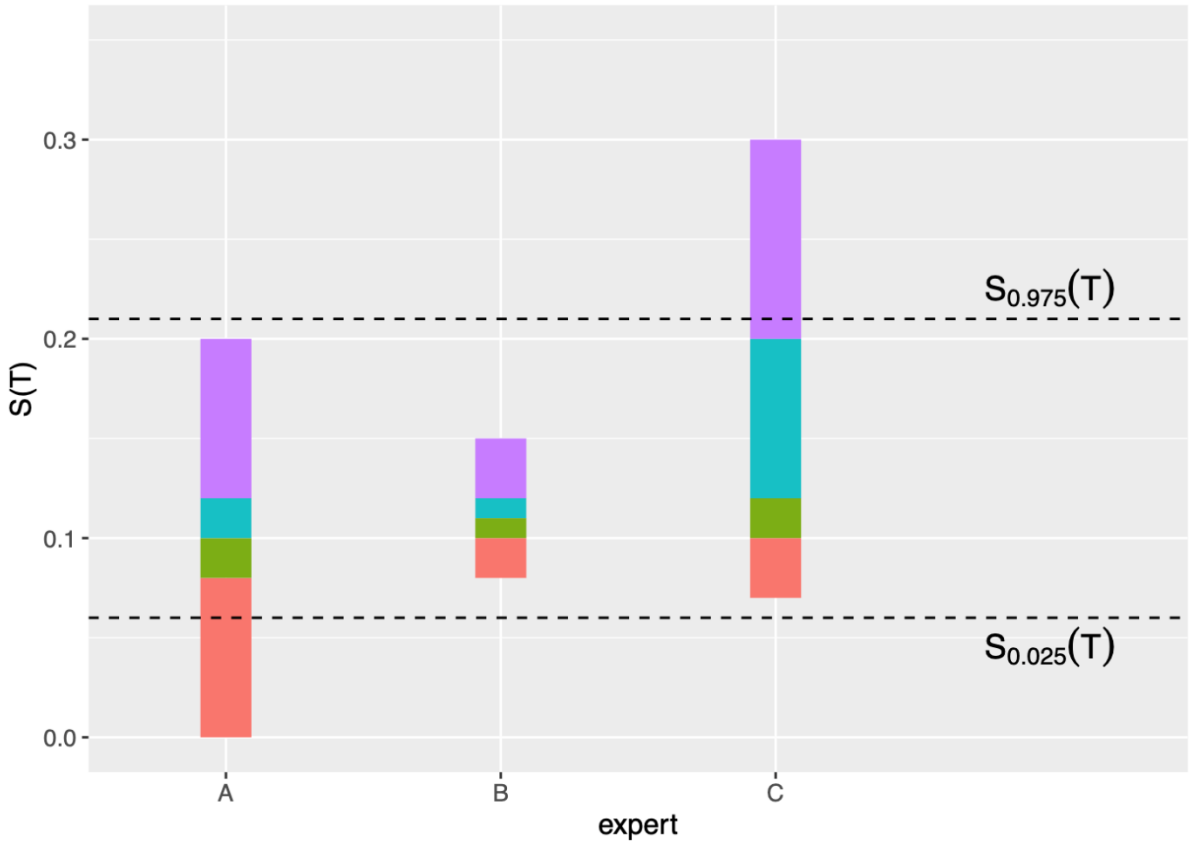

Figure 6: A hypothetical example of elicited individual judgements using the quartile method and the limits of constant hazards derived from scenario testing.

The facilitator should explain the scenario clearly (a constant hazard from some time t^* up to time T), making clear that nothing is claimed about the plausibility of the scenario: it is a point of reference for the experts. The facilitator should also explain how to interpret the credible interval of the scenario (dashed lines); values outside the limits are *only* likely to be the consequence of a change in hazard, but values inside could result from *either* constant or changing hazard.

In the example, the facilitator would check the following:

- that expert A thinks an increasing hazard is possible, otherwise the expert's lower plausible limit L would be too low. No conclusions can be drawn from the individual judgements regarding whether expert A thinks a decreasing hazard is possible;
- that expert C thinks a decreasing hazard is possible, otherwise the expert's upper plausible limit U would be too high. No conclusions can be drawn from the individual judgements regarding whether expert C thinks an increasing hazard is possible.

The facilitator would explain to expert B that because expert B's plausible range lies entirely within the credible interval, but that values within the credible interval can result from either constant or changing hazard, there is no feedback that can be reported to expert B *from this scenario test*. It may be helpful to use phrasing such as, "This method has failed to provide feedback on your judgements," to emphasise that nothing is inferred about expert B's judgements from this analysis. It may also help to emphasise that all experts will be given the opportunity to change their judgements in any case; expert B may wish to do so if they changed their opinions about the hazard.

4.2.3 Optional adjustment of any individual probability judgements

At this point in the discussion, some experts may wish to revise their initial judgements. Possible reasons for this would be

1. an expert has revised their judgements, following the qualitative discussion of hazards;
2. the scenario test suggests an inconsistency between an expert's judgements and their opinions about the hazard function;
3. seeing the judgements of peer experts is sufficient to make an expert change their mind;
4. an expert has misunderstood the elicitation task.

An example of how point 3 might occur relates to the process of eliciting plausible limits. When eliciting, for example, an upper plausible limit U, an expert is asked to imagine a reported estimate of $S(T)$ greater than U. They are asked to consider their reaction; if U really was their upper plausible limit, they might suspect an error in the reported estimate or a flaw in the study. But there is a difference between *imagining* a reported estimate greater than U, and *observing* a peer expert state values greater than U to be plausible. In the latter case, an expert may be more willing to modify their original judgements.

From our experience, we would expect some experts *not* to make any adjustments at this stage (and they should not feel an expectation to do so), but some may wish to. Allowing experts to change their individual judgements and updating the display such as Figure 6 may avoid unnecessary discussion at the next stage.

4.2.4 Identification and discussion of significant disagreements between experts

A plot such as Figure 6 provides a comparison of the experts' individual judgements. There are likely to be some differences between experts that can be thought of as 'noise'. For example, when judging their median value, an expert might report a value of 10% but might struggle to articulate why their median was 10% rather than, say, 13%. Discussions around relatively small differences are unlikely to be fruitful, and we give advice for resolving small differences in the next section.

We suggest focussing the discussion on instances where one expert has reported an upper quartile greater than another expert's upper plausible limit, and where one expert has reported a lower quartile less than another expert's lower plausible limit. In

effect, we classify a disagreement as 'significant' where one expert is giving a probability of at least 0.25 to some interval for $S(T)$, but another expert is ruling that interval out as 'implausible'. An example is given in Figure 7.

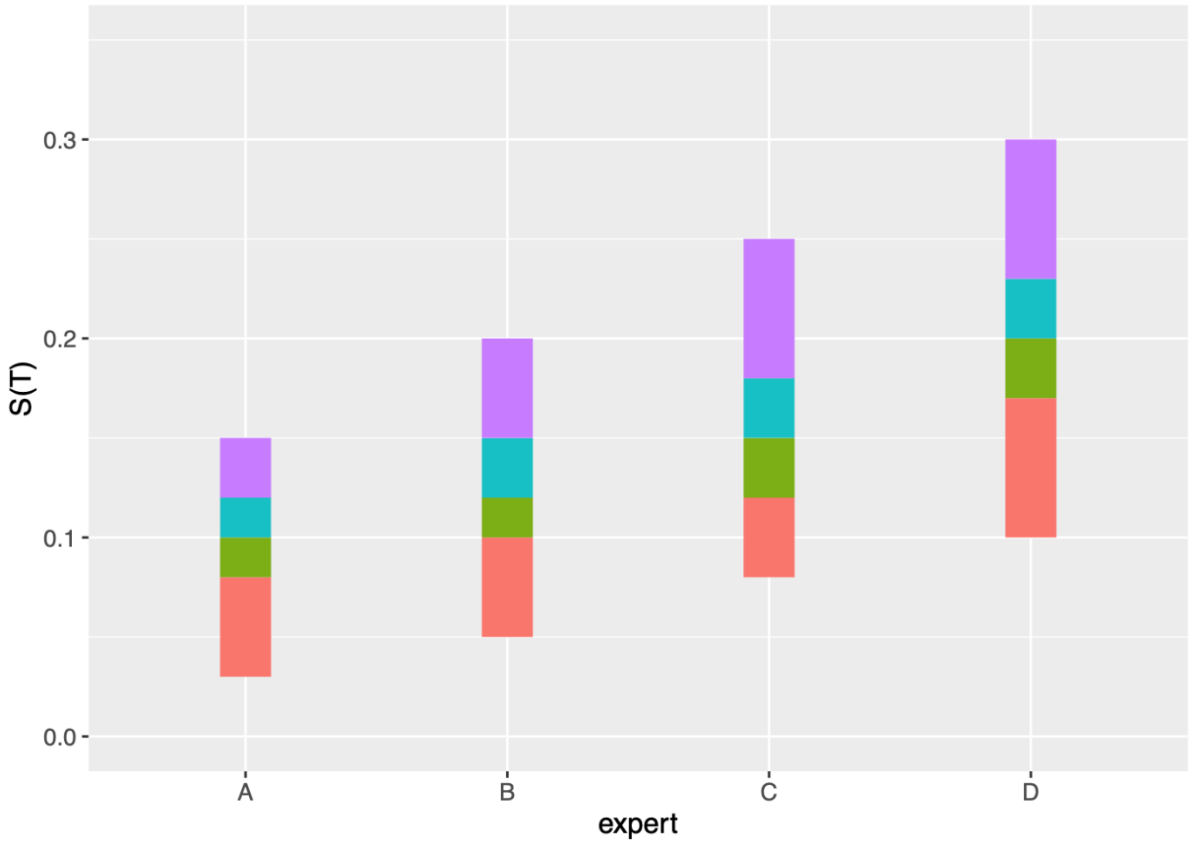

Figure 7: An example of four expert individual judgements with varying degrees of disagreement.

In this example, we would consider there to be significant disagreement between expert A and experts C and D. Expert A has ruled out $S(T)$ exceeding 15% as implausible, but experts C and D have both given substantial probability (0.5 and at least 0.75 respectively) to $S(T)$ exceeding 15%. The facilitator would then ask expert A to justify their choice of U, invite experts C and D to respond, and give other experts an opportunity to comment. A possibility is that expert A is using an assumption or a particular piece of evidence as the basis for their upper plausible limit; group discussion may help establish the appropriateness of the assumption/relevance of the evidence.

Other differences of opinion may be less significant. For example, comparing experts B and C, expert C has given higher quartile values, but discussion may not necessarily result in much insight regarding why one expert has chosen higher values than another.

In summary, important discussion points will involve comparison of the smallest upper plausible limit with the largest upper quartile, and comparison of the largest lower plausible limit with the smallest lower quartile. All experts should be invited to comment on these disagreements, as well as being able to state and debate any other judgements that they think are relevant.

Once the discussion has reached a clear end, with no further points to be made by the experts, the facilitator or recorder should provide a recap of the main discussion, before proceeding to the next stage.

4.2.5 Agreement on a single set of probability judgements

The facilitator now seeks a set of probability judgements from which a single distribution can be fitted. In SHELF, this involves invoking a notion of a "Rational Impartial Observer (RIO)". RIO is assumed to have studied the evidence dossier and observed and understood all the group discussion. RIO is also assumed to be impartial: they will not favour one expert any more than any other, but they may have observed some arguments to have been evidenced more strongly than others. The group is asked to propose and agree probability judgements that such a RIO would make, recognising that this is different to asking any expert what *they* think. If there is no obvious disagreement between the experts to investigate, a linear pool can be chosen instead.

Note that the experts are not being asked to come to a consensus regarding *their own opinions about* $S(T)$; they are only asked to come to an agreement on what probability judgements RIO would make. If there is no expert consensus regarding $S(T)$, this would contribute to greater uncertainty about $S(T)$ from RIO's perspective.

We suggest using fixed interval methods at this point, to help the experts distinguish this part of the process: agreeing on RIO's probabilities, from the earlier step of making their own probability judgments. Note that because the experts have already provided variable interval judgements, it is now easier to identify suitable intervals for fixed interval methods. We suggest considering three values of $S(T)$ based on values discussed thus far: one in the lower tail, one in the upper tail, and one more central value. Suitable values may be apparent from the individual judgements and group discussion. Alternatively, the linear pool functions in the SHELF R package can be used to identify appropriate values of $S(T)$, by obtaining quantiles from the tails of a linear pool distribution. Code to illustrate this is given in Appendix C.1. For example, continuing from Figure 7 the facilitator might ask the experts to agree on RIO's probabilities for

$$P(S(T) \leq 0.1),\ P(S(T) > 0.2),\ P(S(T) \leq 0.15).$$

4.2.6 Distribution fitting and feedback

The facilitator now fits a distribution to the elicited probabilities, for example, using the SHELF R package.[22] A plot of the density function is displayed, and quantiles from the fitted distribution are reported. A beta distribution is a natural choice as $S(T)$ is constrained to lie in [0,1], but other distributions are available in the software if no satisfactory beta fit can be found. In our experience, we have sometimes found it easier to find an appropriate fit using a skew-normal distribution (available in the SHELF R package): this is a three-parameter family and so can offer more flexibility.

As feedback, we suggest reporting the 0.01, 0.1, 0.9 and 0.99 quantiles from the fitted distribution for $S(T)$, which we denote by $\hat{S}_{0.01}(T),\ \hat{S}_{0.1}(T),\ \hat{S}_{0.9}(T),\ \hat{S}_{0.99}(T)$. If the fitted distribution is acceptable to the experts as a representation of RIO's judgements, this implies RIO would judge

$$P\left(S(T) \leq \hat{S}_{0.01}(T)\right) = 0.01,\ \ldots, P\left(S(T) \leq \hat{S}_{0.99}(T)\right) = 0.99,$$

hence, informally, the experts should confirm that

1. RIO would 'rule out' (only give 1% probability to) $S(T)$ being less than $\hat{S}_{0.01}(T)$,

2. but RIO would *not* 'rule out' (they would give 10% probability to) $S(T)$ being less than $\hat{S}_{0.1}(T)$.
3. RIO would 'rule out' (only give 1% probability to) $S(T)$ exceeding $\hat{S}_{0.99}(T)$,
4. but RIO would *not* 'rule out' (they would give 10% probability to) $S(T)$ exceeding $\hat{S}_{0.9}(T)$.

The aim here is to find an appropriate trade-off between avoiding overconfidence, by considering conditions 1 and 3 above, and avoiding underconfidence, by considering conditions points 2, and 4.

Some iteration may be required, with modifications made (and documented) to the RIO probabilities, and re-fitting as appropriate. Once an acceptable distribution is found, the elicitation is concluded.

After the workshop, the SHELF template, available as part of the SHELF resources (https://shelf.sites.sheffield.ac.uk/), is completed to provide a written record of the elicitation and circulated to the experts for approval. The record can also be circulated to other external experts for validation, although we should note that the external experts may not have received similar training in making and interpreting probability judgements, as discussed in Section 2.9.

The protocol is presented here for a single treatment. For multiple treatments, we would conduct all the individual elicitation steps first, so that individual opinions are first established and recorded before any interaction between experts.

4.3 Feasibility of adapting and implementing other protocols

At the time of writing, we have conducted seven structured expert elicitation exercises for survival extrapolation. We based the elicitation on the SHELF protocol in each case (six online workshops, and one in-person workshop). The modification for survival extrapolation evolved as we developed methodology and gained experience, resulting in the protocol presented above. We consider feasibility of using other protocols by considering whether there would have been any difficulties if we had used a different protocol instead of SHELF.

The IDEA protocol, though the detailed implementation is different to SHELF, is broadly similar regarding its organisation, the convening of experts and the general tasks required of them. There would have been no additional difficulties had we chosen to use the IDEA protocol rather than SHELF. The IDEA protocol involves facilitated group discussion and so provides the same opportunity for qualitative discussion of hazard, and scenario testing, if appropriate. The IDEA protocol does not necessarily involve the same distribution fitting step following the aggregated experts' judgements, but this would be a straightforward addition.

Regarding Cooke's classical method, first note that Williams et al. (2021)[36] implemented both SHELF and this method in parallel in a healthcare setting, illustrating their similarity in general terms at an organisational level. Cooke's classical method requires more preparation work for the team conducting the elicitation: suitable "seed" questions must be identified. These questions are test elicitation questions where the answers are known to the facilitator, but not the experts. Here, we would suggest using seed questions based on survival extrapolation from published (and suitably truncated) Kaplan-Meier plots; this is how we developed our training exercise. Cooke's classical method would typically be implemented in a workshop format with discussion sessions between experts, and so there would be opportunity to incorporate qualitative discussion of hazards. Experts could be presented scenario testing results and given the option to adjust their judgements if desired, whilst maintaining the process of experts making their judgements individually.

The MRC protocol can be implemented with either group interaction between experts, or via a Delphi process; the four options listed in Section 2.9 would all be consistent with the MRC protocol. If there is group interaction, then regarding general practicalities, adapting from SHELF to the MRC protocol would be similar to adapting from SHELF to the IDEA protocol.

Had we attempted a Delphi process, this would have changed how the elicitation was conducted more significantly; it is harder for us to assess the feasibility of Delphi based on our experience of using SHELF. If the Delphi method was conducted using one-to-one interviews of each expert (option 3 in Section 2.9), this would be similar to the training and individual elicitation stages we implemented with SHELF. Qualitative

discussion of hazard and scenario testing could also be implemented in a one-to-one interview, though exploring between-expert variability would be more difficult. We would be least confident in judging the feasibility of using Delphi via a survey (option 4 in Section 2.9), in part because of the concerns discussed in that section, and because the exercise would be too far removed from what we have tested.

In summary, based on our experience, we believe all five protocols to be suitable for structured expert elicitation for survival extrapolation, but in the case of Delphi methods, this is on the basis that there are one-to-one interviews of each expert by a facilitator.

5 STRUCTURED EXPERT ELICITATION FOR LONG-TERM SURVIVAL OUTCOMES IN THE BROADER LITERATURE AND NICE TECHNOLOGY APPRAISALS

As part of this TSD, we conducted a systematic literature review on the use of structured expert elicitation for long-term survival outcomes in the broader literature. Additionally, we conducted a pragmatic review of NICE oncology technology appraisals (TAs) to assess how structured expert elicitation for long-term survival outcomes has been applied in NICE submissions. To enhance our findings, we also described studies from both reviews that used expert consultation to inform survival extrapolation as opposed to focussing solely on studies which used more structured approaches. These insights were subsequently used to develop the recommendations presented in Section 6.

Appendix D.1 and D.2 describe the methods including search strategy, eligibility criteria and data extraction process for the reviews of the broader literature and NICE TAs respectively, and we summarise the findings below. The protocol for this review was published on the Open Science Framework.[37]

5.1 RESULTS: REVIEW OF THE BROADER LITERATURE

A total of 354 studies were screened and reviewed, with the selection process summarised in Appendix D.3. Eleven studies were deemed relevant to this review, including six that used structured expert elicitation methods and five that used expert consultation for long-term survival outcomes (see Table 2). Here, as in previous sections, ‘expert consultation’ refers to methods in which experts are presented with model extrapolations or landmark survival predictions and asked to assess their plausibility, as opposed to being asked to provide quantitative estimates with accompanying uncertainty.

All included studies were conducted between 2009 and 2023, with eight focusing on oncology, one on cardiovascular disease, one on COVID-19, and one on renal diseases.

Table 2: Studies included in the review of broader literature.

Authors	Title	Publication Date	SEE/Expert consultation
Miksad et al.[38]	Interpreting trial results in light of conflicting evidence: a Bayesian analysis of adjuvant chemotherapy for non-small-cell lung cancer	2009	Consultation
Moatti et al.[39]	Modeling of experts' divergent prior beliefs for a sequential phase III clinical trial	2013	Consultation
Cope et al.[40]	Integrating expert opinion with clinical trial data to extrapolate long-term survival: a case study of CAR-T therapy for children and young adults with relapsed or refractory acute lymphoblastic leukemia	2019	SEE
Nadal et al.[41]	Clinical and economic impact of current ALK rearrangement testing in Spain compared with a hypothetical no-testing scenario	2021	Consultation
Klijn et al.[42]	What did time tell us? A comparison and retrospective validation of different survival extrapolation methods for immuno-oncologic therapy in advanced or metastatic renal cell carcinoma	2021	SEE
Konidaris et al.[43]	Assessing the value of cemiplimab for adults with advanced cutaneous squamous cell carcinoma: a cost-effectiveness analysis	2021	SEE
Ayers et al.[44]	Structured expert elicitation to inform long-term survival extrapolations using alternative parametric distributions: a case study of CAR T therapy for relapsed/ refractory multiple myeloma	2022	SEE
Federico Paly et al.[45]	Heterogeneity in survival with immune checkpoint inhibitors and its implications for survival extrapolations: a case study in advanced melanoma	2022	SEE
Ruggeri et al.[46]	Estimation model for healthcare costs and intensive care units access for covid-19 patients and evaluation of the effects of remdesivir in the portuguese context: hypothetical study	2022	Consultation
Willigers et al.[47]	The role of expert opinion in projecting long-term survival outcomes beyond the horizon of a clinical trial	2023	SEE
Gao et al.[48]	Temporal change in the remaining life expectancy in people who underwent percutaneous coronary intervention	2023	Consultation

SEE, structured expert elicitation

In Sections 5.1.1 to 5.1.5 we discuss the conduct and reporting of key methodological aspects of structured expert elicitation in the six respective studies. For context we then discuss other, more general, consultation methodologies employed in the remaining five studies in Section 5.1.6.

5.1.1 Reporting of identification and recruitment of experts

All six studies that employed structured expert elicitation reported the number of participating experts, which ranged from one expert (Klijn et al. (2021)[42]) to nine experts (Konidaris et al. (2021)[43]). Four studies (Ayers et al. (2022),[44] Cope et al. (2019),[40] Konidaris et al. (2021),[43] and Willigers et al. (2023)[47]) recruited more than five experts and provided some details of the expert recruitment process, which generally included a prior assessment of the experts' knowledge of the disease area and familiarity with the treatment(s) being evaluated. However, it is unclear whether existing contacts were relied upon or if experts were sourced more broadly from the disease area, for example, through 'cold calling'.

We note that the publication by Konidaris et al. (2021)[43] summarises work undertaken as part of NICE TA592,[49] and additional information regarding the elicitation is available in the appendices of the appraisal. However, these appendices are not publicly available, and therefore all information and conclusions presented in this review are based on the publicly available material only.

The selection and justification of experts were supported in two studies (Ayers et al. (2022)[44] and Cope et al. (2019)[40]) through an analysis of the experts' backgrounds. These studies provided detailed information, including the length of time the experts had been practising in the disease area, the number of patients they had treated with the disease, and the number of patients they had treated with the treatment(s) of interest.

Klijn et al. (2021),[42] Federico Paly et al. (2022)[45] and Willigers et al. (2023)[47] only reported the experts' subject area or specialism, and one study (Konidaris et al. (2021)[43]) described the backgrounds of the experts involved. Lastly, none of the six studies explicitly stated whether conflicts of interest were checked prior to the elicitation workshop.

Konidaris et al. (2021)[43] reported that the exercise was double-blinded, meaning neither the company nor the experts were disclosed to the other party. This, in practice, is unlikely to be feasible within the context of structured expert elicitation for HTA as it

is easily decipherable what company is testing the new intervention. Furthermore, it is likely that the company preparing the submission is likely to conduct the elicitation and thus will contact the experts, removing the possibility for blinding. The other studies using structured expert elicitation did not mention whether blinding was implemented.

5.1.2 Statistical training and briefing of experts

Only Ayers et al., Cope et al. and Willigers et al. (2023)[40,44,47] described any training being provided before experts made their judgements. The training provided by Willigers et al. (2023) included potential effects of cognitive biases.

5.1.3 Quantity of interest

The quantities of interest in the elicitation exercises varied considerably across the six studies (see Table 3). In the studies by Klijn et al. (2021)[42] and Federico Paly et al. (2022)[45], the mean lifetime survival for both arms was a key quantity of interest alongside the probability of survival at specific time points following the observed data. Ayers et al. (2022)[44] elicited the time point at which no patients would remain alive in addition to the probability of survival at multiple time points, providing experts with an anchor for their judgements.

Most studies did not aim to elicit conditional survival quantities. However, Ayers et al. (2022)[44] and Willigers et al. (2023)[47] adopted approaches to do so, albeit in different ways. Ayers et al. (2022)[44] explicitly aimed to elicit survival estimates at 5 and 10 years, conditional on the experts' prior judgements at earlier time points. In contrast, Willigers et al. (2023)[47] requested survival estimates at 20 years based on two pre-defined scenarios: survival at 10 years being either 40% or 70%. These scenarios were not dependent on the experts' previous estimates of survival at 10 years but were instead pre-defined as fixed quantities of interest.

All studies which used structured expert elicitation identified within this review aimed to elicit more than one quantity of interest: the survivor function at multiple time points. These were elicited independently; joint distributions were not elicited using multivariate elicitation methods.

Table 3: Summary of quantities of interest elicited.

Study	Time points that survival was elicited (years)	Conditional survival elicited (Y/N)	Time of 0% survival elicited (Y/N)	Mean lifetime survival (years) (Y/N)
Ayers et al.[44]	3, 5, 10	Y- conditional on the experts' judgements of prior time point	Y	N
Cope et al.[40]	2, 3, 4	N	N	N
Federico Paly et al.[45]	10, 20, 30*, 40*	N	N	Y
Klijn et al.[42]	10, 20	N	N	Y
Konidaris et al.[43]	2, 3, 4, 5	N	N	N
Willigers et al.[47]	10, 20	Y- survival at 20 years conditional on survival at 10 years equal to 40% and 70% respectively	N	N

** quantities listed in the attached materials distributed to experts, but only survival at 10 and 20 years are recorded within the main text of the article.*

5.1.4 Evidence dossier

Federico Paly et al. (2022)[45] included the "evidence pack" (dossier) in the published supplementary material. The information presented within the evidence dossier included a project background, objectives, and a thorough overview of the CheckMate 067 trial, which underpinned the quantities being elicited. Details of the CheckMate 067 trial included the study design, baseline characteristics, overall survival data (with supporting Kaplan-Meier curves at the 28-month and 5-year data cut-offs), objective response rate, and details of subsequent treatments.

Willigers et al. (2023)[47] also developed an evidence dossier (referred to as a "data book") and conducted a literature review in order to populate it, although they noted that the review was not systematic. Willigers et al. (2023) highlighted that parametric extrapolations of the survival data were included within the data book, acknowledging that this may have influenced experts when making their judgements.

Two studies (Klijn et al. (2021)[42] and Konidaris et al. (2021)[43]) did not mention a dossier being formulated or distributed before eliciting the quantities of interest.

The remaining studies did not provide detailed information on the content or the process of compiling the evidence dossier, instead generally stating that it included relevant evidence regarding the patient population and outcomes.[40,44]

5.1.5 Structured expert elicitation design and methodology

5.1.5.1 Protocol and format

Five of the six studies appeared to base elements of their elicitation design on the SHELF methodology.[40,42-45] Willigers et al. (2023)[47] used Cooke's classical method.

The studies by Cope et al. (2019)[40] and Federico Paly et al. (2022)[45] discussed modifications to the SHELF methodology and its design. Cope et al. (2019) adjusted the workshop design to facilitate the elicitation of survival at multiple time points, while Federico Paly et al. (2022) appeared to adapt the workshop format into a remote survey format. The authors acknowledged that despite their attempt to follow the SHELF methodology, there was minimal guidance at the time, to help inform the use of the framework for their defined quantities of interest.

Two of the studies (Ayers et al. (2021)[44] and Cope et al. (2019)[40]) conducted the elicitation workshop online. In contrast, Federico Paly et al. (2022)[45] and Willigers et al. (2023)[47] appeared to have collected the experts' individual judgements using surveys. Two studies (Klijn et al. (2021)[42] and Konidaris et al. (2021)[43]) did not report the format of the exercise.

5.1.5.2 Individual judgement

There is considerable variation in how experts' estimates were obtained. In Ayers et al. (2021)[44] and Cope et al. (2019)[40], experts were asked to provide the most likely value, along with lower and upper plausible limits. The most likely value was typically interpreted as the mean or mode, and the range of plausible limits was used to calculate the variance of a normal distribution centred on the mode, based on the assumption that the limits reflected the 1st and 99th percentiles.

Klijn et al. (2021),[42] Konidaris et al. (2021),[43] and Federico Paly et al. (2022)[45] elicited mean survival values alongside upper and lower limits, though these studies did not provide further statistical interpretation of the mean and limits.

In the study by Willigers et al. (2023),[47] experts were asked to provide the 10th, 50th (median), and 90th percentiles for the quantity of interest, which were described qualitatively as “low”, “medium”, and “high”. A description of the 10th percentile was also provided to experts, indicating that it represented a value for which the expert was 90% confident the true value would be greater than.

5.1.5.3 Aggregation methods

Willigers et al. (2023)[47] used mathematical aggregation, weighted by expert scores on a pre-defined, disease-specific set of calibration questions. The ten calibration questions, provided in the supplementary material, related to the disease of interest and relevant medical areas, were chosen to ensure that a single true answer was available and expected to be broadly known by disease specialists. The authors found that, in general, experts answered the calibration questions accurately. Both Ayers et al. (2021)[44] and Cope et al. (2019)[40] used behavioural aggregation to obtain a distribution reflecting the view of the “Rational Impartial Observer” (RIO).

It was unclear how the judgements of the two experts were aggregated in the elicitation conducted by Federico Paly et al. (2022),[45] but it appears that their judgements were used to inform a plausible range which was in turn used to select plausible extrapolations of the survival data. Similarly, the method of aggregation in Konidaris et al. (2021)[43] was not specified in the available documentation. Klijn et al. (2021)[42] elicited judgements from only one expert, so no aggregation was required, and the authors acknowledged this as a limitation of the elicitation exercise.

5.1.5.4 Reported roles within the exercise

Ayers et al. (2021)[44] and Cope et al. (2019)[40] described the role of a facilitator, noting that their responsibility was to guide experts through the web-based application used to collect individual judgements. Details of how the facilitators guided experts to decide

on the final output of the session, the distribution corresponding to the RIO, were not included within the publications.

5.1.5.5 Description of expert discussion

The general description of expert qualitative discussion within the structured expert elicitations was limited in most studies. Ayers et al. (2021)[44] and Cope et al. (2019)[40] provided additional discussion of the experts' reasoning in the supplementary materials of their publications, whereas the other studies did not include this information. Presumably, in the case of Willigers et al. (2023)[47] and potentially Federico Paly et al. (2022)[45], this was because expert judgements were collected via a remote survey and no subsequent expert discussion workshop appeared to have been organised, and thus experts did not have the opportunity to discuss their judgements fully.

In Sections 3.7.1 and 3.7.2 we discussed the consideration of the hazard during the group discussion. None of the six identified studies, appeared to discuss the hazard with experts.

5.1.6 Expert consultation

To add further context to our review, we include a brief description of identified studies that did not conduct a structured expert elicitation but instead consulted generally with subject experts. Five of the 11 identified relevant studies involved general consultation of experts, regarding long-term survival outcomes (see Table 2). Three of these studies were in oncology (Miksad et al. (2009),[38] Moatti et al. (2013),[39] and Nadal et al. (2021)[41]), one study relating to COVID-19 (Ruggeri et al. (2022)[46]) and one study relating to cardiovascular disease (Gao et al. (2023)[48]).

The number of experts consulted was not reported in the study by Gao et al. (2023),[48] but in the other studies, it ranged from three (Ruggeri et al. (2022)[46]) to 39 (Moatti et al. (2013)[39]) experts. The method of obtaining expert judgements was not detailed in two studies.[41,48] However, in the others, judgements were gathered either via a survey[38,39] or through a group interview.[46]

The studies using expert consultation identified in the review were largely validatory in nature, with minimal detail on how the consultation was conducted. Experts were consulted to assist with model selection for the extrapolation of survival outcomes within economic models,[41,48] to validate sources used in survival model selection[46] or to provide alternative survival estimates via a remote survey for scenario analyses.[38] Moatti et al. (2013) used expert judgements to form a prior distribution by mathematically aggregating the judgements.[39] However, during the collection of expert judgements, no individual expert uncertainty was captured (only central estimates of median overall survival) and therefore, the study was classified as expert consultation rather than a structured expert elicitation.

5.2 RESULTS: REVIEW OF NICE SUBMISSIONS

To supplement the review of the broader literature and assess the current use of structured expert elicitation for survival outcomes in technology appraisals, a review of recent NICE oncology submissions was performed. Thirty-five submissions were included in the review (see Table 4). We note that only the publicly available Document B for each submission was reviewed, due to the unavailability of the supporting appendices: this is a limitation of our review.

The majority of submissions involved some form of interaction with experts through advisory board meetings. Four submissions were identified as having conducted structured expert elicitation for survival outcomes: TA917,[50] TA954,[51] TA967,[52] and TA975[53] and the remaining TAs either employed more general expert consultation or did not involve external experts for long-term survival outcomes. Experts were generally consulted due to a lack of relevant long-term survival data in the literature, and thus experts were recruited to help assess the external validity of selected survival models used within the company's economic model.

Table 4: NICE technology appraisals included in the review.

TA Number	Title	Date Issued	SEE/Expert consultation
TA540	Pembrolizumab for treating relapsed or refractory classical Hodgkin lymphoma	03-Sep-18	Consultation

TA688	Selective internal radiation therapies for treating hepatocellular carcinoma	31-Mar-21	Consultation
TA737	Pembrolizumab with platinum- and fluoropyrimidine-based chemotherapy for untreated advanced oesophageal and gastro-oesophageal junction cancer	20-Oct-21	No external validation of survival extrapolation
TA917	Daratumumab with lenalidomide and dexamethasone for untreated multiple myeloma when a stem cell transplant is unsuitable	25-Oct-23	SEE
TA921	Ruxolitinib for treating polycythaemia vera	18-Oct-23	Consultation
TA927	Glofitamab for treating relapsed or refractory diffuse large B-cell lymphoma after 2 or more systemic treatments	17-Oct-23	Consultation
TA928	Cabozantinib for previously treated advanced differentiated thyroid cancer unsuitable for or refractory to radioactive iodine	01-Nov-23	Consultation
TA930	Lutetium-177 vipivotide tetraxetan for treating PSMA-positive hormone-relapsed metastatic prostate cancer after 2 or more treatments	15-Nov-23	Consultation
TA931	Zanubrutinib for treating chronic lymphocytic leukaemia	22-Nov-23	Consultation
TA944	Durvalumab with gemcitabine and cisplatin for treating unresectable or advanced biliary tract cancer	10-Jan-24	Consultation
TA946	Olaparib with bevacizumab for maintenance treatment of advanced high-grade epithelial ovarian, fallopian tube or primary peritoneal cancer	17-Jan-24	Consultation
TA947	Loncastuximab tesirine for treating relapsed or refractory diffuse large B-cell lymphoma and high-grade B-cell lymphoma after 2 or more systemic treatments	31-Jan-24	No external validation of survival extrapolation
TA948	Ivosidenib for treating advanced cholangiocarcinoma with an IDH1 R132 mutation after 1 or more systemic treatments	31-Jan-24	Consultation
TA950	Nivolumab–relatlimab for untreated unresectable or metastatic melanoma in people 12 years and over	07-Feb-24	Consultation
TA951	Olaparib with abiraterone for untreated hormone-relapsed metastatic prostate cancer	07-Feb-24	Consultation
TA952	Talazoparib for treating HER2-negative advanced breast cancer with germline BRCA mutations	21-Feb-24	Consultation
TA954	Epcoritamab for treating relapsed or refractory diffuse large B-cell lymphoma after 2 or more systemic treatments	06-Mar-24	SEE
TA957	Momelotinib for treating myelofibrosis-related splenomegaly or symptoms	20-Mar-24	Consultation*
TA962	Olaparib for maintenance treatment of BRCA mutation-positive advanced ovarian, fallopian tube or peritoneal cancer after response to first-line platinum-based chemotherapy	28-Mar-24	Consultation

TA963	Dostarlimab with platinum-based chemotherapy for treating advanced or recurrent endometrial cancer with high microsatellite instability or mismatch repair deficiency	03-Apr-24	Consultation
TA967	Pembrolizumab for treating relapsed or refractory classical Hodgkin lymphoma in people 3 years and over	01-May-24	SEE
TA970	Selinexor with dexamethasone for treating relapsed or refractory multiple myeloma after 4 or more treatments	08-May-24	Consultation
TA974	Selinexor with bortezomib and dexamethasone for previously treated multiple myeloma	15-May-24	Consultation
TA975	Tisagenlecleucel for treating relapsed or refractory B-cell acute lymphoblastic leukaemia in people 25 years and under	15-May-24	SEE
TA977	Dabrafenib with trametinib for treating BRAF V600E mutation-positive glioma in children and young people aged 1 year and over	29-May-24	Consultation
TA979	Ivosidenib with azacitidine for untreated acute myeloid leukaemia with an IDH1 R132 mutation	05-Jun-24	Consultation
TA983	Pembrolizumab with trastuzumab and chemotherapy for untreated locally advanced unresectable or metastatic HER2-positive gastric or gastro-oesophageal junction adenocarcinoma	12-Jun-24	Consultation
TA985	Selective internal radiation therapy with QuiremSpheres for treating unresectable advanced hepatocellular carcinoma	03-Jul-24	No external validation of survival extrapolation
TA992	Trastuzumab deruxtecan for treating HER2-low metastatic or unresectable breast cancer after chemotherapy	29-Jul-24	Consultation
TA995	Relugolix for treating hormone-sensitive prostate cancer	14-Aug-24	No external validation of survival extrapolation
TA997	Pembrolizumab with platinum- and fluoropyrimidine-based chemotherapy for untreated advanced HER2-negative gastric or gastro-oesophageal junction adenocarcinoma	29-Aug-24	Consultation
TA1001	Zanubrutinib for treating marginal zone lymphoma after anti-CD20-based treatment	04-Sep-24	Consultation
TA1005	Futibatinib for previously treated advanced cholangiocarcinoma with FGFR2 fusion or rearrangement	11-Sep-24	Consultation
TA1007	Rucaparib for maintenance treatment of relapsed platinum-sensitive ovarian, fallopian tube or peritoneal cancer	17-Sep-24	No external validation of survival extrapolation
TA1008	Trifluridine–tipiracil with bevacizumab for treating metastatic colorectal cancer after 2 systemic treatments	25-Sep-24	Consultation

** described as "expert elicitation" but no hallmarks of structured expert elicitation.*
TA, technology appraisal; SEE, structured expert elicitation

We did not see any discussion in the available documentation regarding the choice of structured expert elicitation versus expert consultation. In Sections 5.2.1 to 5.2.5 we discuss key elements of structured expert elicitation methodology and comment on the conduct in the four TAs that employed this methodology. For context we discuss other, more general, consultation methodologies employed in the remaining appraisals in Section 5.2.6.

5.2.1 Reporting of identification and recruitment of experts

Out of the four identified appraisals that used structured expert elicitation, between three and ten clinicians were recruited.[50-53] In TA917,[50] a detailed breakdown of the experts' backgrounds was not provided, it was noted that eight clinicians were from clinics in England, and two were based in Scotland. The involvement of experts appeared to vary, with two clinicians providing initial feedback, five offering feedback on survival extrapolations after an advisory board meeting, and one expert refraining from providing any feedback.

In TA975,[53] three UK clinicians with experience in the treatments (both the new and existing therapies) of the disease were selected for the elicitation. The two other appraisals (TA954[51] and TA967[52]) did not provide further information on the experts or their backgrounds; however, this information may have been available in the appendices, which were not available on the NICE TA websites.

In TA967,[52] experts were assessed for prior involvement with other company activities, and any such activities were logged to check for potential conflicts of interest. Similarly, in TA975,[53] experts were asked to declare any conflicts of interest. However, TA917[50] and TA954[51] did not comment on the potential conflicts of interest of participating experts in Document B.

5.2.2 Statistical training and briefing of experts

Only TA967[52] described the use of any training resources, namely the STEER training resources, which closely align with the MRC reference protocol.[6] In TA967, it was also stated that a training question was provided to experts to complete remotely as to

familiarise themselves with the user interface when making their judgements. Overall, minimal information was found in Document B of each of the remaining three submissions regarding the training and briefing of experts.

5.2.3 Quantity of interest

Generally, the submissions defined quantities of interest at one or more time points following the data cut-off period of the pivotal trial. In TA917[50] TA954[51] and TA975[53] the quantities of interest were defined as the percentages of patients alive at multiple time points. These time points ranged from 1 year to 20 years and largely reflected disease-specific timescales. In TA967,[52] although similar, the quantity of interest was phrased as to focus on whole numbers of patients rather than survival proportions; for instance the number of patients from an original 100-person cohort (representative of the target population) who would still be alive 4 years after the initiation of standard care.

TA917[50] and TA975[53] elicited expert judgement on quantities relating to both the intervention and control arms. However, TA954[51] only obtained expert judgement for quantities relating to the intervention arm and TA967[52] only obtained expert judgement for the control arm. This approach reflected the companies' methodologies in that respective survival curves were obtained by applying a hazard ratio to either the intervention or control survival curve as opposed to modelling the intervention and comparator arms independently. None of these appraisals elicited conditional survival proportions as the quantity of interest.

5.2.4 Evidence dossier

In all elicitation exercises, “pre-read” materials were sent to experts ahead of the elicitation. In TA967,[52] a systematic literature review was used to populate the evidence dossier, supplemented by literature sourced from hand-searching. However, details of the dossier contents were not provided. Details of the pre-read material was also not provided for TA917[50] but it was stated to have been reviewed by clinicians, this review occurred during the meeting, rather than in advance of the elicitation exercise.

The dossier for TA954[51] was reported to include materials on disease background, pivotal trial data, and economic modelling approaches. The inclusion of "modelling approaches" in the dossier raises some uncertainty about whether experts were shown survival curve extrapolations before making judgements, or whether these approaches pertained to other quantities discussed during the meeting.

In TA975,[53] the dossier was stated to include similar summaries of current treatment options, an overview of the pivotal trial using the latest data cut-off, and Kaplan-Meier curves, suggesting that model extrapolations were not presented to experts at this stage, though they were discussed in a subsequent meeting.

5.2.5 Structured expert elicitation design and methodology

5.2.5.1 Protocol and format

TA917[50] and TA967[52] stated that the MRC protocol[6] was used as the basis for the elicitation of survival outcomes. TA954[51] and TA975[53] did not explicitly cite an elicitation protocol as a basis for the structured expert elicitations conducted.

In TA917,[50] an online advisory board meeting was held in which to conduct the elicitation exercise. Whereas, in TA967,[52] the STEER resources were used to enable the collection of both quantitative and qualitative responses from experts via a remote survey.

5.2.5.2 Individual judgement

In TA917 and TA954,[50,51] experts appeared to be required to make their individual judgements during the group workshop, however it was not entirely clear whether individual experts could discuss their judgements during this stage. For each of the quantities of interest, experts were asked to provide most likely values and associated plausible limits. Aside from the quantities of interest, in TA917, experts were also shown company selected survival extrapolations and asked to rank the extrapolations of progression-free survival and overall survival. From Document B, it is not entirely clear what order this was conducted in.

In contrast, TA967[52] and TA975[53] asked all experts to complete their individual judgements remotely via a survey. As mentioned previously, in TA967 experts used the platforms provided within the STEER resources which included the fixed interval, chips and bins method using the Excel template. In addition to the quantitative expert judgements collected remotely, expert rationale was captured via a free-text box; these rationales were subsequently made available in an appendix of the submission. No details of how expert's individual judgements were obtained were provided for TA975.

5.2.5.3 Aggregation methods

In TA967,[52] it appears that the judgements were mathematically aggregated, though it is unclear how the weighting was applied (i.e., equal weighting or performance-based). Although the expert judgements were collected via a remote survey for TA967, there was subsequent discussion of the pooled distribution which took place at the follow-up advisory board meeting. However, experts were not able to refine or alter their judgements or the group distribution in light of the discussion. It was not described in full how expert judgements in TA917, TA954 or TA975 were aggregated.[50,51,53]

5.2.5.4 Reported roles within the exercise

None of the appraisals explicitly mention the role of the facilitator but it is clear that in all four of the appraisals, there was a degree of expert discussion at some point within the elicitation process, potentially during a subsequent advisory board meeting, which will have needed to have been chaired. In the appraisals where experts made their judgements within the group session, it was not clear whether the chair of the discussions actively assisted experts when making their judgements beyond standard clarification.

5.2.5.5 Description of expert discussion

In TA967,[52] it was highlighted how the discussion element of the elicitation was useful in explaining inter-expert variability. However, it was noted that because experts could not change their judgements following the discussion, this could not subsequently inform the mathematically aggregated distribution. In TA975,[53] experts were asked to

assess model fits in light of their individual judgements and experts' judgements were discussed within the advisory board meeting, but it is unclear what the aggregation method was or how this discussion was used in the decision making. The management of different plausible ranges was not discussed in the main text of the submission.

5.2.6 Expert consultation

Of the remaining 32 appraisals, 29 involved consultations with clinical experts on the external clinical validity of survival extrapolations during advisory board meetings (see Table 4). Rather than obtaining quantitative estimates of survival at specific time points, experts were typically presented with model extrapolations or landmark survival predictions and asked to assess their plausibility. These discussions with experts were subsequently used to justify the selection of specific models for the economic analysis.

Generally, this broader consultation with experts is considered less time-intensive and is often summarised in company submissions with a single sentence, accrediting experts with assessing the external validity of the chosen survival extrapolation. However, some of these company submissions describe the preparation and conduct of the consultation in a manner similar to those that conducted structured expert elicitations. For instance, some appraisals report the number of experts involved and provide a high-level description of their expertise. In Table 5, we summarise the appraisals that include descriptions of various methodological choices which mirror choices made within structured approaches.

Table 5: Description of methodological aspects of expert consultation within NICE technology appraisals.

Methodological component	Appraisals
Details of expert background or recruitment	TA921, TA944, TA947, TA951, TA962, TA963*, TA970, TA974, TA977, TA983, TA992, TA977
Definition of quantities of interest (i.e., not simple model validation)	TA921, TA950, TA983, TA992, TA997, TA1008

Preparation of an evidence dossier/interview guide/information pack	TA944

[*] *declaration of conflicts also obtained and included in appendix of submission.*

From the selection of company submissions that consulted with experts, we highlighted TA688[54] which employed a multi-stage process to obtain expert opinion, utilising a remote survey prior to the advisory board meeting. This meeting then provided an opportunity for experts to discuss their reasoning and judgements in greater detail within a group setting. While this discussion element shares several similarities with more structured exercises, the objective of these discussions and whether a consensus was reached was unclear and thus it was labelled as an expert consultation.

Many appraisals sought general expert input to assist with model fitting and the assessment of extrapolation plausibility. However, in six company submissions (TA921,[55] TA950,[56] TA983,[57] TA992,[58] TA997,[59] TA1008[60]), experts were also asked to provide numerical estimates for defined quantities of interest. These estimates were typically obtained during advisory board meetings, with limited description provided on if/how experts were facilitated in making these judgements. Furthermore, there was generally even less information on the steps taken to minimise potential biases, such as the conduct of the expert selection process or the provision of pre-read materials.

In terms of the consideration of the hazard, we identified TA1001[61] which explicitly stated that experts were involved in qualitative discussions about hazard trends in order to assist with model selection. We did not identify any other studies (consultation or structured expert elicitations) which employed this approach and considered expert opinion on the hazard.

5.3 REVIEW DISCUSSION

It is evident from our review of the broader literature and the review of NICE submissions in the past year, that expert opinion is being used to help provide estimates of long-term survival outcomes, and assess external validity, especially in areas where there is often limited long-term data from clinical trials, such as oncology.

This appears to be a key feature of appraisals, ensuring that the modelling of long-term survival within economic models is clinically plausible and defensible.

From our review, we make the following observations.

- **Structured versus unstructured methods.** Only four out of 36 appraisals we reviewed used structured expert elicitation when incorporating expert opinion. Greater uptake of structured expert elicitation would improve the methodological rigour, in particular ensuring that expert uncertainty is quantified and reported. It was not always clear in the NICE submissions why structured expert elicitation or general consultation was favoured when obtaining expert opinion on long-term survival outcomes. As we see evidence of both approaches within submissions, it is likely that this difference in approach may not relate directly with resource availability but instead familiarity with, or knowledge of, the structured expert elicitation methodology.
- **Choice of protocol.** We found that different protocols have been applied in for eliciting long-term survival outcomes in the broader literature and NICE submissions. Notwithstanding concerns about elicitation via remote surveys, this supports our assessment in Section 4.3 about the feasibility of structured expert elicitation for survival extrapolation based on any of the standard protocols described in Appendix A.1.
- **Choice of quantities for elicitation.** Multiple studies in the broader literature and NICE submissions elicited survival proportions at multiple time points. Whilst this in practice has the potential to be a viable approach, there was generally no formal consideration of dependency between time points, and what additional knowledge can be drawn out through elicitation for additional time points, as discussed in Section 3.5.3.
- **Incorporating qualitative knowledge about hazards.** Within our tailored guidance for elicitation of survival outcomes in Section 3.7, we suggest explicit discussion of the hazard to enable checking of internal consistency. Within our review, we found no evidence of discussion of the hazard when eliciting expert judgement on the survival outcomes.

It is more difficult to comment on the use of evidence dossiers, and the reporting of structured expert elicitation exercises;[62] a limitation of our review of NICE submissions is that we were only able to use publicly available documents. It is also difficult to assess the extent of any discussion between experts, or scrutiny of each other's judgements.

From both reviews, it is evident that there is considerable scope to improve on current practice in using expert opinion for survival extrapolation. Firstly, there is a need to increase the uptake of structured expert elicitation methods. Secondly, there is a need for protocols that are tailored to the specific methodological aspects of expert elicitation for survival extrapolation. This will help ensure that future elicitation of long-term survival outcomes provides credible, accurate, consistent, and transparent expert judgement to support decision-making. Additionally, the formulation of this TSD and the recommendations presented in Section 6, will help to support the use of structured expert elicitations approaches more broadly within NICE submissions as opposed to general expert consultation.

6 RECOMMENDATIONS AND DISCUSSION

6.1 RECOMMENDATIONS

The company submitting the appraisal should be responsible for conducting structured expert elicitation for long-term survival outcomes. Below, we set out recommendations that can be used by both companies when planning and conducting their elicitation exercises, and by NICE External Assessment Groups (EAGs) when assessing the validity and credibility of a structured expert elicitation.

What to obtain from experts

1. As a minimum, we should obtain a probability distribution to quantify expert uncertainty about the value of the survivor function for one time point. For a single point $S(T)$, guidance on the choice of T to achieve maximum value from the elicitation exercise is given in Section 3.5.
2. Experts should be asked for qualitative opinions about how the hazard may change over the extrapolation period (Section 3.7.1). This may assist with both choosing between survival models (Section 3.5.1.1), and with checking for internal consistency in an expert's judgements (Section 3.7.2).
3. Distributions for survivor function values at additional time points may be elicited. In this case, commentary should be provided regarding whether the experts are drawing on additional substantive knowledge, beyond that elicited following Recommendations 1 and 2, or whether the judgements are based primarily on expectations of a 'smooth-looking' survival curve. Discussion of this is given in Section 3.5.3. If the aim is to validate a parametric model, multivariate elicitation methods may be required to assess joint probabilities regarding combinations of values for $S(T_1), \ldots, S(T_n)$.
4. The target population should be specified at the point of developing the economic model; its definition should not be altered in the elicitation exercise and should be stated clearly in the quantity of interest definition.

Recruiting experts

5. General recommendations for recruiting experts are given in Section 2.7. Here, we emphasise that the recruitment pool of experts should not exclude those with knowledge of the trial data (either an expert involved in the trial, or another

expert who has seen a report of the trial); the elicitation exercise should not be designed under the assumption that the experts will not have seen the trial data.

6. Experts who have seen model-based extrapolations should be excluded, unless this makes recruiting an appropriate expert panel infeasible. Further recommendations related to this point are Recommendations 9, 15 and 18.
7. We do not recommend recruiting fewer than three experts, but otherwise think it is difficult to justify a minimum acceptable number. This is discussed further in Section 2.7.

Preparation of the evidence dossier

8. Relevant evidence should be compiled in advance of the elicitation activity and circulated to the participating experts for comment and additions. Advice on the content is given in Section 3.3. Here, we emphasise that the dossier should include:
    a. a Kaplan-Meier plot of the survival data for the trial arm(s)/subgroup(s) of interest, including confidence intervals to indicate uncertainty in the Kaplan-Meier estimate and the number at risks at various time points;
    b. reporting of the corresponding empirical hazard;
    c. tabulated survival function data for the general population, for a cohort matched to the trial arm(s)/subgroup population(s).
9. The evidence dossier should *not* include any model extrapolations. This is to avoid anchoring effects, as discussed in Section 2.3.

Training of experts

10. Experts (once recruited) must be trained in making probability judgements. Recommended content for the training is given in Section 3.4.
11. Additional training should be provided in understanding both survivor and hazard functions.
12. We recommend conducting a practice elicitation exercise that involves survival extrapolation.

Elicitation protocol and conduct of the elicitation exercise

13. The basis of the elicitation protocol should be a standard elicitation protocol for structured expert elicitation, such as Cooke's classical method; Delphi; IDEA;

MRC reference protocol; SHELF. Any of these can be used appropriately, though we would only recommend Delphi methods if conducted via one-to-one interviews.

14. Following the choice of protocol, we recommend modifications to the protocol so that additional steps are specified in advance, in particular, how qualitative judgements about the hazard will be obtained, and how experts might consider these when making quantitative judgements about the survivor function. This is illustrated in Section 4.
15. The protocol may involve presenting model extrapolations to the experts, but this should not happen until each expert has provided their own quantitative judgements about the survivor function value(s) of interest.
16. We give recommendations for providing quality assurance in Section 2.9. We emphasise the importance of critically evaluating expert judgements, either through peer assessment within the elicitation protocol, or performance weighting if Cooke's method is used.

Incorporating opinions about the hazard

17. We recommend the use of a "hazard checklist", discussed in Section 3.7.1, to encourage experts to consider factors related to the patients, disease and treatment that might increase or decrease the hazard over the extrapolation period.
18. After qualitative judgements from each expert have been recorded, an option is to present the experts with extrapolations corresponding to scenarios regarding the hazard, as discussed in Section 3.7.2. This would be to provide the experts with feedback on their judgements, and the choice of whether to incorporate this step should depend on the feasibility of providing useful feedback. As stated in Recommendation 15, experts should only be presented with any such extrapolations after they have provided their own quantitative judgements about the survivor function value(s) of interest.

Using the results to inform the choice of survivor function in the economic model

19. We have described a procedure in Section 3.5.1.1 for using elicited judgements (quantitative and qualitative) to choose a survivor function. In summary, both

data and expert judgement can be used to identify candidate survival models, but there still may be multiple models consistent with both data and expert judgement. In this case, we recommend cost-effectiveness results are presented for each consistent model.

20. Full consideration of the experts' uncertainty about $S(T)$ should be considered, using the elicited distribution. A point estimate extracted from this distribution (such as a median value) should *not* be interpreted as "*the* estimate of $S(T)$ given by the experts", ignoring other values of $S(T)$ supported by the experts' distribution.
21. An alternative approach is to synthesise expert opinion and data within a Bayesian framework. In this case, we recommend incorporating a check to ensure there is no double-counting of data, as discussed in Section 3.6.

Reporting of the structured expert elicitation

22. The reporting of the elicitation should follow the guidelines within the chosen protocol, with additional reporting of the hazard discussion as appropriate. Both individual and aggregated judgements should be reported for all elicitations to ensure transparency of the exercise and selected protocol.
23. In addition, details of the design and conduct of the elicitation exercise should be described in full and included in the company's submission. The methods of expert elicitation should be reported to a standard that would enable the principles of the exercise to be independently replicated. At a minimum, we recommend including a description and summary of Recommendations 1–21, as well as a full description of the data provided to experts as part of the evidence dossier.
24. As per Recommendation 23, we advise full disclosure of the expert recruitment process and emphasise that, for transparency purposes, experts' names, expertise, and conflicts of interest should be reported. Individual expert judgements or qualitative statements should be anonymised.

6.2 PLANNING AND TIMELINES

Based on our experience of implementing structured expert elicitation for survival extrapolation, we give some advice regarding the planning of an exercise and appropriate timelines.

We recommend that companies plan well in advance to ensure successful recruitment of experts who meet the recommendations outlined above. This process can begin after the conceptualisation of the economic model has been finalised and the quantities of interest for the structured expert elicitation have been defined. It is likely that the evidence dossier can be constructed after the systematic literature review of the clinical evidence, which is standard practice as part of NICE submissions, and supplemented with further evidence obtained by hand searching. The evidence dossier, therefore, although requiring a comprehensive review of relevant literature, is unlikely to require significant time and/or resources within the setting of a NICE submission.

Based on our experience to date in face-to-face (either in-person or online) workshops; presenting the background, motivation and training of experts typically takes around 2 hours, including a practice exercise. We would then elicit judgements for all the quantities of interest, independently for each expert, allowing around 15 minutes per quantity. For the remaining facilitated discussion (including hazard trends and scenario testing), we would allow around 1 hour for the first quantity; subsequent quantities may need less time, as the hazard discussion may be common to all quantities.

In cases where it is difficult to gather all experts together for a single workshop, training and individual elicitation could be conducted separately from the group discussion and aggregation stage, as suggested in Section 2.8. This means that the whole workshop can be split into two sessions, making it more manageable for participants and scheduling. We also suggest allowing approximately one to two weeks to finalise the elicitation report to allow experts sufficient time to review it.

6.3 FUTURE RESEARCH

We expect the methodology discussed in this report to evolve in two directions. Firstly, as users gain practical experience of conducting elicitations, we expect recommendations for implementation to evolve accordingly. Protocols and software may also adapt over time. It is useful for such practical experience to be shared; this includes organisational aspects such as the timescale over which an elicitation exercise is designed and concluded.

In our literature review, we found only four technology appraisals that used structured expert elicitation for survival extrapolation, with most appraisals only using experts to validate the choice of survival model. This implies little accumulated experience in *using* the output of a structured expert elicitation for survival extrapolation to support decision-making. Full incorporation within a probabilistic sensitivity analysis would be desirable, but there is a technical challenge of synthesising data and expert judgement for inference about the full survivor function, given there are risks of double-counting data. Methodological developments in this problem of synthesis and inference may then feed back into new developments in the elicitation methodology, so that the process of eliciting and then utilising expert judgement for decision-making is fully aligned.

Nevertheless, structured expert elicitation can provide a more robust source of information for decision-makers regarding the long-term survival of populations. This is likely to be preferable to the presentation of several potentially plausible scenarios based on experts' comments derived from general consultation. In the authors' opinion, the use of structured approaches will provide decision-makers with greater confidence when selecting plausible scenarios and reduce the need to assess numerous scenarios with varying levels of plausibility and credibility.

7 REFERENCES

1. National Institute for Health and Care Excellence. NICE health technology evaluations: the manual. 2022.
2. Aspinall W, Cooke R. Quantifying scientific uncertainty from expert judgement elicitation, in. *Risk and uncertainty assessment for natural hazards*. 2013:64.
3. Cooke RM. *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press, USA; 1991.
4. Dalkey N, Helmer O. An experimental application of the Delphi method to the use of experts. *Management science*. 1963;9(3):458-467.
5. Hemming V, Walshe TV, Hanea AM, Fidler F, Burgman MA. Eliciting improved quantitative judgements using the IDEA protocol: A case study in natural resource management. *PLoS One*. 2018;13(6):e0198468. doi:10.1371/journal.pone.0198468
6. Bojke L, Soares M, Claxton K, et al. Developing a reference protocol for structured expert elicitation in health-care decision-making: a mixed-methods study. *Health Technology Assessment (Winchester, England)*. 2021;25(37):1.
7. Oakley JE, O'Hagan A. SHELF: the Sheffield Elicitation Framework (version 4). https://shelf.sites.sheffield.ac.uk/
8. Latimer N. NICE DSU technical support document 14: survival analysis for economic evaluations alongside clinical trials-extrapolation with patient-level data. *Report by the Decision Support Unit*. 2011;
9. Rutherford M, Lambert P, Sweeting M, Pennington R, Crowther MJ, Abrams KR. NICE DSU technical support document 21: Flexible methods for survival analysis. *Leicester, UK: Department of Health Sciences, University of Leicester*. 2020:1-97.
10. Therneau T. A package for survival analysis in S. *R package version*. 2015;2(7):2014.
11. Garthwaite PH, Kadane JB, O'Hagan A. Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association*. 2005/06/01 2005;100(470):680-701. doi:10.1198/016214505000000105
12. O'Hagan A, Buck CE, Daneshkhah A, et al. Uncertain judgements: eliciting experts' probabilities. 2006;
13. Oakley J. Eliciting univariate probability distributions. *Rethinking risk measurement and reporting*. 2010;1:155-177.
14. Authority EFS. Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal*. 2014;12(6):3734.
15. Morgan MG. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National academy of Sciences*. 2014;111(20):7176-7184.
16. Dias LC, Morton A, Quigley J. Elicitation: State of the art and science. *Elicitation: The science and art of structuring judgement*. 2018:1-14.
17. Soares M, Colson A, Bojke L, et al. Recommendations on the use of structured expert elicitation protocols for healthcare decision making: a good practices report of an ISPOR task force. *Value in Health*. 2024;27(11):1469-1478.
18. Kahneman D. Thinking, fast and slow. *Farrar, Straus and Giroux*. 2011;
19. Wilson KJ. An investigation of dependence in expert judgement studies with multiple experts. *International Journal of Forecasting*. 2017;33(1):325-336.
20. Kynn M. The 'heuristics and biases' bias in expert elicitation. *Journal of the Royal Statistical Society Series A: Statistics in Society*. 2008;171(1):239-264.
21. Gigerenzer G, Hoffrage U, Kleinbölting H. Probabilistic mental models: a Brunswikian theory of confidence. *Psychological review*. 1991;98(4):506.

22. Oakley JE. SHELF: Tools to Support the Sheffield Elicitation Framework. R package version 1.12.0. https://github.com/OakleyJ/SHELF
23. Winkler RL. The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical association*. 1967;62(319):776-800.
24. Genest C, Zidek JV. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*. 1986;1(1):114-135.
25. Mannes AE, Soll JB, Larrick RP. The wisdom of select crowds. *Journal of personality and social psychology*. 2014;107(2):276.
26. Budescu DV, Chen E. Identifying expertise to extract the wisdom of crowds. *Management science*. 2015;61(2):267-280.
27. Bolger F. The Selection of Experts for (Probabilistic) Expert Knowledge Elicitation. In: Dias LC, Morton A, Quigley J, eds. *Elicitation: The Science and Art of Structuring Judgement*. Springer International Publishing; 2018:393-443.
28. Turner RM, Spiegelhalter DJ, Smith GC, Thompson SG. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society Series A: Statistics in Society*. 2009;172(1):21-47.
29. Jackson CH. flexsurv: a platform for parametric survival modeling in R. *Journal of statistical software*. 2016;70
30. Ren S, Oakley JE. Assurance calculations for planning clinical trials with time-to-event outcomes. *Statistics in Medicine*. 2014;33(1):31-45.
31. Garthwaite PH, Al-Awadhi SA, Elfadaly FG, Jenkinson DJ. Prior distribution elicitation for generalized linear and piecewise-linear models. *Journal of Applied Statistics*. 2013;40(1):59-75.
32. Che Z, Green N, Baio G. Blended survival curves: a new approach to extrapolation for time-to-event outcomes from clinical trials in health technology assessment. *Medical Decision Making*. 2023;43(3):299-310.
33. Guyot P, Ades AE, Beasley M, Lueza B, Pignon J-P, Welton NJ. Extrapolation of survival curves from cancer trials using external information. *Medical Decision Making*. 2017;37(4):353-366.
34. Jackson CH. survextrap: a package for flexible and transparent survival extrapolation. *BMC Medical Research Methodology*. 2023;23(1):282.
35. Salsbury JA, Oakley JE, Julious SA, Hampson LV. Assurance methods for designing a clinical trial with a delayed treatment effect. *Statistics in Medicine*. 2024;
36. Williams CJ, Wilson KJ, Wilson N. A comparison of prior elicitation aggregation using the classical method and SHELF. *Journal of the Royal Statistical Society Series A: Statistics in Society*. 2021;184(3):920-940.
37. Forsyth J, Uttley L, Wong R, Ren K. Expert elicitation for survival outcomes in healthcare decision making. 2024/9/18 2024;doi:10.17605/OSF.IO/HJFCR
38. Miksad RA, Gonen M, Lynch TJ, Roberts TG, Jr. Interpreting trial results in light of conflicting evidence: a Bayesian analysis of adjuvant chemotherapy for non-small-cell lung cancer. Research Support, N.I.H., Extramural
Research Support, Non-U.S. Gov't. *J Clin Oncol*. May 1 2009;27(13):2245-52. doi:10.1200/JCO.2008.16.2586
39. Moatti M, Zohar S, Facon T, Moreau P, Mary JY, Chevret S. Modeling of experts' divergent prior beliefs for a sequential phase III clinical trial. *Clin Trials*. Aug 2013;10(4):505-14. doi:10.1177/1740774513493528
40. Cope S, Ayers D, Zhang J, Batt K, Jansen JP. Integrating expert opinion with clinical trial data to extrapolate long-term survival: a case study of CAR-T therapy for children and young adults with relapsed or refractory acute lymphoblastic leukemia.

Research Support, Non-U.S. Gov't. *BMC Med Res Methodol*. Sep 2 2019;19(1):182. doi:10.1186/s12874-019-0823-8
41. Nadal E, Bautista D, Cabezon-Gutierrez L, et al. Clinical and economic impact of current ALK rearrangement testing in Spain compared with a hypothetical no-testing scenario. *BMC Cancer*. Jun 10 2021;21(1):689. doi:10.1186/s12885-021-08407-1
42. Klijn SL, Fenwick E, Kroep S, et al. What Did Time Tell Us? A Comparison and Retrospective Validation of Different Survival Extrapolation Methods for Immuno-Oncologic Therapy in Advanced or Metastatic Renal Cell Carcinoma. Research Support, Non-U.S. Gov't. *Pharmacoeconomics*. Mar 2021;39(3):345-356. doi:10.1007/s40273-020-00989-1
43. Konidaris G, Paul E, Kuznik A, et al. Assessing the Value of Cemiplimab for Adults With Advanced Cutaneous Squamous Cell Carcinoma: A Cost-Effectiveness Analysis. Research Support, Non-U.S. Gov't. *Value Health*. Mar 2021;24(3):377-387. doi:10.1016/j.jval.2020.09.014
44. Ayers D, Cope S, Towle K, Mojebi A, Marshall T, Dhanda D. Structured expert elicitation to inform long-term survival extrapolations using alternative parametric distributions: a case study of CAR T therapy for relapsed/ refractory multiple myeloma. Research Support, Non-U.S. Gov't. *BMC Med Res Methodol*. Oct 15 2022;22(1):272. doi:10.1186/s12874-022-01745-z
45. Federico Paly V, Kurt M, Zhang L, et al. Heterogeneity in Survival with Immune Checkpoint Inhibitors and Its Implications for Survival Extrapolations: A Case Study in Advanced Melanoma. *MDM Policy Pract*. Jan-Jun 2022;7(1):23814683221089659. doi:10.1177/23814683221089659
46. Ruggeri M, Signorini A, Caravaggio S, et al. Estimation Model for Healthcare Costs and Intensive Care Units Access for Covid-19 Patients and Evaluation of the Effects of Remdesivir in the Portuguese Context: Hypothetical Study. *Clin Drug Investig*. Apr 2022;42(4):345-354. doi:10.1007/s40261-022-01128-8
47. Willigers BJA, Ouwens M, Briggs A, et al. The Role of Expert Opinion in Projecting Long-Term Survival Outcomes Beyond the Horizon of a Clinical Trial. Research Support, Non-U.S. Gov't. *Adv Ther*. Jun 2023;40(6):2741-2751. doi:10.1007/s12325-023-02503-3
48. Gao L, Nguyen D, Moodie M, et al. Temporal Change in the Remaining Life Expectancy in People Who Underwent Percutaneous Coronary Intervention. *Am J Cardiol*. Jan 15 2023;187:154-161. doi:10.1016/j.amjcard.2022.10.045
49. Cemiplimab for treating metastatic or locally advanced cutaneous squamous cell carcinoma NICE technology appraisal guidance 592. 2019;
50. Daratumumab with lenalidomide and dexamethasone for untreated multiple myeloma when a stem cell transplant is unsuitable NICE technology appraisal guidance 917. 2023;
51. Epcoritamab for treating relapsed or refractory diffuse large B-cell lymphoma after 2 or more systemic treatments NICE technology appraisal guidance 954. 2024;
52. Pembrolizumab for treating relapsed or refractory classical Hodgkin lymphoma in people 3 years and over NICE technology appraisal guidance 967. 2024;
53. Tisagenlecleucel for treating relapsed or refractory B-cell acute lymphoblastic leukaemia in people 25 years and under NICE technology appraisal guidance 975. 2024;
54. Selective internal radiation therapies for treating hepatocellular carcinoma NICE technology appraisal guidance 688. 2021;
55. Ruxolitinib for treating polycythaemia vera NICE technology appraisal guidance 921. 2023;

56. Nivolumab–relatlimab for untreated unresectable or metastatic melanoma in people 12 years and over NICE technology appraisal guidance 950. 2024;
57. Pembrolizumab with trastuzumab and chemotherapy for untreated locally advanced unresectable or metastatic HER2-positive gastric or gastro-oesophageal junction adenocarcinoma NICE technology appraisal guidance 983. 2024;
58. Trastuzumab deruxtecan for treating HER2-low metastatic or unresectable breast cancer after chemotherapy NICE technology appraisal guidance 992. 2024;
59. Pembrolizumab with platinum- and fluoropyrimidine-based chemotherapy for untreated advanced HER2-negative gastric or gastro-oesophageal junction adenocarcinoma NICE technology appraisal guidance 997. 2024;
60. Trifluridine–tipiracil with bevacizumab for treating metastatic colorectal cancer after 2 systemic treatments NICE technology appraisal guidance 1008. 2024;
61. Zanubrutinib for treating marginal zone lymphoma after anti-CD20-based treatment NICE technology appraisal guidance 1001. 2024;
62. Iglesias CP, Thompson A, Rogowski WH, Payne K. Reporting Guidelines for the Use of Expert Judgement in Model-Based Economic Evaluations. *Pharmacoeconomics*. 2016/11/01 2016;34(11):1161-1172. doi:10.1007/s40273-016-0425-9
63. Cooke RM, Goossens LL. TU Delft expert judgment data base. *Reliability Engineering & System Safety*. 2008;93(5):657-674.
64. Rowe G, Wright G. The Delphi technique as a forecasting tool: issues and analysis. *International journal of forecasting*. 1999;15(4):353-375.
65. Haddaway NR, Grainger MJ, Gray CT. Citationchaser: A tool for transparent and efficient forward and backward citation chasing in systematic searching. *Research Synthesis Methods*. 2022;13(4):533-545.
66. Cooke A, Smith D, Booth A. Beyond PICO: the SPIDER tool for qualitative evidence synthesis. *Qualitative health research*. 2012;22(10):1435-1443.

APPENDICES

APPENDIX A

A.1 EXISTING ELICITATION PROTOCOLS

The ISPOR Task Force on structured expert elicitation for health-care decision-making has identified five of the most frequently used protocols, and we briefly summarise them below.[17] A comparison of Cooke's Classical Method, Delphi and SHELF is given in EFSA (2014).[14] Some experimental work comparing Cooke's Classical Method, SHELF, and equal-weighted linear pooling is reported in Williams et al. (2021).[36]

A.1.1 Cooke's Classical Method

Cooke's method is a performance-based protocol.[3] Experts are assessed using calibration questions (i.e., problems with known answers) to evaluate the accuracy and informativeness of their responses. These performance metrics are used to assign weights to each expert, determining their contribution to the aggregated result in a weighted linear pool. The experts will typically interact, but make their judgements independently. An extensive database of elicitation exercises conducted with this method is also available on the TU Delft expert judgement data base.[63]

A.1.2 Modified Delphi Method

The modified Delphi method is an iterative process that focuses on refining expert judgments through anonymous feedback over multiple rounds.[14] Experts provide initial estimates independently, which are then aggregated and shared with the group. Each participant revises their input after reviewing this feedback, and the process continues until estimates converge. The original Delphi method was invented in the 1950s at the RAND corporation, and a review of Delphi studies is given in Rowe and Wright (1999).[64] The "modified" method referred to here was proposed in EFSA (2014) for the purposes of obtaining a probability distribution.[14] Delphi methods were designed more generally for surveying opinion, not specifically for eliciting distributions.

A.1.3 Investigate, Discuss, Estimate, Aggregate (IDEA) Protocol

The IDEA protocol is described in Hemming et al. (2018).[5] It involves interaction between the experts via a facilitated discussion and uses mathematical aggregation.

Experts first provide individual estimates independently, which are then shared anonymously to facilitate structured group discussions. Following these discussions, a second round of individual estimates is conducted. Quantile aggregation is used: the mean for each judgement is computed (e.g., “best estimate”, “lower estimate”, “upper estimate”) and reported, rather than fitting distributions to each expert’s judgements and then aggregating the distributions. Equal weights can be used in the aggregation, or performance weights can be used, as in Cooke’s classical method.

A.1.4 Medical Research Council (MRC) Protocol

This is a protocol presented in Bojke et al. (2021, Chapter 10),[6] and was developed in an MRC funded project on expert elicitation for health-care decision-making. The protocol stipulates individual elicitation for each expert, with group interaction optional; this may be conducted face-to-face, or via a Delphi process. Distributions are fitted individually to each expert’s judgements and aggregated using linear pooling.

A.1.5 Sheffield Elicitation Framework (SHELF)

SHELF involves both individual judgments and group discussion.[7] Experts provide probability judgements individually. These individual assessments are then presented in a facilitated group discussion, with the aim of investigating and understanding differences between the experts. The output is a single distribution that is agreed by the experts to represent the perspective of a “Rational Impartial Observer” (RIO). SHELF includes step-by-step protocols for specific univariate and multivariate elicitation problems. Although a behavioural aggregation method, the supporting software also includes tools for mathematical aggregation.

A.2 Scenario testing: theory for the constant hazard scenario

We denote the individual patient observations for the study group of interest by $x_1, \ldots, x_n$. Each observation is either a survival time, or in the case of censoring, an observation that the survival time for patient i is at least as large as x_i. We denote D to be the observations $x_1, \ldots, x_n$ plus additional information about which observations are censored, if any.

Let X denote the survival time for a patient in the study group population. We choose a time t^* after which we suppose that the hazard will remain constant, so that

$$X - t^* | X > t^* \sim Exponential(rate = \lambda).$$

The proportion surviving to at least time T can then be expressed as

$$S(T) = S(t^*)P(X > T | X > t^*) = S(t^*)\exp\big(-\lambda(T - t^*)\big).$$

Both $S(t^*)$ and λ are uncertain; we make a further assumption that these are independent, so that we can sample from the distribution of $S(T)$, conditional on the available data D, by sampling independently from the distributions of $S(t^*)|D$ and $\lambda|D$ and then multiplying the corresponding terms. An approximate 95% credible interval is obtained by drawing a large sample from the distribution of $S(T)|D$, and then obtaining the sample 2.5th and 97.5th percentiles.

We derive approximate posterior distributions $S(t^*)|D$ and $\lambda|D$, assuming weak prior information. We set

$$\log S(t^*)|D \sim N\left(\log \hat{S}(t^*), \frac{1}{\hat{S}(t^*)^2}\widehat{var}\left(\hat{S}(t^*)\right)\right) I_{[0,1]}\big(S(t^*)\big),$$

where $\hat{S}(t^*)$ is the Kaplan-Meier estimate of $S(t^*)$, and $\widehat{var}\left(\hat{S}(t^*)\right)$ is Greenwood's estimator of the variance of $\hat{S}(t^*)$, and $I_{[0,1]}\big(S(t^*)\big)$ is an indicator term truncating $S(t^*)$ to the interval $[0,1]$. Note that this approach corresponds to the default method for obtaining Kaplan-Meier confidence intervals in R using the survival::survfit() function.

For λ, we construct a reduced data set $\widetilde{D}$ obtained by discarding any observations less than t^*, and subtracting t^* from each remaining observation (so that we are modelling $X - t^*$). We set

$$\lambda|\widetilde{D} \sim N\left(\hat{\lambda}, \widehat{var}\big(\hat{\lambda}\big)\right) I_{[0,\infty)}(\lambda),$$

where $\hat{\lambda}$ is the maximum likelihood estimate of λ, and $\widehat{var}\big(\hat{\lambda}\big)$ is the estimated variance of the maximum likelihood estimator, assuming asymptotic normality, and $I_{[0,\infty)}(\lambda)$ is an indicator term truncating λ to the interval $[0, \infty)$.

To interpret the results of the constant hazard scenario test, it can help to make comparisons with hypothesis testing. Firstly, we consider a null and alternative hypothesis:

- $H_{0,1}$: $h(t) \geq h(t^*)$ for all $t \in [t^*, T]$;
- $H_{A,1}$: $h(t) < h(t^*)$ for at least one $t \in (t^*, T]$,

i.e., the null hypothesis is no increase in hazard. The two hypotheses are mutually exclusive and exhaustive regarding possible hazard functions over the interval $[t^*, T]$. For a fixed hazard function over the interval, $[0, t^*]$, the maximum possible value of $S(T)$ that can occur under $H_{0,1}$ is obtained by assuming the exponential model with constant hazard $h(t) = h(t^*)$ for $t \in [t^*, T]$. This is because, assuming $H_{0,1}$ to be true,

$$S(T) = \exp\left(-\int_0^T h\,(t)dt\right) \leq \exp\left(-\int_0^{t^*} h\,(t)dt - \int_{t^*}^T h\,(t^*)dt\right).$$

Hence, if an expert judges a non-negligible probability that $S(T)$ will exceed $S_{0.975}(T)$, that would imply support for the hypothesis $H_{A,1}$, as no model under $H_{0,1}$ is likely to result in $S(T)$ exceeding $S_{0.975}(T)$.

However, whilst only models consistent with $H_{A,1}$ can result in relatively large values of $S(T)$ compared to $S_{0.975}(T)$, both hypotheses include models that can result in relatively smaller values of $S(T)$; we cannot use a judgement of non-negligble probability of $S(T)$ less than $S_{0.975}(T)$ to infer support for one hypothesis over the other.

Similar considerations apply for the lower credible limit $S_{0.025}(T)$. Here, the relevant hypotheses are

- $H_{0,2}$: $h(t) \leq h(t^*)$ for all $t \in [t^*, T]$;
- $H_{A,2}$: $h(t) > h(t^*)$ for at least one $t \in (t^*, T]$.

In this case, if an expert judges a non-negligible probability that $S(T)$ will be less than $S_{0.025}(T)$, this would imply support for the hypothesis $H_{A,2}$, as no model under $H_{0,2}$ is likely to result in $S(T)$ being below $S_{0.025}(T)$.

Additionally, whilst only models consistent with $H_{A,2}$ can result in relatively small values of $S(T)$ compared to $S_{0.025}(T)$, both hypotheses include models that can result in relatively larger values of $S(T)$; we cannot use a judgement of non-negligible probability of $S(T)$ above $S_{0.025}(T)$ to infer support for one hypothesis over the other.

It may be possible to find other hazard or modelling assumptions such that $S(T)$ is *only* likely to fall within the interval $\left(S_{0.025}(T), S_{0.975}(T)\right)$ under these assumptions. If these assumptions are meaningful to the expert, reporting them could provide useful feedback, e.g., "You have judged that $S(T)$ is certain to lie somewhere inside the interval $\left(S_{0.025}(T), S_{0.975}(T)\right)$. This is only possible if…", with the corresponding assumptions stated.

APPENDIX B

B.1 AN EVIDENCE DOSSIER TEMPLATE

The template can be downloaded from the online supplementary material.

APPENDIX C

C.1 SOFTWARE INSTALLATION AND CODE USE FOR THE EXAMPLES

Using the SHELF R package

The supporting software for SHELF is provided as an R package. R can be installed for free from

- https://cran.r-project.org/

The SHELF R package can be installed from R with the command

```
install.packages("SHELF")
```

The package includes various shiny apps. The apps can also be run online, and are listed at https://shelf.sites.sheffield.ac.uk/software.

The main app for survival extrapolation is run with the command

```
SHELF::elicitSurvivalExtrapolation()
```

Code to produce Figure 3, Section 3.5.1.

```
# Make an example data set
sdf <- survival::veteran[, c("time", "status", "trt")]
colnames(sdf) <- c("time", "event", "treatment")
sdf$treatment <- factor(sdf$treatment,
                        labels = c("standard", "test"))

# Produce the extrapolation plot for the "test" treatment group
SHELF::survivalModelExtrapolations(sdf,
                                   tEnd = 500,
                                   group = "test",
                                   tTruncate = 100)
```

Code to produce Figure 4, Section 3.7.2.

```
# Make an example data set
sdf <- survival::veteran[, c("time", "status", "trt")]
colnames(sdf) <- c("time", "event", "treatment")
sdf$treatment <- factor(sdf$treatment,
                        labels = c("standard", "test"))
```

```
SHELF::survivalScenario(tLower = 0,
                        tUpper = 150,
                        expLower = 100,
                        expUpper = 150,
                        tTarget = 250,
                        survDf = sdf,
                        expGroup = "standard")
```

Linear pooling

Linear pooling can be implemented via the app

```
SHELF::elicitMultiple()
```

but we show the command line implementation here, specifically, the use of linear pooling to identify candidate probabilities for RIO judgements, as discussed in Section 4.2.5.

We use the example judgements shown in Figure 6.

```
# Lower plausible limits for each expert
l <- c(3, 5, 8, 10 )/100


# Upper plausible limits for each expert
u <- c(15, 20, 25, 30)/100

# Quartile judgements for each expert, arranged in a matrix,
# one column per expert

v <- matrix(c(8, 10, 12,
              10, 12, 15,
              12, 15, 18,
              17, 20, 23),
            3, 4)/100

# Fit distributions to each expert's judgements
individualFits <- SHELF::fitdist(vals = v,
                                 probs = c(0.25, 0.5, 0.75),
                                 lower = l,
                                 upper = u)
```

For the lower tail, we can report the 10th and 30th percentiles (0.1 and 0.3) quantiles of the fitted values, in this example, using a beta fitted distribution

```
# Obtain quantiles from linear pool, with beta distributions fitted
SHELF::qlinearpool(individualFits,
                   q = c(0.1, 0.3),
                   d = "beta")

## [1] 0.082 0.111
```

Hence, we might ask the experts to propose a RIO probability $Pr(S(T) \leq x)$ for choosing x to be somewhere in the range 8% to 11%.

Similarly, for the upper tail, we can try the 70th and 90th percentiles (0.7 and 0.9) quantiles of the fitted values

```
# Obtain quantiles from linear pool, with beta distributions fitted
SHELF::qlinearpool(individualFits,
                   q = c(0.7, 0.9),
                   d = "beta")

## [1] 0.168 0.218
```

Hence, we might ask the experts to propose a RIO probability $Pr(S(T) \geq x)$, choosing x to be somewhere in the range 17% to 22%.

This is intended to give the facilitator some suggestions for what RIO probabilities to ask for; it shouldn't be shown to the experts! The group discussion may provide other prompts for appropriate values.

APPENDIX D

D.1 Review search strategies and data extraction of the broader literature

A protocol for the review of the wider literature is available via the Open Science Framework.[37] We highlight the key methodology of the review here, further details of the search strategy or methods can be found in the published protocol.

D.1.1 Search strategy

Three literature databases were included within the search: MEDLINE(R) via Ovid, Embase, and Web of Science. These three databases were selected to ensure a broad coverage of the wider literature, whilst retaining focus on likely studies within the health-care decision-making setting. The search algorithm was designed with an Information Specialist. The search algorithms can be found from the published protocol.[37]

All databases were searched from inception until either 25th July 2024 (for Medline and Web of Science) or 16th Augst 2024 (for Embase). This time difference was due to subsequent addition of the Embase database to ensure full coverage of the literature.

To supplement the search of the literature databases, three key papers identified prior to the search, Ayers et al. (2022),[44] Cope et al. (2019)[40] and Willigers et al. (2023),[47] were used to identify further relevant literature via forwards and backwards citation searches. This was implemented via the web version of citationchaser on the 21st August 2024.[65]

D.1.2 Inclusion and exclusion criteria

Due to the nature of the research question, we followed the Sample, Phenomenon of Interest, Evaluation, Research type (SPIDER) framework for defining the inclusion and exclusion criteria as an alternative to the PICOS framework.[66]

Table D.1.1: Eligibility criteria for studies in the review using the SPIDER framework.

Inclusion criteria	
Sample	Clinical and methodological experts
Phenomenon of Interest	I. Use of expert elicitation for survival outcomes. II. Use of more general expert consultation for survival outcomes.
Design	Primary empirical studies
Evaluation	I. Documentation of the use of structured expert elicitation for survival outcomes in the HCDM context. II. Mention of consultation with experts specifically surrounding long-term survival outcomes.
Research Type	Empirical studies published in the English language in peer-reviewed journals.
Exclusion criteria	
Discussion pieces, conference abstracts and reviews.	

D.1.3 Study selection

Study selection was performed by two independent researchers both at the initial screening stage and at the full-text review stage.

D.1.4 Data extraction and synthesis

Data was extracted from studies conditional on whether structured expert elicitation or general expert consultation was used for the long-term survival outcomes. In cases where this was unclear, it was assumed that general consultation was performed due to a lack of methodological information. Studies which had any of the following characteristics were deemed to have used general expert consultation methodology to obtain the expert's input:

- Qualitative opinion sought only, no quantitative judgements or estimates;
- No quantification of individual expert uncertainty;
- No use of an existing framework, or reference to existing expert elicitation methodologies.

Relevant data items for extraction of papers where structured expert elicitation for survival outcomes was used included: first author name, publication year, clinical area, number of experts, the expert selection process, (base) elicitation framework,

elicitation setting (in-person or online), roles listed (e.g., facilitator), declaration of conflicts of interest, details of expert backgrounds/specialties, details of training provided, details of evidence dossier preparation/content, quantities of interest (QoI), individual judgement method (e.g., tertiles, quartiles, roulette), aggregation of expert judgements (e.g., RIO or mathematical aggregation), form of QoI (e.g., range, fitted distribution), rationale for expert judgements provided, discussion of the hazard, discussion of limitations/benefit of expert judgements. These items are all key components of a structured expert elicitation and thus we sought to extract as much information relating to the exercise as possible.

Relevant data items for extraction of papers where general expert consultation was conducted for survival outcomes included: first author name, publication year, clinical area, number of experts, method (e.g., advisory board meeting, survey) and purpose (e.g., model selection or external validity assessment).

Data were extracted from each included study alongside a narrative synthesis to allow comparisons to be made across studies.

D.2 Review search strategies and data extraction of NICE oncology technology appraisal submissions

D.2.1 Search strategy

To supplement the findings from the broader literature and obtain a perspective on the use of experts for long-term survival outcomes within NICE submissions, we also searched recent submissions to NICE for reference to expert involvement within the extrapolation of survival. To identify candidate submissions, the NICE published guidance was searched on the NICE website, using date filters (1st October 2023 to 14th October 2024) and selecting only “Technology appraisal guidance documents”.

Any appraisals which had been terminated were excluded. Only the full company submission documents (“Document B”) were included within our assessment and therefore any studies without a downloadable version of Document B were also excluded.

Subsequently, the submission documents were reviewed and searched for the following phrases in order to identify use of experts with particular relevance to survival outcomes;

1. Elicit*
2. Expert
3. Extrap*
4. If searches 1-3 did not provide any results, then the document was manually searched for evidence of elicitation or expert consultation on survival outcomes.

Identified sections were then reviewed for more details of the methodology employed. If references to the submission appendices were made, these were not reviewed as they are not publicly available.

D.2.2 Data extraction

Due to the scarcity of information generally included within the submissions relating to expert involvement, all data relating to expert consultation or conduct of elicitation was recorded and summarised in a narrative format.

D.3 PRISMA DIAGRAM OF THE BROADER LITERATURE REVIEW

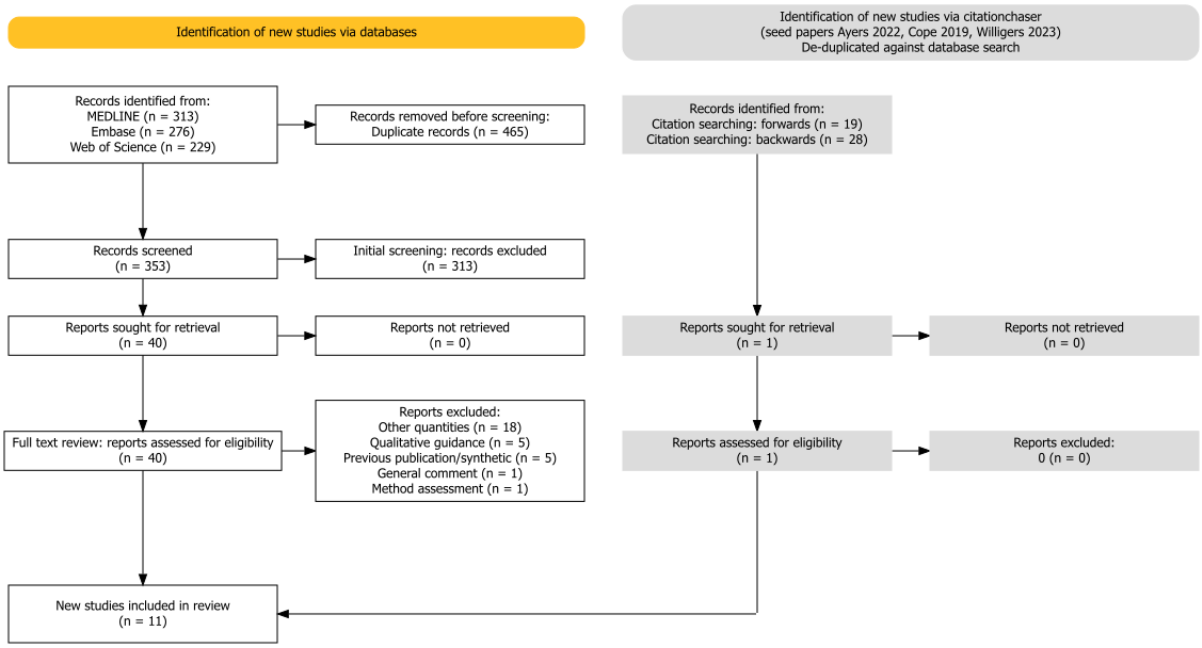

Figure D.3.1: PRISMA diagram of the selection process for the review of the broader literature.