OPEN ACCESS

University of BRISTOL

Peer reviewed version

University of Bristol - Explore Bristol Research
General rights

Extrapolation of survival curves from cancer trials using external information

Patricia Guyot[1, 2] PhD, A.E. Ades[1] PhD, Matthew Beasley[3] MD, Béranger Lueza[4,5] MSc, Jean-Pierre Pignon[4,5] MD, Nicky J. Welton[1] PhD

[1]School of Social and Community Medicine, University of Bristol, Bristol, UK

[2]Mapi, Houten, the Netherlands

[3]Bristol Haematology and Oncology Centre, Bristol, UK

[4]Gustave Roussy, Université Paris-Saclay, Service de biostatistique et d'épidémiologie / Ligue Nationale Contre le Cancer meta-analysis plateform, F-94805 Villejuif, France

[5]Université Paris-Saclay, Univ. Paris-Sud, UVSQ, CESP, INSERM, F-94085 Villejuif, France

Corresponding Author: Patricia Guyot, School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS UK

Tel: +33 (0)4 72 13 51 46

Fax: +33 (0)4 72 13 51 40

Email: pguyot@mapigroup.com

Abstract

Background: Estimates of life-expectancy are a key input to cost-effectiveness analysis (CEA) models for cancer treatments. Due to limited follow-up in Randomized Controlled Trials (RCTs), parametric models are frequently used to extrapolate survival outcomes beyond the RCT period. However, different parametric models that fit the RCT data equally well may generate highly divergent predictions of treatment-related gain in life expectancy. **Objectives:** We investigate the use of information external to the RCT data to inform model choice and estimation of life-expectancy. **Methods**: We used Bayesian multi-parameter evidence synthesis to combine the RCT data with external information on general population survival, conditional survival from cancer registry databases, and expert opinion on the time course of the treatment effect. We illustrate with a 5 year follow-up RCT of cetuximab plus radiotherapy versus radiotherapy alone for head and neck cancer. **Results:** Standard survival time distributions were insufficiently flexible to simultaneously fit both the RCT data and external data on general population survival. Using spline models we were able to estimate a model that was consistent with the trial data and all external data. A model integrating all sources of internal and external evidence achieved an adequate fit and predicted a 4.7 months (95% CrL: 0.4; 9.1) gain in life expectancy due to Cetuximab. **Conclusions:** Long-term extrapolation using parametric models based on RCT data alone is highly unreliable and these models are unlikely to be consistent with external data. External data can be integrated with RCT data using spline models to enable long-term extrapolation. Conditional survival data could be used for many cancers and general population survival may have a role in other conditions. The use of external data should be guided by knowledge of natural history and treatment mechanisms.

<285 words>

Introduction

In randomized controlled trials (RCTs) reporting survival outcomes, progression-free survival and overall survival are generally graphically displayed with the Kaplan-Meier method (KM) (1) and analyzed by Cox Regression (2). However, neither method is sufficient for the purposes of Cost Effectiveness Analysis (CEA) because they do not provide a model which can be extrapolated beyond the RCT period. Instead, parametric survival models are required to estimate expected survival (3). Models most often used in submissions to the National Institute for Health and Care Excellence (NICE) have a parametric model, such as Exponential, Weibull, Log-logistic, or Log-normal (4-5) distributions, for the control arm and a constant hazard ratio (proportional hazards) to predict the treatment arm (3). Parametric distributions differ in their flexibility: the exponential requires a constant hazard, the Weibull has monotonically increasing or decreasing hazards, while the log-normal and log-logistic allow for "bowl" or "hat" shaped hazards (1). Sometimes the proportional hazard assumption is relaxed by fitting an accelerated failure time (AFT) model, or by fitting unrelated models to each arm, also known as Fitted Separately to Each Arm (FSEA) (3). Model selection is not straightforward, and models which fit equally well to the observed RCT data may give very different estimated mean survival gains (6). This is due to mean survival being very sensitive to the tails of the survival distribution, which are usually not captured within RCT follow-up periods.

There is a well-recognized need to improve extrapolation of survival data for use in the context of cost-effectiveness analysis (7), and there have been several attempts to do this using data external to the trial. One approach has been to use external data to inform the choice of parametric model to extrapolate the trial data (8-11). Other investigators have replaced the control arm by external data, with (12-14) or without (15-17) adjustment, while maintaining the treatment effect from the trial. Most methods using data external to the trial require a process of matching the trial and external populations (18), but investigators have differed in the relative weight given to each, and in how uncertainty in each source of information is reflected in the final model.

This paper presents a method for using external information to extrapolate survival curves in the specific area of cancer trials. Cancer is an interesting and important area in this respect for several reasons. Firstly there is a wealth of information about long term survival from cancer registries. Second, the

difficulties in choosing parametric models for extrapolation is well-understood (4), and in the UK these have contributed to a series of controversies in decisions as to which new cancer treatments should be used in the National Health Service (19-21).

A trial comparing radiotherapy plus cetuximab (Erbitux) (22) to radiotherapy alone for head and neck cancer patients is used to illustrate our approach. We consider two types of external information to inform extrapolation of the control arm: population data on overall survival of an age- and sex-matched cohort, and cancer registry information on conditional survival of a matched cohort of head and neck cancer patients. A third external source is information on the relative treatment effect, derived from an analysis of the literature and treatment mechanisms.

We begin with a description of the illustrative dataset and, to motivate the paper, we compare the performance of a series of standard parametric survival models on goodness of fit and predicted gain in life expectancy. We then briefly review what is known about the clinical epidemiology of head and neck cancer and the mechanisms underlying treatment with radiotherapy and cetuximab, and provide some preliminary analyses of data from cancer registries and recent meta-analyses, which justify our use of conditional survival data and the external information on treatment effect. The statistical methods are presented in outline, with further details and WinBUGS code appearing in a Web appendix. In the discussion section we consider the general properties of our approach compared to previous uses of external data in the literature, and consider the generalizability of the methods to other cancers, and to extrapolation of survival curves more widely.

Motivating example

Bonner et al (2006) (22) conducted an RCT to compare radiotherapy plus cetuximab versus radiotherapy alone, with overall survival recorded over a 5-year follow-up period. We have used a method (23) that accurately reconstructs the life-table data from the Kaplan-Meier curves published in that paper. These data are used throughout the paper and are shown in the Web appendix.

In the analyses that follow we consider how to extrapolate both arms in order to inform a potential CEA of cetuximab, aimed at a target population represented by the Bonner trial. The trial was carried out

between 1999 and 2002 on treatment-naïve patients in several countries. The patient sample was 80% male with median age of 57 years, range 34 to 83. The most common site of primary tumors was oropharynx (59.5%), followed by larynx (25.5%) and hypopharynx (15%). All patients were classified as stage III or IV. Other factors of potential clinical and cost-effectiveness relevance are the EGFR (Epidermal Growth Factor Receptor) status (22), the performance status, Karnofsky or ECOG (Eastern Cooperative Oncology Group), and the severity of comorbidities (24-25).

Results with standard survival models without external data

Table 1 shows Deviance Information Criteria (DIC) (26), posterior mean deviance, $\bar{D}$, and gain in life expectancy predicted by 12 commonly used parametric survival models. $\bar{D}$ measures overall model fit, whereas the DIC is a composite measure of fit and complexity useful to identify the most parsimonious model, trading off fit and complexity. Models with lower $\bar{D}$ and DIC are preferred although differences less than 3 (or even 5) (27) are not considered important. Models with similar goodness of fit generate very different estimates of life expectancy gain (Table 1). Visually, the AFT and FSEA Log-normal models have an almost identical fit to the KM curves (Figure 1) and similar DIC (Table 1), but the predicted gain in life expectancy due to cetuximab is 80.4 months for the FSEA and only 32.3 for AFT. Similarly, the Generalised Gamma (28) AFT gives similar DIC to the Log-normal FSEA, however the predicted survival gain is only 13.9 months, nearly 6 times lower. The log-normal and Generalised Gamma models present the best fit to the data according to the $\bar{D}$ and DIC values.

Clinical epidemiology of head and neck cancer and mechanism of treatments

Head and neck cancer usually displays a rapidly increasing mortality over the first 36 months from diagnosis, which then levels off (29). Survival rates differ markedly according to the site of the cancer. Cancer of the larynx is typically diagnosed earlier than cancers of the hypopharynx and oropharynx, and patients usually have a better chance of surviving their cancer (24-25, 29). Longer-term registry-based studies from the US and Netherlands (30, 31) show that relative survival of head and neck cancer stabilizes after 5 or 6 years. Hence, UK guidance advises regular examination of the neck during the first two years after treatment, and discharge from routine follow-up after five years (32). Subsequently, patients can be considered as essentially "cured" from their head and neck cancer; although they continue to experience excess mortality due to risk factors associated with head and neck cancer such

as alcohol abuse, tobacco and Human Papillomavirus (HPV) (33, 34). According to the above registry studies this excess mortality persists for at least 15 years.

To confirm the relevance of these results in the present context, we constructed a cohort of cancer patients matched to the Bonner trial population for age-, gender-, - cancer site and date of diagnosis (35) using registry data from the Surveillance Epidemiology and End Results (SEER) database (36). One-year conditional survival in this cohort is shown alongside conditional survival in the trial (Figure 2A). The results accord closely with previous work reported above, and show the close agreement of trial and SEER data. Also shown is the conditional survival of an age- and sex- matched general population cohort based on US survival statistics (37). (See Web Appendix A and B details of how these matched cohorts were constructed). This confirms that conditional relative survival remains approximately constant and less than one for at least 20 years after the initial 5 years from diagnosis.

Cetuximab is administered over a single 8-week period concurrently with radiotherapy. Its mode of action is to enhance the effect of radiotherapy (38), by increasing the proportion of cancer cells that are sensitive to radiotherapy. For this reason, the time course of the effect of cetuximab can be expected to be the same as the time course of the effect of radiotherapy. Similarly, we would expect that the effect of both therapies in reducing head and neck cancer mortality should be limited to the initial 5 or 6 years during which mortality due to the head and neck cancer predominantly occurs.

To further check our interpretation of the clinical literature on head and neck cancer, we undertook additional analyses of data incorporated in the Pignon (2009) meta-analysis comparing loco-regional treatment (radiotherapy and surgery plus postoperative radiotherapy) versus the same loco-regional treatment plus chemotherapy (39-40). 5-year overall survival rates in the meta-analysis control arm and the Bonner control arm were very close (p=0.74), in spite of some differences in stage distribution and site. Further, this independent source of data confirms the same pattern of conditional mortality as previous literature (Figure 2B): a rising conditional mortality over the first 5 years, reflecting an increasing proportion of deaths from causes other than head and neck cancer, followed by stable conditional mortality in years 6-8.

Also, confirming our understanding of the treatment mechanisms, the advantage of chemotherapy in conditional survival can be seen over years 1-5, but disappears after that. To our knowledge, there is currently no evidence associating cetuximab with benefits, or toxicities, after five years.

Based on these findings, we make the following three assumptions

*1. General population survival. S*urvival in the control arm of the trial will remain less than survival in the matched general population cohort over the entire time horizon of the CEA model.

2. Conditional survival: One-year conditional survival in the trial control arm will converge to that of a matched cancer cohort, five or six years after diagnosis.

3. Relative Treatment effect: The hazard ratio changes over time as a smooth function with one turning point (expected to decrease initially and then increase). The hazard ratio is assumed to begin at one when treatment begins, and return to unity by 6 years.

Statistical Methods

We begin by illustrating the impact of incorporating external data constraints on parametric survival models. We present results for the AFT Log-normal and the FSEA Generalised Gamma models. These were chosen as the most compatible with the RCT data, as indicated by DIC and $\bar{D}$ (Table 1), and because they varied in flexibility (number of parameters). Similar issues arose with all of the parametric survival models.

As a more flexible alternative, cubic spline functions were chosen to model the relationship between log cumulative hazards and log-time (41). These are cubic polynomials fitted to successive sets of two points, or internal knots, in a way that guarantees they are continuously differentiable. Further details are given in Web appendix C. Two spline models were created on the log cumulative hazard scale, one for the control arm and one for the additional relative effect of treatment.

Boundary knots were placed at the extreme ends of the data. We also specified one internal knot in the RCT segment of the data, and another internal knot in the period between the end of the trial (5 years) and the end of the external data (40 years). The knots were placed in the mid-point on the log-time scale.

External information was added incrementally: data on general population and the SEER database for the control arm were first included separately, then together, and finally information was also added on the hazard ratio.

External information is introduced by specifying the relationship between the parameters estimated by the external information, and the parameters of the survival distributions, whether these are parametric or spline functions. This is implemented by writing the likelihoods for the external data in terms of the parameters of the (extrapolated) survival model that also gives the likelihood for the trial data. In this way a single survival model is estimated from all sources of data (RCT and external) simultaneously.

Estimation

Estimation was carried out by Bayesian Markov Chain Monte Carlo (MCMC) simulation using WinBUGS (42) and WBDev (43). Observed and fitted survival, conditional survival and relative hazard ratio were visually compared. $\bar{D}$ and DIC statistics were computed within the MCMC simulation and transformed using a constant to the scale of the parametric models to allow comparisons (10). These statistics were recorded, separately for the RCT data and each type of external data. This allows us to compare model fit to the different data types. The threshold for choosing one model over another was a 5 point difference in the DIC (27) on the RCT data. Details regarding choice of initial values, convergence checks, burn-in period, and posterior sampling are given in Web appendix D.

Introducing external data on absolute survival in the general population

The age- and gender-matched general population is expected to have overall survival, $S_{GP}(t)$, that is no lower than survival $S_{0,RCT}(t)$ in the RCT control arm at any time t. We cannot put constraints directly on $S_{0,RCT}(t)$, because this is a complex function of the survival model parameters. Instead we introduce the external data in a way that imposes the constraint on the extrapolated curves. We do this at a single time point at 40 years, when the matched cohort had $r_{GP}(40) = 1{,}660$ survivors out of a denominator of $n_{GP}(40) = 158{,}858$ persons in the matched cohort at time 0 (see Web Appendix A), corresponding to 40-year survival 1.045% [95CrL: 1.036; 1.054]. Assuming a binomial distribution for this data, we implement the belief that survival in the control arm of the RCT at 40 years, $S_{0,RCT}(40)$, is no better

than that in the matched general population, $S_{0,RCT}(40) \leq S_{GP}(40)$, by giving a Binomial likelihood for the general population data:

$$r_{GP}(40) \sim Binomial(S_{GP}(40), n_{GP}(40))$$
where $S_{GP}(40)$ is constrained to be greater than $S_{0,RCT}(40)$ so that   (1)
$$S_{GP}(40) = S_{0,RCT}(40) + \beta, \quad \beta > 0$$

Additional constraints are required to ensure that $S_{GP}(40)$ lies in the interval 0,1.

Introducing external data from SEER on conditional survival

We assumed that the RCT control arm population 1-year conditional survival at time *t* conditional on being alive at time (*t*-1), $CS_{0,RCT}(t|t-1)$, is no different to the matched SEER population conditional survival, $CS_{SEER}(t|t-1)$, from 6 years onwards until 26 years (last timepoint available in the SEER database). The methods for calculating the numbers of person alive and at risk in the SEER population between time t and t-1, respectively $r_{SEER}(t|t-1)$ and $n_{SEER}(t|t-1)$, are shown in the Web appendix B, along with the data itself. Assuming a binomial likelihood for 1-year conditional survival probabilities from the SEER population, we implement this by specifying:

$$r_{SEER}(t|t-1) \sim Binomial(CS_{SEER}(t|t-1), n_{SEER}(t|t-1))$$
where $CS_{SEER}(t|t-1)$ is contrained to be equal to $CS_{0,RCT}(t|t-1)$ so that   (2)
$$CS_{SEER}(t|t-1) = CS_{0,RCT}(t|t-1), \quad 6 \leq t \leq 26 \ years$$

Introducing both general population survival data and conditional survival data

When adding the SEER data on conditional survival (Equation [2]), we observed that the 1-year conditional estimated survival curve for the RCT population crossed that from the general population at 33 years. To avoid conflict between the two sources of external data, we assumed that 1-year conditional survival in the control arm of the RCT was no different to that in the general population at one time point. In addition to equation (2) we therefore specify:

$$r_{GP}(35|34) \sim Binomial(\text{CS}_{\text{GP}}(35|34), n_{GP}(35|34))$$

where $CS_{GP}(35|34)$ is contrained to be equal to $CS_{0,RCT}(35|34)$ so that $\qquad$ (3)

$$CS_{GP}(35|34) = CS_{0,RCT}(35|34)$$

Introducing external data on the relative treatment effect

We expect that the hazard ratio is a smooth function with a monotonic decrease followed by a monotonic increase over years 0-5. This is captured by having a single knot in this interval. To reflect that there is some uncertainty about whether the hazard ratio returns exactly to 1, we introduce external data on the hazard ratio taking the value of 1 at time points $t = 6...35\,years$ each with standard error 0.1. This is

achieved using a normal likelihood: $HR(\text{t}) \sim \text{N}\left(\dfrac{h_{1,RCT}(t)}{h_{0,RCT}(t)}, 0.1^2 \right), \quad t = 6...35\,years$

$\qquad$ (4)

where $h_{0,RCT}(t)$ and $h_{1,RCT}(t)$ are the hazards on the control and treatment arms of the RCT population, respectively.

Winbugs code and data are presented in Appendix D.

Results

Parametric models with external data on general population survival

Figure 3A shows estimated curves from an AFT 2-parameter Log-normal model for (i) the unconstrained model with no external information, (ii) the model constrained by overall survival in the general population (Equation [1]), together with the KM-curves. It is evident that the external data is very far from being compatible with the trial data under this model. The fit of the trial data deteriorates markedly with the inclusion of external data, with a DIC increasing from 2,316 to 2,329 and a $\bar{D}$ from 2,313 to 2,327 (Table 1). Similar results were obtained with the FSEA Generalised Gamma (Figure 3B). Neither of these distributions is sufficiently flexible to comply with both the RCT and the general population data (Equation [1]). The same problem arose with all of the parametric models reported in Table 1. Further

external data cannot improve the flexibility of the parametric models, and so we do not present results including further external data.

Spline models

With restricted cubic splines with 2 internal knots, DIC results for the unconstrained model were between 2,304 and 2,306, depending on the sets of initial values. When the data on general population survival was added, DIC was 2,304 (Table 2). Convergence of the MCMC simulations was not satisfactory for all parameters in the model, due to the lack of data between the end of the trial and year 40, making it impossible to identify spline parameters (see Web appendix E). These numerical problems were overcome when further external constraints were incorporated.

When the external data on conditional survival was incorporated, the DIC value for the RCT data from the spline model was found equal to 2,307, very close to DIC without external data (Table 2). After the RCT period, the extrapolated trial curves were visually close to the SEER data (DIC value of 98). The SEER data fully identify the extrapolated control arm survival curve. The spline model therefore produces a survival estimates that are consistent with both RCT and all external evidence sources.

When conditional survival in the general population (Equation [3]) and SEER data (Equation [2]) are both applied, DIC for the RCT and the SEER data barely changed (Table 2). The survival curves estimated from the RCT and the general population now cross at 35 years after the start of the RCT. We could have extended this general population constraint by adding time-points between 36 and 60 years. However, considering the low percentage of patients still alive 35 years after the start of the RCT (median age at randomization was 57 years), we did not try to incorporate more general population data.

When the data on the hazard ratio was also added (Equation [4]), DIC for trial, SEER, and general population data again barely change (Table 2). Figure 4 suggests a visually good fit of the spline models to each source of data. The estimates comply with the RCT data during the RCT period and with SEER, general population and expert data after the end of the RCT. Using the flexible spline models, we have therefore managed to incorporate all external data, without deterioration in model fit for any of the different evidence sources.

The life-expectancy gain when all external data sources were incorporated was 4.7 months [95%CrL: 0.4; 9.1] (Table 1).

Discussion

We begin this concluding section by reviewing the choice of survival function (splines, parametric models, and other options). We then compare our use of external data to previous work in the literature from a technical point of view, with a particular focus on uncertainty propagation. Third, we consider the modeling choices facing investigators using external data, and suggest extensions and sensitivity analyses that could be used in practical applications. Finally, we consider the particular type of external information used here, and what this implies about the generalizability of the methods to other cancers and other conditions.

Survival models

The study adds to a growing literature on the difficulties associated with extrapolation of survival curves using standard parametric models (3, 6). Even the models that fitted the RCT data best *without* external data fitted the trial quite poorly compared to spline models *with* external data. Indeed, although it seems a weak requirement that the extrapolated survival in a cancer trial control arm must be less than would be observed in the general population, for this dataset it was easily strong enough to rule out the standard parametric curves. These distributions do not provide the possibility of "local" parameters: the behavior of the tails is strongly determined by data fit at the very beginning, and vice versa. It remains to be seen whether other flexible distributions such as fractional polynomials (43), mixture distributions (44-46), or join-point techniques could represent valid alternatives to the splines, which we found gave rise to numerical problems in some contexts (see Web appendix).

Uncertainty propagation

There are many examples of the use of external data on the general population (9, 11, 13) and also use of registries, meta-analyses or prospective observational studies of cancer cohorts (8, 10, 12-13, 15-17, 47). Many applications are tailored to specific circumstances and it is difficult to give an overall review. There are, however, marked differences in the way uncertainty is propagated. Our approach to the combination of trial and external data is an example of Bayesian multiple parameter evidence

synthesis (MPES) (48). Typically, vague priors are assigned to the basic parameters (49) and information on parameters or functions of parameters is introduced by the data likelihoods. Simultaneous estimation from all these data sources ensures a coherent model that is consistent with all the evidence included, and that fully reflects the statistical uncertainty in the evidence for propagation into a cost-effectiveness model. We would also claim that by using the highly flexible cubic splines we have also appropriately accounted for uncertainty in choice of survival model, conditional of course on our assumptions about the number of inflexions implied by the number of internal knots.

In terms of uncertainty propagation, we therefore believe our general approach has some clear technical advantages over other ways of using external data. For example, use of external data to choose a parametric model (8-9) fails to account for uncertainty in model choice. Methods that "import" an estimate of the relative effect from the trial and overlay this on a model of the control arm, will mis-represent the uncertainty in the treatment effect estimate, because it was generated from a different model of the control arm. In our approach we advocate all parameters are estimated *simultaneously*, using combined data from the RCT, the general population survival data, conditional survival data, and information on the relative treatment effect.

The main difficulty with importing a treatment effect estimate from a different model is that in virtually every case, investigators have relied on a proportional hazard assumption. Based on a logical consideration of how treatments work, our view is that the proportional hazards assumption is highly implausible in many cancer treatment trials. Instead, the effect of a treatment on mortality risk can be expected to accelerate over an initial period, and then decelerate as non-cancer causes of death begin to predominate. Although we have made quite strong assumptions about the relative treatment effect, they are reasonably well-grounded in evidence and in theory, and far weaker than the routinely accepted proportional hazards assumption.

In view of the wild variation in estimated life expectancy gain between parametric models (Table 1), it is interesting to observe that the relative uncertainty in our final estimate of mean survival gain from the spline model with all sources of external data (4.7 months with 95%CrI 0.4 to 9.1 month), is commensurate with the log hazard ratio in the original Bonner trial (22): -0.30 (-0.56 to -0.03). Both estimates exclude the null effect by a narrow margin, and have a similar coefficient of variation (standard

error divided by mean), based on the 95% intervals: 0.47 from the present study compared to 0.45 from the trial.

An earlier attempt to extrapolate survival in the Bonner trial appeared in the manufacturer's submission to NICE (47). A cure model with a logistic link was used to model the survival of head and neck cancer patients. For the non-cured patients, a Log-Normal distribution was chosen. The non-cured fraction was extrapolated for both control and treated group, while subsequent mortality in the cured fraction was based on UK population data adjusted down using a hazard ratio derived from the Pignon meta-analysis (40). This generated an expected gain of 10.6 months. A credible interval was not published. This approach has several similarities to what we are advocating here, in assuming a time-limited treatment effect and stable conditional survival subsequently, but it relies on a specific parametric form for the cured fraction and fails to incorporate uncertainty in model choice.

Modeling choices and sensitivity analyses

In constructing what is primarily a methodological exercise to demonstrate the feasibility of a method in principle, we have the luxury of not having to face the modelling choices that investigators in a real decision making context must confront. In this section we consider how the present methods could be varied or extended to different target populations, and what kinds of sensitivity analyses might be required in practice.

While we believe our approach appropriately reflects statistical uncertainty in the combined data, as well as uncertainty in model choice, it cannot in itself propagate uncertainty about the relevance or applicability of the external evidence to the target population. Bayesian theory distinguishes strictly between the "subjective" prior and the "objective" data likelihood, but this overlooks the subjectivity in the interpretation of the data – specifically the assumption that the data are providing unbiased estimates of the target parameters. For example we assume that conditional survival is exactly the same in the target (trial) population as in the matched SEER cohort. Further analyses of SEER and other registries, such as those shown in Figure 2A, can be used to show how sensitive conditional survival might be to imbalances in cancer site mix, age, and date of diagnosis and Karnofsy score.

We used the US SEER data and hence US population data to supply external data on survival and conditional survival because of the comprehensiveness of this database. Decision makers in other

jurisdictions would preferentially use locally relevant population and cancer registration data. If this was unavailable, or partially available, it might be still be possible to use SEER data, perhaps making suitable adjustments. Note however that conditional survival or relative conditional survival is more likely to generalize across jurisdictions than overall survival. Alternatively, the similarity of SEER and other national cancer registries could be investigated to inform sensitivity analysis. Similarly, we assumed that the target population was identical to the trial population, which is frequently done in Health Technology Assessments, rather than replacing the control arm with appropriately matched external data that precisely represents the chosen target population. In this case, unless the external data is drawn from a cancer registry, it may still be necessary to use conditional survival register data in the way we have to extrapolate lifetime survival. Note that the same flexible spline approach we have proposed for the time course of the *relative* treatment effect can still be applied, whether the control arm is constructed from extrapolated trial data or based entirely on external data.

In a practical application, investigators would carry out a series of sensitivity analyses. Placement and number of internal knots in the spline function might seem obvious candidates, but general experience with splines is that the location of knots has little impact on estimates, whereas too few or too many knots may degrade the goodness-of-fit (41, 50). The strategy we have followed has been to place internal knots in the middle of the (log) range, to use the same points in all models, and to have just a single knot for each segment of data. These modeling choices avoid post-hoc trawling, and are *a priori* the most easily defended. The choice of a single internal knot in the trial period is informed by our assumption that the hazard ratio would fall monotonically then rise, and that changes in conditional relative survival over time would be smooth. In a practical application, however, some sensitivity analysis around the placement of knots would be expected. One would also expect a sensitivity analysis around the assumption that the time at which the hazard ratio is assumed to return to unity.

The present analysis relied on a period-analysis (35) of survival and conditional survival. This only partly allows for improvements over time in cancer survival. Another form of sensitivity analysis, or perhaps an extension to our approach, might consider a more sophisticated model of the registry data that allowed for the continual improvement in cancer survival rates into the future. A further issue is the time between diagnosis and randomization. Patients in the Bonner trial would have been recruited soon after diagnosis, but the methods could be modified to allow for other scenarios.

More broadly, our construction of "external information" on the relative treatment effect is based on a whole series of assumptions about head and neck cancer and treatment mechanisms. This should not be regarded as a weakness of the analysis: similar assumptions would have to be made in *any* decision analysis concerned with cancer treatments. Investigators need to be explicit about their choice of external data and their interpretation of it, backing this up if possible with supporting analyses, such as those illustrated in Figure 2. The assumptions must be open to examination, debate, and sensitivity analyses, as in any other analysis of clinical or cost effectiveness.

Limitations and generalizability to other cancers and conditions

Any use of data to make predictions requires subjective judgment about the relevance and applicability of the data, but among the limitations of the method is the degree of reliance on subjective judgment about the clinical epidemiology of the condition and the treatment effect. "Uncertainty" regarding what are in effect structural assumptions is difficult to express. Another limitation is the technical difficulty in fitting spline models, particularly in evidence-sparse situations. We have experienced this problem mainly when fitting models incrementally: the problem is much less acute when all sources of external data are included. A great deal could be learned from applying these methods to extrapolate a range of survival curves in other cancers.

Besides the uncertainty propagation properties of the method, which could be applied very generally, a more substantive contribution of the paper is the suggestion that relative conditional survival, based on cancer registers or meta-analyses of cancer trials, provides a relatively accurate way of extrapolating cancer trials. Studies of registry data from many countries shows that, while conditional survival is sensitive to age at diagnosis and stage at diagnosis, for many cancers it tends to stabilize 5 to 8 years after diagnosis (51-54).

We do not expect all cancers and all treatments to behave in precisely the same way as head and neck cancer. An examination of the available literature and registry data guided by clinical experts, alongside preliminary analyses such as Figure 2, is a procedure that could be adopted in other circumstances. For example, cetuximab is usually administered in a single relatively brief course: treatments involving a longer course, several courses, and other modes of treatment like surgery or radiotherapy might

require a somewhat different approach. Similarly, if there was a suspicion of later toxicity, a sensitivity analysis could be readily constructed to allow for this.

In the present case, external data on general population survival did not represent much of a constraint on extrapolation, although it served to rule out a large number of commonly used models quite decisively. In the UK, where manufacturers submit evidence to NICE to obtain approval for the use of drugs in the NHS, uncertainty about which parametric model to apply when extrapolating cancer survival curves has, on several occasions, contributed to controversy that lead eventually to an appeal process (18-20). The use of external information, in the way suggested above, both for extrapolating the control arm and the treatment effect, would contribute to obtaining better evidence-based estimates of gain in life-expectancy.

< 5353 words>

References

1. Collett D. Modelling survival data in medical research. 2nd edition: CHAMAN & HALL/CRC; 2003. 391 p.
2. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. British journal of cancer 2003; 89: 232-238.
3. Guyot P, Welton NJ, Ouwens NJNM, ADES AE. Survival times outcomes in randomized, controlled trials and meta-analyses: the parallel universes of efficacy and cost-effectiveness. Value in Health 2011; 14(5):640-6.
4. Jackson CH, Sharples LD, Thompson SG. Structural and parameter uncertainty in bayesian cost-effectiveness models. Journal of the Royal Statistical Society Series C – Applied Statistics. 2010; 59:233-53.
5. Johnson NL,Kotz S, Balakrishnan N. Continuous univariate distributions. Vol. 1. Wiley Series in Probabilities and Mathematical Statistics: Applied Probability and Statitics (2nd ed). New York: John Wiley & Sons. 1994.

6. Latimer NR. Survival analysis for economic evaluations alongside clinical trials – Extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. Medical Decision Making. 2013; 33(6):743-54.

7. Grieve R, Hawkins N, Pennington M. Extrapolation of survival data in cost-effectiveness analyses: improving the current state of play. Medical Decision Making. 2013; 33:740.

8. Dundar Y, Haycox A, bagus A, Hil R, Dickson R, McLeod C, et al. Pemetrexed disodium for the treatment of malignant pleural mesothelioma: a systematic review and economic evaluation. Issued in January 2008. Available from:
   http://www.nice.org.uk/nicemedialive/11698/36841/36841.pdf. [Accessed january 1, 2010].

9. Demiris N, Lunn D, Sharples LD. Survival extrapolation using the poly-Weibull model. Stat Methods Med Res. 2011.

10. Jackson CH, Sharples LD, Thompson SG. Survival models in health economic evaluations: balancing fit and parsimony to improve prediction. The international journal of biostatistics. 2010; 6(1):34.

11. Green C, Bryant J, Takeda A, et al. Bortezomib for the treatment of multiple myeloma patients. Issued: October 2007. Available from:
    http://www.nice.org.uk/nicemedia/pdf/MultipleMyeloma_Evidencereviewgroupreport.pdf.
    [Accessed january 1, 2010].

12. Demiris N, Sharples LD. Bayesian evidence synthesis to extrapolate survival estimates in cost-effectiveness studies. Statistics in Medicine. 2006; 25(11):1960-75.

13. Neyt M, Van Brabandt H, Devriese S, Van de Sande S. A cost-utility analysis of transcatheter aortic valve implementation in Belgium: focusing on a well-defined and identifiable population. BMJ open. 2012; 2(3).

14. Knight C, Hind D, Brewer N, Abbott V. Rituximab (MabThera) for aggressive non Hodgkin's lymphoma: systematic review. Issued in May 2003. Available from:
    http://www.nice.org/nicemedia/live/11506/32682/32682.pdf. [Accessed january 1, 2010].

15. Dundar Y, Hounsome J, McLeod C, Bagust A, Boland A, Davis H, et al. Rituximab for the first line treatment of stage III-IV follicular non-Hodgkin's lymphoma. Issued September 2006. Available from: http://www.nice.org.uk/nicemedia/live/11591/46698/46698.pdf. [Accessed january 1, 2010].

16. Mc Kenna C, Burch J, Suekarran S, Walker S, Bakhai A, Witte K, et al. A systematic review and economic evaluation of the clinical effectiveness and cost-effectiveness of aldosterone antagonists for postmyocardial infarction heart failure. Health Technology Assessment. 2010; 14(24):1.

17. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2001; 46: 399-424.

18. Connock M, Round J, Baykiss S, Tubeuf S, Greenheld W, Moore D. Sorafenib for advanced hepatocellular carcinoma. The evidence review group's report. Available from: https://www.nice.org.uk/guidance/TA189/documents/evidence-review-group-report2 [Accessed October 2015]

19. Bagust A, Boland A, Davis H, Dundar Y, Fleeman N, Greenhalgh J, O'Reilly S, Oyee J, Trevor N. Eribulin for the treatment of locally advanced or metastatic breast cancer. The evidence review group's report. Available from: https://www.nice.org.uk/guidance/TA250/documents/breast-cancer-advanced-eribulin-evidence-review-group-report2 [Accessed October 2015]

20. Wade R, Duarte A, Simmonds M, Rodriguez-Lopez R, Duffy S, Spackman E, Woolacoot N. Aflibercept in combination with irinotecan and fluorouracil-based therapy for the treatment of metastatic colorectal cancer which has progressed following prior oxaliplatin-based chemotherapy. The evidence review group's report. Available from: https://www.nice.org.uk/guidance/TA307/documents/colorectal-cancer-metastatic-aflibercept-evaluation-report2 [Accessed October 2015]

21. Cetuximab. Wikipedia. Available from: https://en.wikipedia.org/wiki/Cetuximab

22. Bonner JA, Harari PM, Giralt J, Azaarnia N, Shin DM, Cohen RB, et al. Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck. New England Journal of Medicine. 2006; 354(6): 567-78.

23. Guyot P, Ades AE, OUwens MJNM, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published kaplan-Meier survival curves. BMC Medical Research Methodology. 2012; 12.

24. Piccirillo JF, Costas I, Reichman ME. SEER survival monograph: cancer survival among adults: U.S. SEER program, 1988-2001, patient and tumor characteristics. Chapter 2 cancers of the head and neck. National Cancer Institute, SEER program, NIH Pub. No. 07-6215, Bethesda, MD, 2007.

25. Piccirillo JF, Costas I. SEER survival monograph: cancer survival among adults: U.S. SEER program, 1988-2001, patient and tumor characteristics. Chapter 8 cancer of the larynx. National Cancer Institute, SEER program, NIH Pub. No. 07-6215, Bethesda, MD, 2007.

26. Spiegelhalter DJ, Best NG, Carlin BR, van der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B – Statistical Methodology. 2002; 64:583-616.

27. Lunn D. WinBUGS Development Interface (WBDev) Last Updated in September 2004. Available from http://www.winbug-development.org.uk/wbdev.html. [Accessed january 1, 2010].

28. ReliaSoft. Life data analysis reference book. Last access in June 2013. Available from http://reliawiki.org/index.php/Life_Data_Analysis_Reference_Book. [Accessed january 1, 2010].

29. Blanchard P, Baujat B, Holostenci V, Bourredjem A, Baey C, Bourhis J, et al. Meta-analysis of chemotherapy in head and neck cancer (MACH-NC): a comprehensive analysis by tumour site. Radiotherapy and oncology. 2011; 100(1):33-40.

30. Fuller CD, Wang SJ, Thomas CR Jr., Hoffman HT, Weber RS, Rosenthal DI. Conditional survival in head and neck squamous cell carcinoma – Results from the SEER dataset 1973-1998. Cancer. 2007; 109(7):1331-43.

31. Van der Schroeff MP, van de Schans SAM, Piccirillo JF, Langeveld TPM, de Jongg RJB, Janssen-Heijnen MLG. Conditional relative survival in head and neck squamous cell carcinoma: permanent excess mortality risk for long-term survivors. Journal for the Sciences and Specialties of the Head and Neck. 2010; 32(12): 1613-8.

32. Guidance on cancer services. Improving outcomes in head and neck cancers. The manual. National Institue for Clinical Excellence. 2004. Available from:

http://www.nice.org.uk/nicemedia/pdf/csghn_the manual.pdf. [Accessed january 1, 2010].

33. Hashibe M, Brennan P, Benhamou S, Castellsague X, Chu C, Curado MP, et al. Alcohol drinking in never users of tobacco, cigarette smoking in never drinkers, and the risk of head and neck cancer: pooled analysis in the international head and neck cancer epidemiology consortium. Journal of the National Cancer Institute. 2007; 99(10):777-89.

34. MCKaig RG, Baric RS, Olshan AF. Human papillomavirus and head and neck cancer: epidemiology and molecular biology. Head & neck. 1998; 0(3):250-65.

35. Brenner H, Gefeller O, Hakulinen T. Period analysis for 'up-to-date' cancer survival data: theory, empirical evaluation computational realisation and applications. European Journal of Cancer. 2004; 40(3):326-35.

36. Incidence – SEER 18 Regs Research Data + Hurricane Katrina Impacted Lousiana Cases, Nov 2011 Sub (1973-2009 varying) – Linked to County Attributes – Total U.S., 1969-2010 Counties. Surveillance Research Program National Cancer Institue. Available at: SEER*Stat software (www.seer.cancer.gov/seerstat) version 8.0.2.

37. Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic research (Germany). Data downloaded on 17/04/2012. Available from www.mortality.org or www.humanmortality.de. [Accessed january 1, 2010].

38. Baselga J. The EGFR as a target for anticancer therapy – focus on cetuximab. European Journal of Cancer. 2001; 37:S16-S22.

39. Pignon JP, Bourhis J, Domenge C, Designe L, MACH-NC Collaborative Group. Chemotherapy added to locoregional treatment for head and neck squamous cell-carcinoma: three meta-analyses of updated individual data. Lancet. 2000; 355(9208):949-55.

40. Pignon JP, le Maitre A, Maillard E, Bourhis J, MACH-NC Collaborative Group. Meta-analysis of chemotherapy in head and neck cancer (MACH-NC): an update on 93 randomised trials and 17,346 patients. Radiotherapy and Oncology. 2009; 92(1):4-14.

41. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling an destimation of treatment effects. Statistics in Medicine. 2002; 21(15):2175-97.

42. Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual Version 1.4. January 2003. Available from http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf.

43. Jansen JP. Network meta-analysis fo survival data with fractional polynomials. BMC Medical Research Methodology. 2011; 11.

44. Erisoglu U, Erisoglu M, Erol H. Mixture model approach to the analysis of heterogeneous survival data. Pakistand Journal of Statistics. 2012; 28(1):115-30.

45. De Angelis R, Capocaccia R, Hakulinen T, Soderman B, Verdecchia A. Mixture models for cancer survival analysis: application to population-based data with covariates. Statistics in Medicine. 1999; 18(4):441-54.

46. Lambert PC, Dickman PW, Weston CL, Thompson JR. Estimating the cure fraction in population-based cancer studies by using finite mixture models. Journal of the Royal Statistical Society Series C-Applied Statistics. 2010; 59:35-55.

47. Griffin S, Walker S, Sculpher M, White S, Erhorn S, Brent S, et al. Cetuximab plus radiotherapy for the treatment of locally advanced squamous cell carcinoma of the head and neck cancer. The evidence review group's report. Available from:
http://nice.org.uk/nicemedia/pdf/HeadandNeckCancer_cetuximab_ERG.pdf. [Accessed january 1, 2010].

48. Ades AE, Sutton AJ. Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. Journal of the Royal Statistical Society A. 2006; 169, Part 1:5-35.

49. Eddy DM, Hasselblad V, Shachter R. Meta-Analysis by the Confidence Profile Method. Academic Press, London; 1992

50. Nelson CP, Lambert PC, Squire IB, Jones DR. Flexible parametric models for relative survival, with application in coronary heart disease. Statistics in Medicine. 2007; 26(30): 5486-98.

51. Janssen-Heijnen MLG, Gondos A, Bray F, Hakulinen T, Brewster DH, Brenner H, et al. Clinical relevance of conditional survival of cancer patients in Europe: age-specific analyses of 13 cancers. Journal of clinical oncology. 2010; 28(15): 2520-8.

52. Janssen-heijnen MLG, houterman S, Lemmens VEPP, Brenner H, Steyerberg EW, Coebergh JWW. Prognosis for long-term survivors of cancer. Annals of Oncology. 2007; 18(8):1408-13.

53. Merill RM, Hunter BD. Conditional survival among cancer patients in the United States. Oncologist. 2010; 15(8):873-82.

54. Gloeckler RLA, Reichman ME, Lewis DR, hankey BF, Edwards BK. Cancer survival and incidence from the Surveillance, Epidemiology and End Results (SEER) program. The oncologist. 2003; 8(6):541-52.

55. Royston P, Lambert PC. Flexible parametric survival analysis using stata: beyond the Cox model. Press S, editor 2011.

56. Atkinson KE. An introduction to numerical analysis. 2nd edition. Sons JW, editor 1989.

Table 1. Model fit statistics (posterior mean deviance, $\bar{D}$, and Deviance Information Criteria, DIC) and estimated differences in life expectancy between the 2 arms of Bonner 2006 RCT for different survival models, with and without external data

	Total $\bar{D}$	Total DIC	Gain in life expectancy due to cetuximab (months)	
			Point estimate	95% Cr I
Standard parametric models, no external data				
2-parameter Gamma FSEA	2,343	2,345	21.0	[2.6; 44.5]
2-parameter Gamma AFT	2,342	2,345	18.0	[1.7; 36.9]
Weibull FSEA	2,341	2,344	23.3	[0.7; 54.5]
Weibull PH	2,341	2,343	19.4	[1.6; 40.9]
Exponential FSEA	2,342	2,342	17.0	[2.1; 33.4]
Exponential PH	2,342	2,343	17.0	[2.0; 33.4]
Log-logistic FSEA	2,322	2,325	195.5	[-6895.0; 6860.0]
Log-logistic AFT	2,322	2,325	82.5	[-5.7; 487.8]
Log-normal FSEA	2,311	2,314	80.4	[2.0; 237.0]
Log-normal AFT	2,313	2,316	32.3	[-3.1; 78.6]
Generalised Gamma FSEA	2,308	2,313	50.9	[-19.2; 179.4]
Generalised Gamma AFT	2,310	2,313	13.9	[-4.3; 48.2]
Standard parametric models with external data				
Log-normal AFT, with general population data	2,327	2,329	32.4	[14.11 ; 55.73]
Generalised Gamma FSEA with general population data	2,317	2,322	31.7	[-26.4; 162.3]
Spline model with external data				
Splines with general population survival, conditional and relative treatment effect	2,303	2,306	4.7	[0.4; 9.1]

CrI: Credible Interval; DIC: Deviance Information Criterion; FSEA: Fitted Separately to Each Arm; PH: Proportional Hazards; AFT: Accelerated Failure Time.

For each parameter, or log parameter if the parameter was required by definition to be positive, the prior was assumed to follow a normal distribution with a mean of 0 and a variance of 1000.

Table 2. Global goodness of fit statistics (posterior mean deviance, $\bar{D}$, and Deviance Information Criteria, DIC) for the internal and external data elements.

	RCT data		SEER data		General population data		Relative treatment effect	
	$\bar{D}$	DIC	$\bar{D}$	DIC	$\bar{D}$	DIC	$\bar{D}$	DIC
Log logistic AFT using Equation [1]	2,327	2,329	NA	NA	10	11	NA	NA
Gen Gamma FSEA using Equation [1]	2,317	2,322	NA	NA	10	11	NA	NA
Splines, no external data used	2,300	2,307	NA	NA	NA	NA	NA	NA
Splines using Equation [1]	2,300	2,307	NA	NA	10	11	NA	NA
Splines using Equation [2]	2,302	2,307	96	98	NA	NA	NA	NA
Splines using Equation [2] and Equation [3]	2,302	2,307	96	97	11	12	NA	NA
Splines using Equation [2], Equation [3] and Equation[4]	2,303	2,306	97	98	11	12	-62	-60

Figure 1. FSEA and AFT Log-normal models compared with Kaplan-Meier curves

KM: Kaplan-Meier; FSEA: Fitted Separately to Each Arm; PH: Proportional Hazards; AFT:

Accelerated Failure Time.
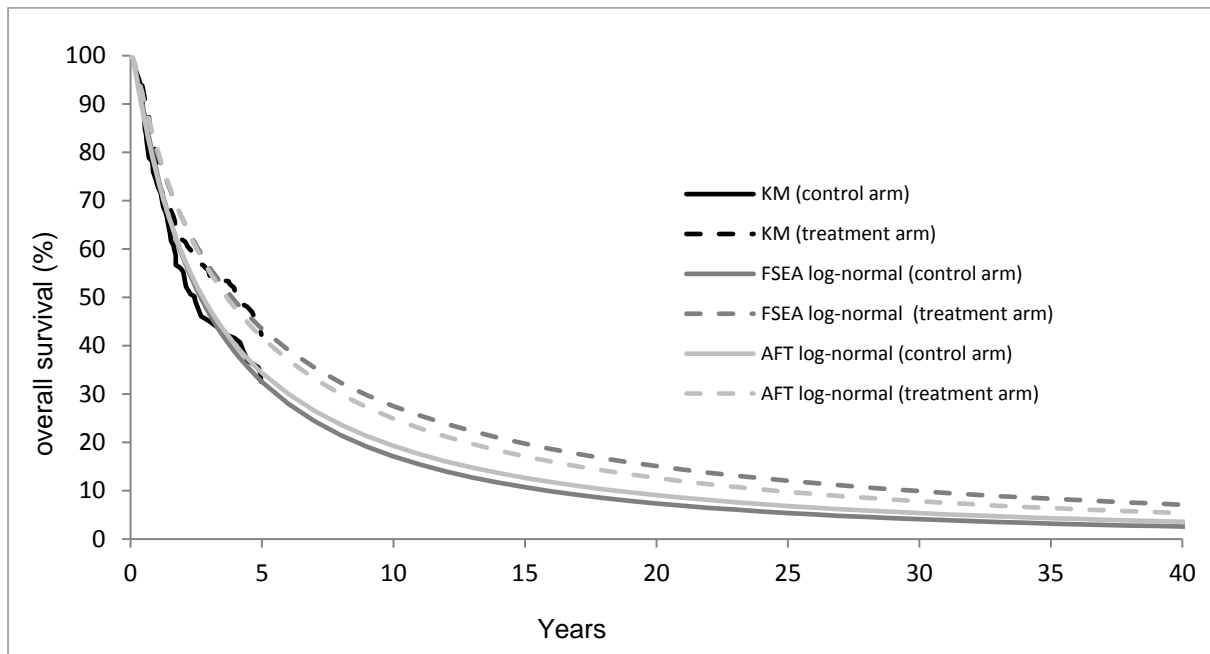
Figure 2. 1-year conditional survival predicted in Bonner 2006 and in the Surveillance, Epidemiology, and End Results (SEER) database (A). 1-year conditional survival predicted in Bonner 2006 and in Pignon 2009 (B).
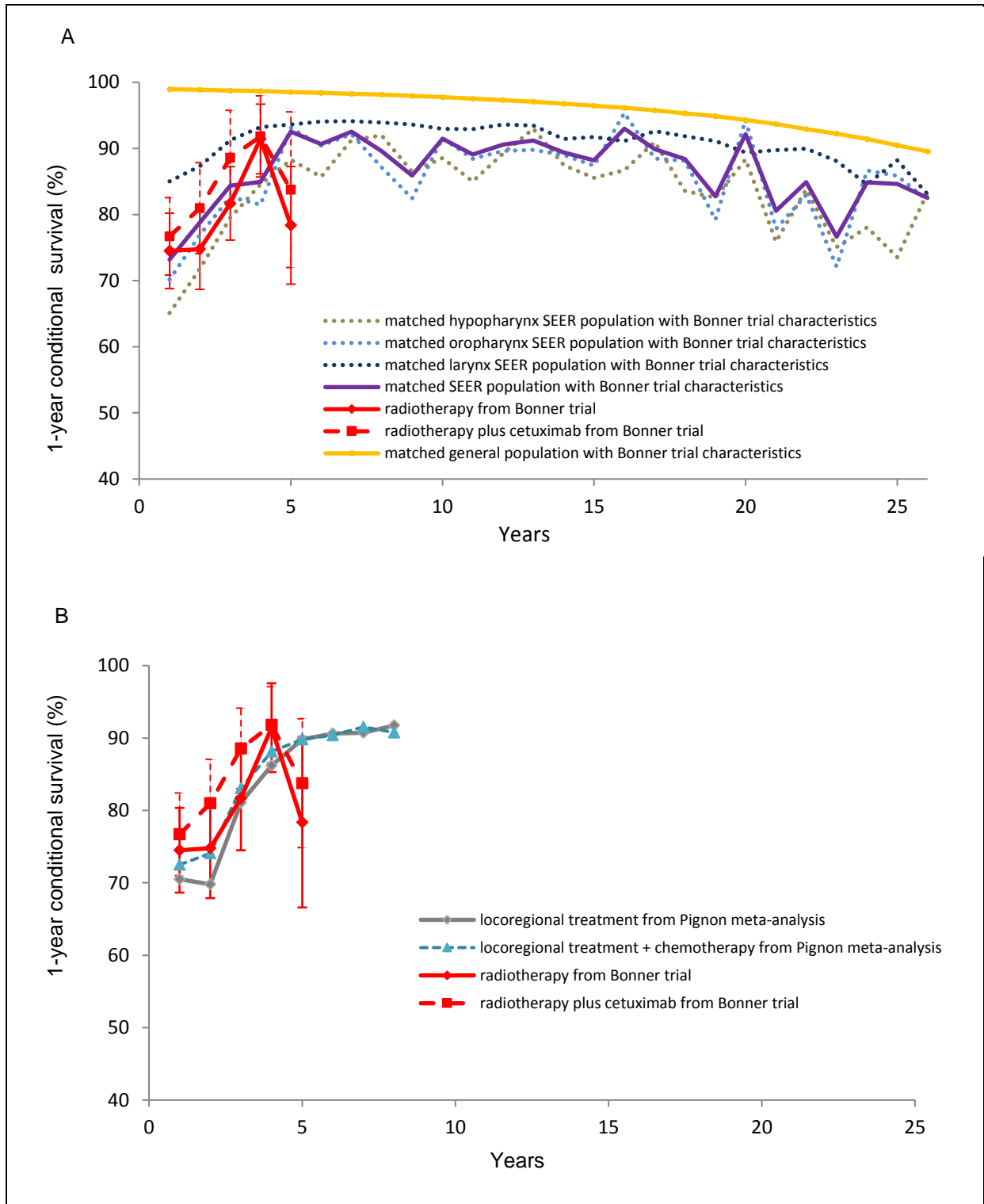
Figure 3. Overall survival predicted by KM, unconstrained AFT Log-normal, and AFT Log-normal constrained by general population data, using Equation [1] (A). Overall survival predicted by KM, unconstrained FSEA Generalized Gamma, and FSEA Generalized Gamma constrained by general population data, using Equation [1] (B).

KM: Kaplan-Meier; FSEA: Fitted Separately to Each Arm; AFT: Accelerated Failure Time.
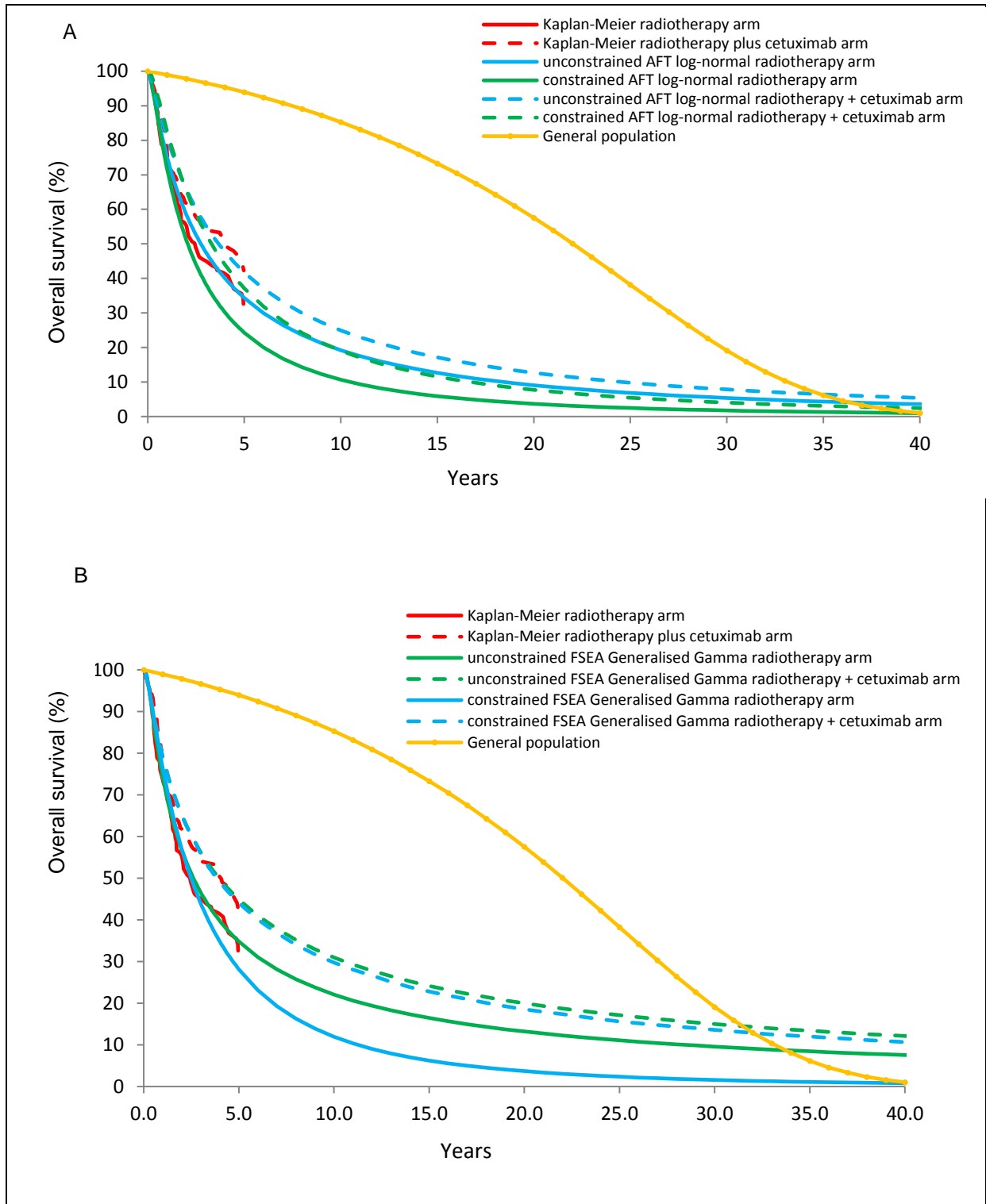
Figure 4. Overall survival predicted by KM, unconstrained splines, and splines constrained by general population data using Equation [3], SEER conditional survival data using Equation [2] and expert data using Equation [4] (A). 1-year conditional survival (B). Hazard ratio (C).

KM: Kaplan-Meier; FSEA: SEER: Surveillance, Epidemiology, and End Results
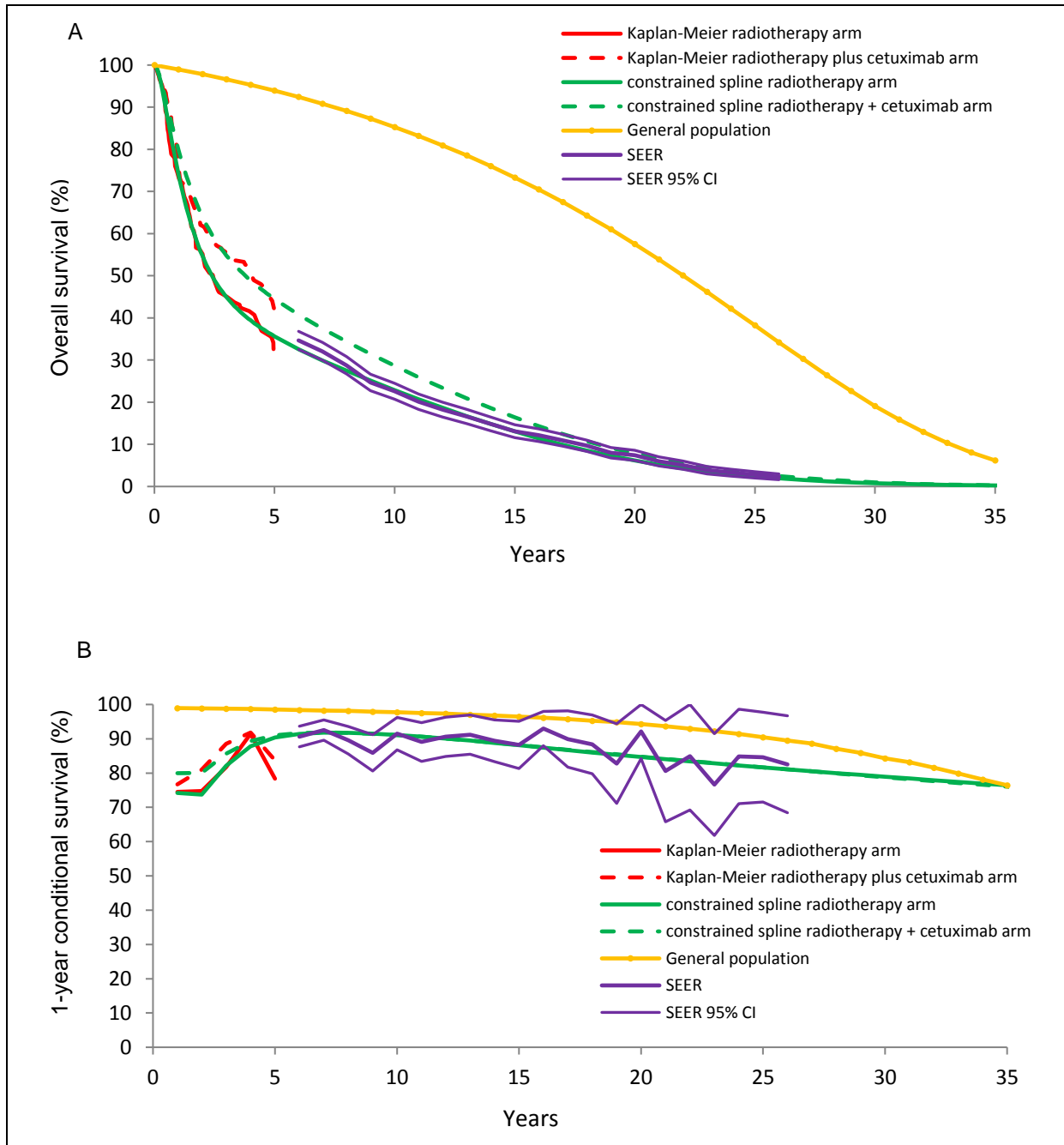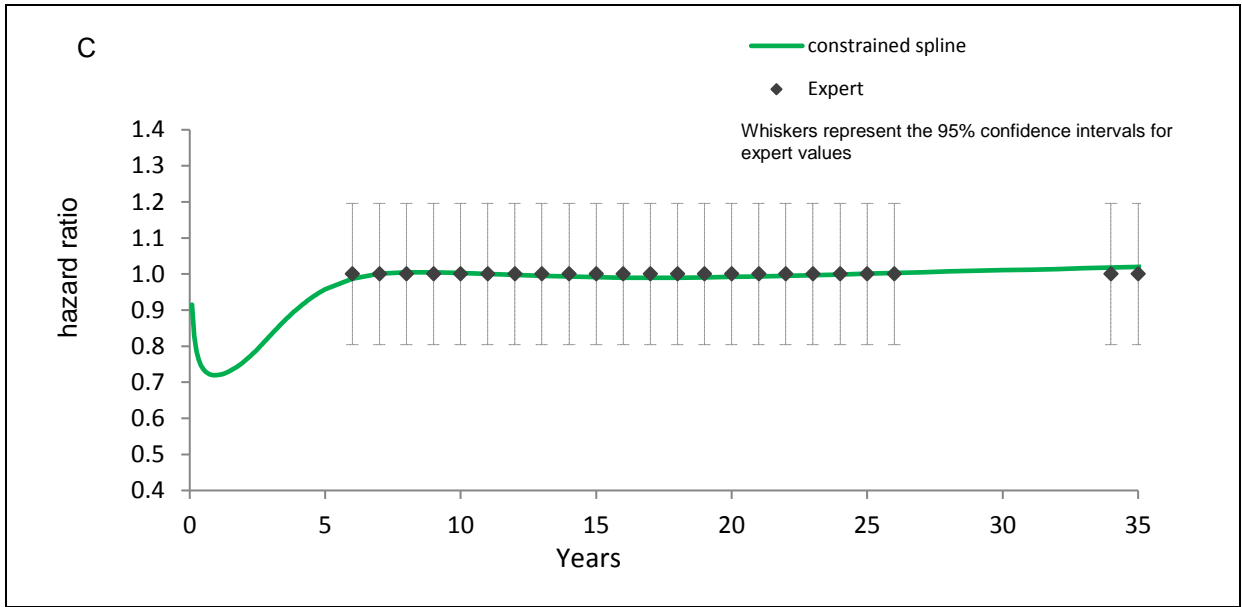
Web Appendix

A: Construction of a general population matched to the Bonner RCT population

General population survival data are available in the human mortality database (37). This database provides the number of persons exposed to the risk of death, the number of deaths and the probabilities of deaths, by age, from 0 to 109 years, gender, for years 1933-2009 and many countries.

The objective is to generate a general population cohort matched to the trial cohort for age at the onset of the trial and gender that differs from the trial population *only* in that it does not have head and neck cancer. P_{a_k}, P_g and P_c are the proportions of RCT patients in each age group a_k, gender g and country c. These figures are available in the Bonner publication. We assume that these factors are independent.

For age we used the published mean age and standard deviation, and divided the age into quartiles. The proportion of patients entering the trial at each of the N-quantiles defined is then set to 1/(N-1).

We compose our matched general population with survival data using the same calendar year than the study entry time of the RCT. For instance, if the study recruitment has been initiated in April 1999, the survival data for each wanted subgroup of the general population is extracted for the year 1999. By doing this, we perform a period analysis instead of a cohort analysis; in order to provide more sensible predicted estimates (27).

The average number of years lived during year t, for people dying during that year, is assumed to be 0.5. Using this half-cycle correction, the numbers of person alive between times t and t+1 for each subgroup G of persons having the same characteristics, $r_{general,t+1|t,G}$, are equal to the following:

$$r_{general,G,t} = ExposedtoRisk_{general,G,t} - 0.5 * Deaths_{general,G,t}$$

.

Where $ExposedtoRisk_{general,G,t}$ is the number of patients exposed to the risk of dying at time t for the subgroup G and $Deaths_{general,G,t}$ is the number of deaths between time t and t+1 for the subgroup G.

We refer to 1-year conditional survival probabilities: an interval-specific measure that gives the chance of surviving at time $t+1$, conditional on having survived to time $t : CS(t+1|t) = S(t+1|t)/S(t)$, whereby $S(t)$ denotes the probability of survival up to time t > 0.

The survival probabilities at time t+1 years conditional on being alive at time t years, in the matched general population, $CS_{general,t+1|t}$, are calculated as follows:

$$CS_{general,t+1|t} = \sum_{G} CS_{general,G,t+1|t} * W_{G}$$

Where $CS_{general,G,t+1|t}$ is the survival probabilities at time t+1 years conditional on being alive at time t years in the subgroup G and W_{G} the weight of the subgroup G.

Their corresponding variances are equal to the equation below:

$$V(CS_{general,t+1|t}) = \sum_{G} \frac{CS_{general,G,t+1|t} * (1 - CS_{general,G,t+1|t})}{r_{general,G,t+1|t}} * W^2$$

The number of persons alive during the entire period, $r_{general,t+1|t}$ and the number of persons at risk, $n_{general,t+1|t}$, between times t and t+1, in the matched general population, are obtained by the following equations:

$$n_{general,t+1|t} = \frac{CS_{general,t+1|t} * (1 - CS_{general,t+1|t})}{V(CS_{general,t+1|t})} .$$

$$r_{general,t+1|t} = CS_{general,t+1|t} * n_{general,t+1|t}$$

Carrying this out generates a numerator of 1660 survivors at *t*=40

$$r_{general,40|39} = 1,660$$

$$S_{general,t+1} = CS_{general,t+1|t} * S_{general,t}$$

Carrying this out generates a survival probability of 1.045% at t=40

$$n_{general,t+1} = r_{general,t+1|t} / S_{general,t+1}$$

Carrying this out generates a denominator of 158,858 at risk at t=0

$$n_{general,0} = 158,858$$

B. Construction of a head and neck cancer cohort matched to the Bonner trial population

Relative conditional survival probabilities and numbers at risk were extracted from the SEER database for each of the primary cancer sites reported in the Bonner RCT. Then, the relative conditional survival probabilities for each site were multiplied by the age- and gender-matched general conditional survival to obtain conditional survival predictions, and finally pooled to form a site-weighted average using the cancer site information in the RCT. A period analysis of 5 years (1997-2002) was applied to extract data from the SEER database. Data were available from 1970 onwards; as a consequence, a follow-up of 26 years was available. We did not adjust for stage or treatment, as these factors were less often recorded in the database than primary sites. Figure 2A shows the 1-year conditional survival probabilities obtained for each primary site and for a primary site-matched SEER population. 1-year conditional survival increases until around 5-years, and then stabilizes before slowly falling. RCT results are consistent with this trend.

The conditional survival data is presented in Table B.1.

Table B.1. Matched SEER conditional survival data.

Time t in years from randomization	Numerator	Denominator

6	325	358
7	285	308
8	198	221
9	143	167
10	122	133
11	105	117
12	91	100
13	86	94
14	87	98
15	75	85
16	93	101
17	47	52
18	47	54
19	34	41
20	40	44
21	22	27
22	17	20
23	24	31
24	22	26
25	25	29
26	23	28

For the analyses including both SEER and general population conditional survival information, the general population numerator at 34 years was 23,080, and the denominator 30,196: representing an estimated 76.4% 1-year conditional survival.

C. Cubic Splines

We fit a cubic spline model to the log-cumulative hazard, where all parameters except the intercept, α, depend on treatment k (1=treatment, 0 = control). There is therefore a spline for the control arm, and an additional spline for the relative treatment effect on the log-cumulative-hazard scale. All patients are assumed to start the RCT with the same risk of death.

$$x = \ln(t), \quad t \text{ in months}$$

$$\ln(H_k(t)) = s(x, \alpha, \gamma_k) = \alpha + \gamma_{k,1}x + \gamma_{k,2}v_1(x) + \ldots + \gamma_{k,m+1}v_m(x) \text{ with } k = 0 \text{ for the control arm and}$$

$k = 1$ for the treatment arm

The j- th basis function is defined for $j = 1,...,m$ as

$$v_j(x) = (x - L_j)_+^3 - \lambda_j(x - L_{min})_+^3 - (1 - \lambda_j)(x - L_{max})_+^3$$

and $\lambda_j = \dfrac{L_{max} - L_j}{L_{max} - L_{min}}$

$$(x - a)_+ = \max(0, x - a).$$

Boundary knots, L_{min} and L_{max}, and internal knots, L_1 and L_2 were placed at the same location for each arm. L_{min} was set equal to the log-time of the first death in the RCT data, i.e. log(1)=0. L_{max} was set equal to the log-time of the last assumed death in the general population data, 60 years, log(720)=6.58 (time in months). The first internal knot k_1 was placed in the middle of the RCT period i.e. log(30)=3.39 and the second internal knot k_2 was chosen approximately half way through the external data period on the log scale i.e. log(187)=5.23.

D. Winbugs code and data

The cubic splines model was defined as in Royston-Lambert 2011 [55].

 # Winbugs model

model {

##Definition of the knots

lambda1<-(kmax-k1)/(kmax-kmin)
lambda2<-(kmax-k2)/(kmax-kmin)

##Cubic splines model

C<-10

for (i in 1:N) {
x[i]<-log(t[i])

##Definition of the splines variables
rcs1[i]<-x[i]
rcs2[i]<-pow(max(0,x[i]-k1),3)-lambda1*pow(max(0,x[i]-kmin),3)-(1-lambda1)*pow(max(0,x[i]-kmax),3)
rcs3[i]<-pow(max(0,x[i]-k2),3)-lambda2*pow(max(0,x[i]-kmin),3)-(1-lambda2)*pow(max(0,x[i]-kmax),3)

##Definition of the derivatives of the splines variables
drcs1[i]<-1
drcs2[i]<-3*pow(max(0,x[i]-k1),2)-3*lambda1*pow(max(0,x[i]-kmin),2)-3*(1-lambda1)*pow(max(0, x[i]-kmax),2)
drcs3[i]<-3*pow(max(0,x[i]-k2),2)-3*lambda2*pow(max(0,x[i]-kmin),2)-3*(1-lambda2)*pow(max(0, x[i]-kmax),2)

Restricted cubic splines functions are not listed under the standard distributions in Winbugs. However, the "zeros trick" describes in the Winbugs manual allows arbitrary sampling distributions to be used

zeros[i]<-0

eta[i] is the linear predictor which includes the intercept gamma[1] and 3 further parameters gamma[2], gamma[3], gamma[4] for the control arm of the 3 spline variables rcs1, rcs2 and rcs3 and 3 further parameters beta[1], beta[2], beta[3] for the treatment arm of the 3 spline variables of rcs1, rcs2 and rcs 3.

eta[i]<-gamma[1]+gamma[2]*rcs1[i]*(1-treat[i])+gamma[3]*rcs2[i]*(1-treat[i])+gamma[4]*rcs3[i]*(1-treat[i])+beta[1]*treat[i]*rcs1[i]+beta[2]*treat[i]*rcs2[i]+beta[3]*treat[i]*rcs3[i]

d.spi[i] is the linear predictor for the derivative of the spline variables. This has the same parameters as in the linear predictor but uses the derivative variables drcs1, drcs2 and drcs3.

d.sp[i]<-gamma[2]*drcs1[i]*(1-treat[i])+gamma[3]*drcs2[i]*(1-treat[i])+gamma[4]*drcs3[i]*(1-treat[i])+beta[1]*treat[i]*drcs1[i]+beta[2]*treat[i]*drcs.treat2[i]+beta[3]*treat[i]*drcs.treat3[i]

lnL[i] is the log-likelihood
lnL[i]<- -(d[i]*log(max((1/t[i])*d.sp[i]*exp(eta[i]), 0.00001))-exp(eta[i])) + C

zeros[i]~dpois(lnL[i])

```
##Predicted survival of trial population at times of external data
for (i in 1:24) {
```

##Definition of the splines variables at times of external data

```
x.RCT.Text[i]<-log(T.ext[i])
rcs1.RCT.Text[i]<-x.RCT.Text[i]
rcs2.RCT.Text[i]<-pow(max(0,x.RCT.Text[i]-k1),3)-lambda1*pow(max(0,x.RCT.Text[i]-kmin),3)-(1-
lambda1)*pow(max(0,x.RCT.Text[i]-kmax),3)
rcs3.RCT.Text[i]<-pow(max(0,x.RCT.Text[i]-k2),3)-lambda2*pow(max(0,x.RCT.Text[i]-kmin),3)-(1-
lambda2)*pow(max(0,x.RCT.Text[i]-kmax),3)
```

##Definition of the derivatives of the splines variables at times of external data

```
drcs1.RCT.Text[i]<-1
drcs2.RCT.Text[i]<-3*pow(max(0,x.RCT.Text[i]-k1),2)-3*lambda1*pow(max(0,x.RCT.Text[i]-kmin),2)-
3*(1-lambda1)*pow(max(0, x.RCT.Text[i]-kmax),2)
drcs3.RCT.Text[i]<-3*pow(max(0,x.RCT.Text[i]-k2),2)-3*lambda2*pow(max(0,x.RCT.Text[i]-kmin),2)-
3*(1-lambda2)*pow(max(0, x.RCT.Text[i]-kmax),2)
```

##Definition of the linear predictors at times of external data for each arm, c=control, t=treatment

```
eta.RCT.Text.c[i]<-gamma[1] + gamma[2]*rcs1.RCT.Text[i] + gamma[3]*rcs2.RCT.Text[i]+
gamma[4]*rcs3.RCT.Text[i]
eta.RCT.Text.t[i]<-gamma[1] + beta[1] *rcs1.RCT.Text[i]+beta[2]*rcs2. RCT.Text[i]+
beta[3]*rcs3.RCT.Text[i]
```

##Definition of the linear predictors at times of external data for each arm, c=control, t=treatment

```
d.sp.RCT.Text.c[i]<-
gamma[2]*drcs1.RCT.Text[i]+gamma[3]*drcs2.RCT.Text[i]+gamma[4]*drcs3.RCT.Text[i]
d.sp.RCT.Text.t[i]<-beta[1]*drcs1.RCT.Text[i]+beta[2]*drcs 2.RCT.Text[i]+beta[3]*drcs3.RCT.Text[i]
```

##Definition of the hazard functions at times of external data for each arm, c=control, t=treatment

```
h.RCT.Text.c[i]<-1/T.ext[i]*d.sp.RCT.Text.c[i]*exp(eta.RCT.Text.c[i])
h.RCT.Text.t[i]<-1/T.ext[i]*d.sp.RCT.Text.t[i]*exp(eta.RCT.Text.t[i])
```

##Definition of the survival functions at times of external data for each arm, c=control, t=treatment

```
S.RCT.Text.c[i]<-max(0.00001,min(exp(-exp(eta.RCT.Text.c[i])),0.99999))
S.RCT.Text.t[i]<-max(0.00001,min(exp(-exp(eta.RCT.Text.t[i])),0.99999))
}

for (i in 1:21)
{
```

##Definition of the conditional survival function in the control arm at times of external data
```
CS.RCT.c[i]<-max(0.00001,min(S.RCT.Text.c[i+1]/S.RCT.Text.c[i],0.99999))
```

##Model for the SEER conditional survival data
```
r.b[i] ~ dbin (CS.ext.b[i],n.b[i])
```

Constraint using the SEER conditional survival data
```
CS.ext.b[i]<-max(0.00001,min(CS.RCT.c[i],0.99999))
}
```

```
for (i in 1:1){

##Model for the general population conditional survival data
r.GP[i] ~ dbin (CS.GP[i],n.GP[i])

##Constraint using the general population conditional survival data
CS.RCT.c[i+21]<-max(0.00001,min(S.RCT.Text.c[(24)]/S.RCT.Text.c[(23)],0.99999))CS.GP[i]<-
max(0.00001,min(CS.RCT.c[i+21],0.99999))
}

for (i in 1:23)
{

##Model for the external hazard ratio data
HR.ext.m[i]~dnorm(HR[i],HR.prec[i])

##Constraint using the external hazard ratio data
HR[i]<-h.RCT.Text.t[i+1]/h.RCT.Text.c[i+1]
}

#Priors for the 4 gamma parameters
for (j in 1:4){
gamma[j]~dnorm(0,0.01)
}

#Priors for the 3 beta parameters
for (j in 1:3) {
beta[j]~dnorm(0,0.01)
}

# To compute results
for (i in 1:68) {

x.pred[i]<-log(t.annual[i])
rcs1.pred[i]<-x.pred[i]
rcs2.pred[i]<-pow(max(0,x.pred[i]-k1),3)-lambda1*pow(max(0,x.pred[i]-kmin),3)-(1-
lambda1)*pow(max(0,x.pred[i]-kmax),3)
rcs3.pred[i]<-pow(max(0,x.pred[i]-k2),3)-lambda2*pow(max(0,x.pred[i]-kmin),3)-(1-
lambda2)*pow(max(0,x.pred[i]-kmax),3)

drcs1.pred[i]<-1
drcs2.pred[i]<-3*pow(max(0,x.pred[i]-k1),2)-3*lambda1*pow(max(0,x.pred[i]-kmin),2)-3*(1-
lambda1)*pow(max(0, x.pred[i]-kmax),2)
drcs3.pred[i]<-3*pow(max(0,x.pred[i]-k2),2)-3*lambda2*pow(max(0,x.pred[i]-kmin),2)-3*(1-
lambda2)*pow(max(0, x.pred[i]-kmax),2)

eta.pred.c[i]<-gamma[1]+gamma[2]*rcs1.pred[i]+gamma[3]*rcs2.pred[i]+gamma[4]*rcs3.pred[i]
eta.pred.t[i]<-gamma[1]+ beta[1] *rcs1.pred[i]+ beta[2]*rcs.treat2.pred[i]+ beta[3]*rcs.treat3.pred[i]

d.sp.pred.c[i]<-gamma[2]*drcs1.pred[i]+gamma[3]*drcs2.pred[i]+gamma[4]*drcs3.pred[i]
d.sp.pred.t[i]<-beta[1]*drcs1.pred[i]+beta[2]*drcs.treat2.pred[i]+beta[3]*drcs.treat3.pred[i]

h.c[i]<-1/t.annual[i]*d.sp.pred.c[i]*exp(eta.pred.c[i])
h.t[i]<-1/t.annual[i]*d.sp.pred.t[i]*exp(eta.pred.t[i])

# Predicted annual hazard ratio

HR.a[i]<-h.t[i]/h.c[i]

# Predicted annual survival
```

```
S.c[i]<-exp(-exp(eta.pred.c[i]))
S.t[i]<-exp(-exp(eta.pred.t[i]))
}
```

 # Predicted annual conditional survival

```
for (i in 2:68) {
CS.c[i]<-S.c[i]/max(S.c[i-1],0.000001)
CS.t[i]<-S.t[i]/max(S.t[i-1],0.00000001)
}
```

 # Predicted mean in each arm, c=control, t=treatment
```
mean.surv.c<-Mean3k(endL,nsubdiv,gamma[1],gamma[2],gamma[3],gamma[4],kmin,kmax,k1,k2)
mean.surv.t<-
Mean3k(endL,nsubdiv,gamma[1],beta[1],beta[2],beta[3],kmin,kmax.treat,k.treat1,k.treat2)
```

 # Predicted mean difference
```
diff.mean<-mean.surv.t-mean.surv.c


}
```


 # Data used


```
list(N=424,nsubdiv=1000, endL=900, kmin=0, kmax=6.58, k1=3.39, k2=5.23,
T.ext=c(60,72,84,96,108,120,132,144,156,168,180,192,204,216,228,240,252,264,276,288,300,312,
408, #Note this time-point is 1-year prior to 1st GP CS data-point
420),
r.GP=c(23080),
n.GP=c(30196),
n.b=c(358,308,221,167,133,117,100,94,98,85,101,52,54,41,44,27,20,31,26,29,28),
r.b=c(325,285,198,143,122,105,91,86,87,75,93,47,47,34,40,22,17,24,22,25,23),
HR.ext.m=c(1,1,1,1,1, 1,1,1,1,1, 1,1,1,1,1, 1,1,1,1,1, 1,1,1), HR.prec=c(100,100,100,100,100,
100,100,100,100,100, 100,100,100,100,100, 100,100,100,100,100, 100,100,100))
```

```
# IPD reconstructed from the KM curves
```

t[]	d[]	treat[]
1	1	0
1	1	0
1.93	1	0
1.93	1	0
2.54	1	0
2.54	1	0
2.54	1	0
3.77	1	0
3.77	1	0
3.77	1	0
4.31	1	0
4.31	1	0
4.47	1	0
4.93	1	0
4.93	1	0
5.31	1	0
5.31	1	0
5.31	1	0
5.54	1	0
5.54	1	0
5.54	1	0

5.54	1	0
5.54	1	0
6.24	1	0
6.24	1	0
6.39	1	0
6.39	1	0
6.39	1	0
6.39	1	0
6.62	1	0
6.62	1	0
6.62	1	0
7.01	1	0
7.01	1	0
7.32	1	0
7.32	1	0
7.32	1	0
7.85	1	0
7.85	1	0
7.85	1	0
8.39	1	0
8.39	1	0
8.39	1	0
8.39	1	0
10	1	0
10	1	0
10.3	1	0
10.3	1	0
10.3	1	0
10.3	1	0
10.7	1	0
11.4	1	0
11.4	1	0
12.3	1	0
12.3	1	0
12.3	1	0
12.8	1	0
13.8	1	0
13.8	1	0
14.3	1	0
14.3	1	0
14.3	1	0
15	1	0
15	1	0
15	1	0
15.9	1	0
15.9	1	0
16.6	1	0
16.6	1	0
17.1	1	0
17.1	1	0
17.1	1	0
17.6	1	0
17.6	1	0
18	1	0
18	1	0
18.4	1	0
18.4	1	0
18.4	1	0
19	1	0
19.6	1	0

20	1	0
20	1	0
20.6	1	0
20.6	1	0
20.7	1	0
20.7	1	0
20.7	1	0
20.7	1	0
22.6	1	0
23.4	1	0
24.1	1	0
24.9	1	0
24.9	1	0
24.9	1	0
24.9	1	0
25.3	1	0
25.3	1	0
27.4	1	0
27.4	1	0
27.4	1	0
29.1	1	0
29.6	1	0
29.6	1	0
30.9	1	0
30.9	1	0
30.9	1	0
31.7	1	0
31.7	1	0
32.2	1	0
34.4	1	0
36.3	1	0
40	1	0
40	1	0
42.6	1	0
43	1	0
47.4	1	0
50	1	0
51.4	1	0
51.4	1	0
52.5	1	0
53.5	1	0
58.4	1	0
59.5	1	0
59.4	1	0
1.465	0	0
2.235	0	0
3.425	0	0
5.12	0	0
6.315	0	0
7.585	0	0
8.89	0	0
12.55	0	0
14.65	0	0
17.35	0	0
23	0	0
26.85	0	0
30.45	0	0
31.3	0	0
31.95	0	0
33.3	0	0

33.3	0	0
33.3	0	0
33.3	0	0
35.35	0	0
35.35	0	0
35.35	0	0
37.3	0	0
37.3	0	0
37.3	0	0
37.3	0	0
39.15	0	0
39.15	0	0
40.55	0	0
40.55	0	0
41.85	0	0
41.85	0	0
41.85	0	0
42.8	0	0
45.05	0	0
45.05	0	0
45.05	0	0
45.05	0	0
45.05	0	0
45.05	0	0
45.05	0	0
45.05	0	0
45.05	0	0
47.25	0	0
48.7	0	0
48.7	0	0
48.7	0	0
48.7	0	0
48.7	0	0
50.9	0	0
50.9	0	0
51.75	0	0
52.3	0	0
52.8	0	0
53.3	0	0
55.8	0	0
55.8	0	0
55.8	0	0
55.8	0	0
55.8	0	0
55.8	0	0
55.8	0	0
55.8	0	0
58.25	0	0
58.95	0	0
59.45	0	0
62.55	0	0
62.55	0	0
62.55	0	0
62.55	0	0
62.55	0	0
62.55	0	0
65.1	0	0
65.1	0	0
65.1	0	0
65.1	0	0

65.1	0	0
65.1	0	0
65.1	0	0
65.1	0	0
65.1	0	0
65.1	0	0
65.1	0	0
65.1	0	0
65.1	0	0
65.1	0	0
65.1	0	0
65.1	0	0
0.417	1	1
1.11	1	1
1.46	1	1
1.63	1	1
2.32	1	1
2.32	1	1
2.67	1	1
3.02	1	1
3.36	1	1
3.54	1	1
4.06	1	1
5.1	1	1
5.27	1	1
5.62	1	1
5.79	1	1
5.96	1	1
6.31	1	1
6.31	1	1
6.48	1	1
7	1	1
7.18	1	1
7.18	1	1
7.18	1	1
7.7	1	1
8.22	1	1
8.39	1	1
8.39	1	1
8.74	1	1
8.74	1	1
8.74	1	1
8.91	1	1
9.08	1	1
9.6	1	1
10.1	1	1
10.1	1	1
10.3	1	1
10.6	1	1
10.6	1	1
10.8	1	1
11.2	1	1
11.3	1	1
11.3	1	1
11.5	1	1
11.5	1	1
11.9	1	1
11.9	1	1
11.9	1	1
12	1	1

12	1	1
12.4	1	1
12.4	1	1
12.4	1	1
12.4	1	1
12.4	1	1
12.5	1	1
12.7	1	1
13.2	1	1
13.2	1	1
14.3	1	1
14.5	1	1
15	1	1
16	1	1
16.2	1	1
17.2	1	1
17.6	1	1
18.1	1	1
18.4	1	1
18.8	1	1
19.3	1	1
19.7	1	1
20	1	1
20.3	1	1
20.3	1	1
20.5	1	1
20.9	1	1
22.1	1	1
22.4	1	1
22.8	1	1
22.9	1	1
24.7	1	1
25.4	1	1
25.9	1	1
26.9	1	1
27.8	1	1
28.5	1	1
29.5	1	1
29.5	1	1
30.2	1	1
31.1	1	1
33.2	1	1
34.2	1	1
35.6	1	1
35.8	1	1
36.1	1	1
36.3	1	1
44.8	1	1
45.3	1	1
47.2	1	1
47.6	1	1
47.7	1	1
48.9	1	1
49.5	1	1
53.3	1	1
55.5	1	1
56.7	1	1
58.8	1	1
59.7	1	1
5.01	0	1

43

11.95	0	1
14	0	1
15.9	0	1
18	0	1
22.5	0	1
24.95	0	1
27.4	0	1
31	0	1
31.9	0	1
32.75	0	1
33.6	0	1
34.5	0	1
35.5	0	1
36.4	0	1
38.35	0	1
38.35	0	1
38.35	0	1
40.65	0	1
40.65	0	1
40.65	0	1
42.8	0	1
42.8	0	1
42.8	0	1
42.8	0	1
42.8	0	1
42.8	0	1
42.8	0	1
42.8	0	1
42.8	0	1
42.8	0	1
42.8	0	1
44.9	0	1
45.2	0	1
45.4	0	1
45.9	0	1
46.1	0	1
46.4	0	1
46.8	0	1
47.1	0	1
47.5	0	1
47.8	0	1
48	0	1
48.3	0	1
48.7	0	1
49	0	1
49.4	0	1
49.7	0	1
50.4	0	1
50.6	0	1
51.8	0	1
51.8	0	1
51.8	0	1
51.8	0	1
51.8	0	1
51.8	0	1
51.8	0	1
53	0	1
53.35	0	1
53.7	0	1
53.9	0	1

54.2	0	1
54.6	0	1
54.9	0	1
55.1	0	1
55.45	0	1
55.8	0	1
56.1	0	1
56.3	0	1
56.65	0	1
57	0	1
57.2	0	1
57.5	0	1
57.9	0	1
58.2	0	1
58.4	0	1
58.7	0	1
59.1	0	1
59.4	0	1
59.6	0	1
60.65	0	1
63.65	0	1
63.65	0	1
63.65	0	1
63.65	0	1
63.65	0	1
63.65	0	1
63.65	0	1
66.3	0	1
66.3	0	1
66.3	0	1
66.3	0	1
66.3	0	1
66.3	0	1
66.3	0	1
66.3	0	1
66.3	0	1
66.3	0	1
66.3	0	1
66.3	0	1
66.3	0	1
66.3	0	1
66.3	0	1
66.3	0	1
END

Subroutine Mean3k called by the WinBUGS program above
Blue color represents changes from the blackbox module template

```
MODULE WBDevMean3k;

    IMPORT
        WBDevVector,
        WBDevScalar,
        Math,
        WBDevSpecfunc;

    TYPE
        Function = POINTER TO RECORD (WBDevScalar.Node) END;
```

```
                    Factory = POINTER TO RECORD (WBDevScalar.Factory) END;

           VAR
               fact-: WBDevScalar.Factory;

           PROCEDURE (func: Function) DeclareArgTypes (OUT args: ARRAY OF CHAR);
           BEGIN
```

#number of parameters=10
```
                args := "ssssssssss";
           END DeclareArgTypes;

           PROCEDURE (func: Function) Evaluate (OUT value: REAL);
           CONST
```

#Definition of the parameters
```
                timeupper = 0; nsubdiv = 1; gamma1 = 2; gamma2= 3; gamma3= 4; gamma4=5; kmin= 6;
kmax= 7; k1= 8;k2=9;

           VAR
```
#Definition of the variables
```
                i,j,n: INTEGER;
                t,h,a,b,c,kamin,kamax,ka1,x,lambda1, rcs1,
rcs2,eta,xka1,xkamin,xkamax,d,ka2,xka2,lambda2,rcs3: REAL;

           BEGIN
```
#Definition of the function: the area under the curve is computed using the Simpson's rule [60]. This
rule is a method for numerical integration that provides numerical approximation of definite integrals.

```
                a := func.arguments[gamma1][0].Value();
                b:= func.arguments[gamma2][0].Value();
                c:= func.arguments[gamma3][0].Value();
                d:= func.arguments[gamma4][0].Value();
                kamin:= func.arguments[kmin][0].Value();
                kamax:= func.arguments[kmax][0].Value();
                ka1:= func.arguments[k1][0].Value();
                ka2:= func.arguments[k2][0].Value();
                n:= SHORT(ENTIER(func.arguments[nsubdiv][0].Value()));
                t := func.arguments[timeupper][0].Value();
                i:=0;
                h:=t/n;
                value:=1;
                INC(i);
                j:=0;

                lambda1:=(kamax-ka1)/(kamax-kamin);
                lambda2:=(kamax-ka2)/(kamax-kamin);

                WHILE i < n DO;

                x:=Math.Ln(i*h);
                rcs1:=x;

                IF (x-ka1>0) THEN;
                xka1:=x-ka1;
                ELSE;
                xka1:=0;
                END;

                IF (x-ka2>0) THEN;
```

```
        xka2:=x-ka2;
        ELSE;
        xka2:=0;
        END;

        IF (x-kamin>0) THEN;
        xkamin:=x-kamin;
        ELSE;
        xkamin:=0;
        END;

        IF (x-kamax>0) THEN;
        xkamax:=x-kamax;
        ELSE;
        xkamax:=0;
        END;

        rcs2:=WBDevSpecfunc.Power(xka1,3)-lambda1*WBDevSpecfunc.Power(xkamin,3)-(1-
lambda1)*WBDevSpecfunc.Power(xkamax,3);

        rcs3:=WBDevSpecfunc.Power(xka2,3)-lambda2*WBDevSpecfunc.Power(xkamin,3)-(1-
lambda2)*WBDevSpecfunc.Power(xkamax,3);

        eta:=a+b*rcs1+c*rcs2+d*rcs3;


        IF (i=2*j+1) THEN;
            value:= value + 4*(Math.Exp(eta-Math.Exp(eta)));
                INC(j);
        ELSE;
            value:= value + 2*(Math.Exp(eta-Math.Exp(eta)));

        END;
        INC(i);

        END;

        x:=Math.Ln(t);
        rcs1:=x;

        IF ((x-ka1)>0) THEN;
        xka1:=x-ka1;
        ELSE;
        xka1:=0;
        END;

        IF (x-ka2>0) THEN;
        xka2:=x-ka2;
        ELSE;
        xka2:=0;
        END;

        IF ((x-kamin)>0) THEN;
        xkamin:=x-kamin;
        ELSE;
        xkamin:=0;
        END;

        IF ((x-kamax)>0) THEN;
        xkamax:=x-kamax;
```

```
            ELSE;
            xkamax:=0;
            END;

            rcs2:=WBDevSpecfunc.Power(xka1,3)-lambda1*WBDevSpecfunc.Power(xkamin,3)-(1-
lambda1)*WBDevSpecfunc.Power(xkamax,3);

            rcs3:=WBDevSpecfunc.Power(xka2,3)-lambda2*WBDevSpecfunc.Power(xkamin,3)-(1-
lambda2)*WBDevSpecfunc.Power(xkamax,3);

            eta:=a+b*rcs1+c*rcs2+d*rcs3;

            value:= value + Math.Exp(eta-Math.Exp(eta));

            value:= (t/n)*value/3

        END Evaluate;

        PROCEDURE (f: Factory) New (option: INTEGER): Function;
        VAR
            func: Function;
        BEGIN
            NEW(func); func.Initialize; RETURN func;
        END New;

        PROCEDURE Install*;
        BEGIN
            WBDevScalar.Install(fact);
        END Install;

        PROCEDURE Init;
        VAR
            f: Factory;
        BEGIN
            NEW(f); fact := f;
        END Init;

    BEGIN
        Init;
END WBDevMean3k.
```

E: Initial values, convergence, and posterior sampling

We assessed convergence by running multiple chains with different initial values, then inspecting history plots, density plots for key model outputs (survival estimates) and the gelman-rubin statistics.

For the model based on RCT data alone, we present results based on a burn-in period of 20,000 simulations, and posterior inference was based on 350,000 samples for the splines model, 150,000 for the AFT log-normal model, and 300,000 for the generalised gamma model.

For the constrained models, we present results based on a burn-in period of 20,000 simulations, and posterior inference was based on 70,000 samples for the splines model, 30,000 for the AFT log-normal model, and 60,000 for the generalised gamma model.

Convergence was problematic in some cases, and care was required in the choice of initial values to avoid the chain sampling from infeasible curves (see below).

Initial values

Spline function parameters of splines have no direct interpretation, making it difficult to choose a realistic set of initial values. Furthermore, due to the extremely flexible nature of the spline functions, convergence can be problematic if there is not sufficient data to determine the exact shape of the curves between the knots, and especially in the tail of the curves. This is essentially a problem of identifiability. We found that some choices of initial value found a solution that gave infeasible predicted survival curves, especially when we did not include all evidence sources. In such cases it is often suggested to use the Maximum Likelihood Estimate (MLE) to choose initial values [56], however due to the complexity of the relationships between our evidence and the model, it is not possible to evaluate the MLE. Instead we used the following strategy to identify realistic initial values:

1. Run the model with multiple chains of initial values.
2. Choose the chain with the minimum Dbar (i.e. the chains that fit the data best, and closest to MLE),
3. We then used descriptive statistics from the joint distribution from the chain with minimum Dbar to identify new sets of initial values (covering the full range of jointly plausible values)
4. Repeat steps 1-3 until Dbar seems to stabilise, i.e. when Dbar is not decreasing anymore after a reasonable number of attempts.

The sets of initial values identified are then used to check for convergence and to obtain results, as described above.