INTRODUCTION AND GOALS

Progression-free survival (PFS) and overall survival (OS) are the most important clinical outcomes used in the assessment of clinical effectiveness and cost-effectiveness of oncology products for reimbursement decisions. The National Institute for Health and Care Excellence (NICE) in the UK, one of the most influential health technology assessment (HTA) agencies in Europe, requires that the time horizon of the cost-effectiveness analyses is long enough to capture all relevant differences between health interventions. This is supported by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and the Society for Medical Decision Making (SMDM) guidelines. Thus oncology products need to be evaluated over a lifetime horizon. However, these outcomes, particularly OS, are often incomplete—during the follow-up period of the trial not all patients experience an event. Thus, to comply with guidelines, long-term projection of data is required. Long-term extrapolation of trial data is rarely straightforward.

As demonstrated by several HTAs and specific papers, different methods of extrapolation may lead to different conclusions about the mean-life expectancy of the patients, and consequently about the cost-effectiveness of interventions.

In recognition of the significant impact of the choice of method and lack of sufficient documentation of the techniques applied, the NICE Decision Support Unit (DSU) issued a technical guideline on survival analysis for economic evaluations alongside...
clinical trials in June 2011. The DSU publication focuses on the case where individual patient level data is available for the analysts. Additional papers have since been published on aspects of the current extrapolation practices and approaches. The basic steps for extrapolation with parametric models are similar in the various recommendations (see Figure 1). The objective of this current study is to assess the effect of the DSU guidance and these recommendations in the extrapolation of OS and PFS in oncology technology appraisals conducted by NICE.

METHODS

NICE technology appraisals
A review of all NICE technology appraisals completed between June 2011 and August 2013 for oncology drugs was conducted, and manufacturer submissions were reviewed and extracted. NICE ERG (Evidence Review Group) reports were also reviewed to identify any criticisms of the approach chosen by the manufacturer and alternative methods recommended. The next step was to assess if the criticism and recommendations were applied in the ERG’s models developed for Multiple Technology Assessments (MTAs) or within the manufacturer’s model in Single Technology Appraisals (STAs).

Data extraction
The following data were extracted and reviewed for both the data submitted by the manufacturer and the final data accepted by the Review Committee:

- Details of the technology appraisal
  - Disease area and line of therapy
  - ERG
  - Issue time of the guidance
  - Modelling approach
  - Model time horizon and mean/median age of patients
- Details of the extrapolation of PFS and OS
  - If data was extrapolated
  - If yes, what was the final methodology applied
  - How was the extrapolation method chosen
  - How the choices are validated
  - Criticism and conclusions of the ERG
  - Final decision of the committee regarding the drug appraised

RESULTS

Appraisals
In total 21 technology appraisals (TAs) were identified. Of these 21 TAs, one for bone metastasis was excluded, and 20 were extracted. There were 15 STAs and 5 MTAs—including 33 separate models altogether—7 by ERGs and 26 submitted by manufacturers. Four models, which were part of an MTA, were excluded.

MAIN STEPS FOR EXTRAPOLATION WITH PARAMETRIC SURVIVAL CURVES

Source: Remak, et al. ISPOR workshop 2012, Berlin, Germany

figure 1
as publicly available descriptions were insufficient to assess the extrapolation techniques applied. Indication for the TAs included breast cancer (5 TAs); haematological cancers (5 TAs); ovarian cancer, lung cancer, prostate cancer and melanoma (2 TAs each); and, colorectal and transitional cell urothelial tract carcinoma (1 TA each). Ten of the solid tumour TAs included advanced and/or metastatic disease.

**Characteristics of the models**

In the TAs for solid and haematological tumours, model structures were different. Models for solid tumours included the following three main health states (see Table 1):

- Stable or pre-progression health state defined mostly by the PFS curve (with or without adverse events)
- Post progression or progressed health state
- Death defined by OS

Although labelled differently in the submissions (e.g., state-transition model, survival partition model or Markov or semi-Markov model), the underlying structures were similar, with PFS and OS describing disease progression modelled independently of each other.

The models presented for some haematological cancers included health states for various phases of the disease, as well as response status and, therefore, had considerably more complex structures.

**Survival modelling approaches**

For the intervention of interest, PFS and OS were modelled at least partially based on patient-level trial data in 75% of the models. The remaining quarter of models were prepared by ERG groups without access to patient level data and relied on published literature, including plots of Kaplan-Meier curves submitted by manufacturers. In oncology the comparators included in the trial may not be the relevant comparator in the UK due to regional variation in treatments and rapid change in treatment patterns. As a result, literature and data from meta-analyses also played an important role in modelling PFS and OS of comparators.

The extent of extrapolation (i.e., the difference between model time horizon and time span of trial data) was on average 15.5 years (ranging from 2.6 to 29.2 years). Data from Kaplan-Meier curves were applied directly in 24% of the models. However, apart from one submission where data was fully mature, some form of parametric extrapolation was applied for the part of the time horizon not covered by the trial data.

Parametric extrapolation was applied in 75.9% of the models, with the most commonly used distributions being Weibull and exponential (see Table 2). Usually the same type of distributions were chosen across treatment arms, however, in a small proportion of cases, the distributions differed. When different distributions provided the best fit for the treatment arms, best-fitting distributions were in some cases rejected in favour of using the same distribution based on clinical expert opinion. The treatment arms were mainly modelled separately, with joint models fitted in only 25% of cases for both PFS and OS.

Approaches to statistical modelling of PFS seemed better documented than that of OS. The final choice was mostly supported by results.
Statistical testing such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) and visual inspection were rarely mentioned and even more rarely presented. This may partly be due to lack of space or not publishing appendices submitted by the manufacturer. Validation approaches that were reported or presented are shown in Table 3. External validity of extrapolations, plausibility with clinical practice and biological/clinical explanation were rarely explored. No clinical rationale was provided for the modelling approaches in most cases. Structural uncertainty was explored by assessing the effect of applying various extrapolation methods on the cost-effectiveness results with corresponding ICERs reported in 31% of models.

### Characteristics of the 29 Models Extracted

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model duration (mean, range)</td>
<td>25.2 (2.9-50)</td>
</tr>
<tr>
<td>Data duration in years (mean, range)</td>
<td>2.90 (0.8-8.0)</td>
</tr>
<tr>
<td>Survival partition model with 3 health states, n (%)</td>
<td>17 (60.7%)</td>
</tr>
<tr>
<td>Extrapolation used?</td>
<td></td>
</tr>
<tr>
<td>Yes (%)</td>
<td>28 (96.6%)</td>
</tr>
<tr>
<td>No (%)</td>
<td>1 (3.4%)</td>
</tr>
<tr>
<td>Data sources for PFS/ OS, n (%) / n (%)</td>
<td></td>
</tr>
<tr>
<td>Clinical trial data</td>
<td>21 (72%) / 16 (55%)</td>
</tr>
<tr>
<td>Literature</td>
<td>5 (17%) / 5 (17%)</td>
</tr>
<tr>
<td>Other (not reported; mix of trial data plus literature)</td>
<td>3 (10.4%) / 8 (27.6%)</td>
</tr>
</tbody>
</table>

**Table 1**

### Description of Final Parametric Extrapolation

<table>
<thead>
<tr>
<th></th>
<th>PFS</th>
<th>OS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-parametric techniques only</td>
<td>1 (3.4%)</td>
<td>1 (3.4%)</td>
</tr>
<tr>
<td>Parametric techniques</td>
<td>22 (75.9%)</td>
<td>22 (75.9%)</td>
</tr>
<tr>
<td>Mix of non-parametric and parametric approach</td>
<td>6 (20.7%)</td>
<td>6 (20.7%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distributions tested in number (%) of submissions</th>
<th>PFS</th>
<th>OS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>15 (51.7%)</td>
<td>14 (48.3%)</td>
</tr>
<tr>
<td>Weibull</td>
<td>16 (55.2%)</td>
<td>13 (44.8%)</td>
</tr>
<tr>
<td>Gompertz</td>
<td>8 (27.6%)</td>
<td>5 (17.2%)</td>
</tr>
<tr>
<td>Lognormal</td>
<td>8 (27.6%)</td>
<td>7 (24.1%)</td>
</tr>
<tr>
<td>Loglogistic</td>
<td>8 (27.6%)</td>
<td>7 (24.1%)</td>
</tr>
<tr>
<td>Gamma</td>
<td>4 (13.8%)</td>
<td>4 (13.8%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fitting of treatment arms?*</th>
<th>PFS</th>
<th>OS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint</td>
<td>11 (37.9%)</td>
<td>8 (27.6%)</td>
</tr>
<tr>
<td>Separate</td>
<td>15 (51.7%)</td>
<td>16 (58.6%)</td>
</tr>
</tbody>
</table>

*Note: Does not sum to 100%, included mixed approach in both counts, remaining items include “not reported”

**Table 2**
Comments on survival analysis by the ERG

Modelling of OS and PFS and their extrapolation was an important topic in all ERG reports due to its critical impact on results. Alternative scenarios for survival modelling were explored and implemented in the submitted models by the ERG in several STAs. The most important comments and criticisms were that:

1. choice of survival function was not justified;
2. the parametric distribution could not capture changes in hazard expected during the course of the disease, therefore, a piecewise model would have been preferred;
3. no clinical rationale was provided for the modelling approach;
4. the long-term extrapolation of OS was highly uncertain; and,
5. use of extrapolation methods applied in prior TAs as guidance without exploring the data.

Criticisms were consistent for ERG groups and the ERGs often had strong views about the appropriate extrapolation methods. However, there were often differences between the views of the different ERGs.

DISCUSSION

Based on the examined evidence, methods of selecting the extrapolation approach in oncology TAs by manufacturers and ERGs were heterogeneous despite the available guidance.

Several assessments incorporated some form of parametric modelling for the extrapolation of survival data, either in the form of a single curve or as piecewise models. Kaplan-Meier curves were also often relied on for the duration of the pivotal trial in the assessment, with the distributions incorporated only from the end of the follow-up period. For the final choice of the approach, the majority of submissions depended mainly on the statistical goodness-of-fit criteria and visual assessment.

Beyond statistical goodness of fit and visual assessment, clinicians’ opinions about the shape of the OS curves and pragmatic modelling aspects were also taken into account. A pragmatic aspect is important in the case when indirect treatment comparison is only available via incorporation of a hazard ratio estimates. That most often led to selection of the Weibull model despite its worse fit in terms of statistical measures.

Biological/clinical explanation was discussed in very limited number of cases—making it the biggest gap in extrapolation practices. However, validation by key opinion leaders was more often sought in TAs published in 2013, particularly for the selection of the base case distribution for OS as the extent of the extrapolation, and therefore the uncertainty about the tail of the curve is much greater than for PFS.

Although the recently published methodological guidelines recommend various steps to reduce this uncertainty, the implementation of these is still rare. However as a welcome new trend, for some more recent models, actual cost-effectiveness outcomes are presented not only for the base case extrapolation but for alternatives, addressing a key structural uncertainty in modelling.

Muszbek and colleagues along with Grieve, et al., suggest that large registries may be a good source of data for testing plausibility. However, such a validation comes with its own issues, such as how to handle differences between the registry data and the trial population, and how to account for changing treatment patterns over time in registries.
Ishak, et al., review the most commonly used statistical distributions, and describe an objective process of identifying the most suitable parametric distribution in a given dataset that can be applied with both individual-patient data and with survival probabilities derived from published Kaplan-Meier curves. Grieve and colleagues highlight some weak points of the DSU guidelines and encourage further debate. Bagust and Beale, from the Liverpool Evidence Review Group for NICE, aim to provide a “practical guide” to the broad health technology assessment (HTA) community about extrapolation. They also criticize some points of the DSU guidance, including the use of log cumulative hazard-log time plot for diagnostics, and recommend the cumulative hazard-time plot instead; and recommend a piecewise approach, whereby the parametric function is only fitted to later parts of the Kaplan-Meier curve. The disagreement between researchers at the Liverpool ERG and the authors of the DSU guidance is tangible in the HTA reports assessed by this specific group. Evaluations issued by the Liverpool group criticized manufacturers for following the approach outlined by the DSU. This can be disorienting for manufacturers preparing a submission.

In light of the above discussions within and outside NICE, it would be helpful if further specific guidance would be developed on:

- How to carry out external validation/plausibility testing, including guidance on external validation
  - for clinical/biological plausibility
  - with the help of additional datasets, including registry data
- The relative importance of the various elements of testing (statistical criteria; clinician opinion, external data)

The present analysis has important limitations. It relies on information reported in the published TA documents. Potentially not all validation work was reported; e.g., diagnostic tests for survival analyses may have been conducted and not reported, or presented only in appendices to the main body of the manufacturer submission and therefore not publicly available. As a consequence, practices may be closer to the guidelines than reported here. Second, several changes were made to the extrapolation approach during the appraisal process, and these changes were not incorporated in the data extraction.

**CONCLUSION**

Since the publication of various publications on survival extrapolation, and the publication of NICE technology assessment reports, the practice and description of extrapolations have improved within the oncology technology appraisals in the UK, contributing to more transparent decision-making. However, there are still several areas where further discussion and more specific guidance would be welcome.

**Acknowledgments**

The authors would like to thank Linda Hortobagyi, Zsofia Kiss, Panagiotis Orfanos, Rebecca Baggaley, and Eszter Tichy at Evidera for their help in the data extraction.

For more information, please contact Agnes.Benedict@evidera.com or Noemi.Muszbek@evidera.com.

**References**