ScienceDirect

Contents lists available at **sciencedirect.com**
Journal homepage: **www.elsevier.com/locate/jval**

Comparison of Parametric Survival Extrapolation Approaches Incorporating General Population Mortality for Adequate Health Technology Assessment of New Oncology Drugs

Ilse van Oostrum, MSc, Mario Ouwens, PhD, Antonio Remiro-Azócar, MSc, Gianluca Baio, PhD, Maarten J. Postma, PhD, Erik Buskens, MD, PhD, Bart Heeg, PhD

ABSTRACT

Objectives: Survival extrapolation of trial outcomes is required for health economic evaluation. Generally, all-cause mortality (ACM) is modeled using standard parametric distributions, often without distinguishing disease-specific/excess mortality and general population background mortality (GPM). Recent National Institute for Health and Care Excellence guidance (Technical Support Document 21) recommends adding GPM hazards to disease-specific/excess mortality hazards in the log-likelihood function ("internal additive hazards"). This article compares alternative extrapolation approaches with and without GPM adjustment.

Methods: Survival extrapolations using the internal additive hazards approach (1) are compared to no GPM adjustment (2), applying GPM hazards once ACM hazards drop below GPM hazards (3), adding GPM hazards to ACM hazards (4), and proportional hazards for ACM versus GPM hazards (5). The fit, face validity, mean predicted life-years, and corresponding uncertainty measures are assessed for the active versus control arms of immature and mature (30- and 75-month follow-up) multiple myeloma data and mature (64-month follow-up) breast cancer data.

Results: The 5 approaches yielded considerably different outcomes. Incremental mean predicted life-years vary most in the immature multiple myeloma data set. The lognormal distribution (best statistical fit for approaches 1-4) produces survival increments of 3.5 (95% credible interval: 1.4-5.3), 8.5 (3.1-13.0), 3.5 (1.3-5.4), 2.9 (1.1-4.5), and 1.6 (0.4-2.8) years for approaches 1 to 5, respectively. Approach 1 had the highest face validity for all data sets. Uncertainty over parametric distributions was comparable for GPM-adjusted approaches 1, 3, and 4, and much larger for approach 2.

Conclusion: This study highlights the importance of GPM adjustment, and particularly of incorporating GPM hazards in the log-likelihood function of standard parametric distributions.

Keywords: additive hazards, general population mortality, incremental life-years, parametric modeling.

VALUE HEALTH. 2021; ■(■):■–■

Introduction

Cost-effectiveness analyses for health economic evaluations require estimating the difference in mean survival between competing interventions. Nevertheless, because data from randomized clinical trials (RCTs) are often immature at the time of the economic evaluation (eg, for both arms median survival is not reached), mean survival can only be estimated by extrapolating survival beyond the follow-up of the trial.[1,2] Guidance by local health authorities such as National Institute for Health and Care Excellence (NICE), Pharmaceutical Benefits Advisory Committee, and Canadian Agency for Drugs and Technologies in Health recommends extrapolating survival by fitting standard parametric distributions to patient-level data.[1-4] It is important that the considered distributions capture the hazard profile accurately, within both the observed trial follow-up period and the long-term extrapolation period. Therefore, recent NICE guidance (technical support document [TSD] 21, November 2020) recommends incorporating general population background mortality (GPM) in survival models or using general population survival data to examine the face validity of survival extrapolations.[5]

Incorporating GPM in survival models is especially important beyond the observed follow-up period of the clinical trial. During the trial, mortality is often predominantly disease-driven. Therefore, standard parametric extrapolations of clinical trial data will likely be disease-driven as well. During the extrapolation period, there is an increase in GPM hazards as patients become older. Therefore, GPM may explain a larger part of mortality than the disease itself in the long term. Standard parametric distributions based solely on the trial data may produce biased extrapolations

and poor cost-effectiveness estimates. This bias is most apparent among older populations, where the GPM hazards will rise owing to age-related mortality, and in the study of diseases or treatments with a large fraction of cured patients or long-term survivors, where disease-specific hazards captured by the standard parametric models may decrease to zero. Extrapolations that are purely trial-based are also problematic where the data are immature, in which case the hazard produced by standard parametric models may drop below the GPM hazard, either in the trial follow-up or in the extrapolation period.

To address these issues, Andersson et al[6] and the latest NICE TSD (TSD21)[5] propose decomposing all-cause mortality (ACM) into disease-specific/excess mortality (DSM) and GPM by adding GPM hazards in the log-likelihood function of a parametric distribution. This approach has been referred to as "relative survival modeling" or "excess mortality modeling."[5,6] We describe it as modeling additive hazards "internally" and refer to it as "internal additive hazards." Within the extended likelihood function, mortality unrelated to the disease is captured by GPM hazards and the parametric function hazards capture DSM. To our knowledge, so far, this method has not been applied in health technology assessments.[5]

Additional extrapolation adjustments that explicitly account for GPM are categorized by Jackson et al.[7] In the "converging hazards" approach, patients have a higher initial mortality compared to the general population, but this decreases until, at some point, the mortality rate of the patient population converges to GPM.[7-18] In the additive hazards approach, GPM hazards are added to the trial mortality hazards, implying that the disease-specific patient population has a constant additive excess hazard compared to the general population.[7,19-22] Finally, in the proportional hazards approach (also labeled as "standardized mortality risk"), the hazard ratio between the patient and general populations is constant over time.[7,23-25]

All these approaches can be used to estimate mean survival and are likely to produce different results. This is important for the computation of the incremental cost-effectiveness ratio (ICER), and thus for health technology assessment in general. The objective of this article is to compare the outcomes from different approaches, with and without GPM adjustment, to assess the importance of adjusting for GPM. We investigate whether the methodologies are appropriate for extrapolating survival in trials with different levels of data maturity, patient populations (older versus younger), and types of cancer (hematologic versus solid tumor). In line with NICE TSD21 recommendations for future research, we use a Bayesian modeling framework and incorporate external GPM information to the analysis of RCT data.[5] Five approaches are compared. Survival extrapolations under each approach are assessed by comparing predicted mean (incremental) life-years and corresponding uncertainty, statistical fit, face validity, and uncertainty over the use of different parametric distributions.

Methods

Selection, Description, and Preparation of the Data Sets

We consider 2 published data cuts of an RCT in multiple myeloma and 1 data cut of an RCT in patients with breast cancer.[26-28] These illustrate distinct types of scenarios, representing different cancer types (hematologic versus solid tumor) and different levels of data maturity. RCT data were used, because this is typically submitted to health technology assessment agencies.

For the multiple myeloma data set, 2 overall survival data cuts were available (30 and 75 months of follow-up), representing immature (70%-80% alive at end of follow-up) and mature (30%-40% alive at end of follow-up) data.[26,27] In the multiple myeloma phase 3 trial, 682 previously untreated patients, who were ineligible for high-dose therapy plus stem-cell transplantation, were randomized to receive melphalan and prednisone with or without bortezomib. Median survival was not reached at 30 months (first data cut) and was 56.4 in the bortezomib group versus 43.1 months in the control group at the 75-month data cut. The median age was 71 years (range: 48-91), and approximately 50% of the patients were male.

For the breast cancer data set, a mature progression-free survival data cut (64 months of follow-up with 10%-20% progression-free at the end of follow-up) was available.[28] In the breast cancer trial, 808 patients with human epidermal growth factor receptor 2-positive metastatic breast cancer were randomized to receive pertuzumab plus trastuzumab and docetaxel (pertuzumab group) or placebo plus trastuzumab and docetaxel (control group). The median age of the patients in the breast cancer trial was 54 years.

As individual patient-level data were unavailable for both of these trials, survival data were reconstructed by digitizing the published Kaplan-Meier curves using Engauge Digitizer.[29] Subsequently, individual patient-level data were generated in R based on the validated algorithm of Guyot et al.[30] Figure 1 presents the reconstructed Kaplan-Meier data for both trials. In Appendix Figure 1 (see Appendix Fig. 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008), the corresponding cumulative hazards plots are presented.

GPM Data

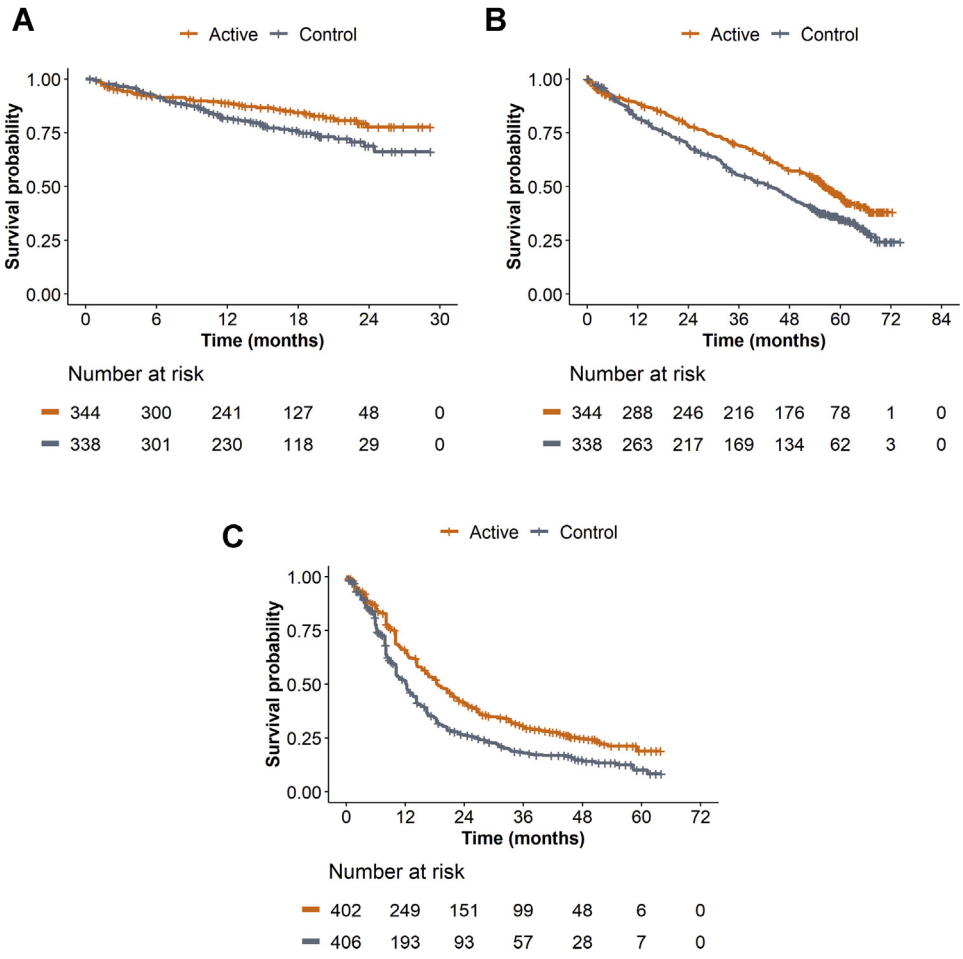
In our analysis, GPM data were sourced from life tables of the United States, which present the annual mortality rates observed within the general population of the country.[31] Appendix Figure 1 (in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008) presents the trial-specific GPM hazards, based on the median age and the proportion of men and women within the multiple myeloma and breast cancer trials.

Compared Approaches

The internal additive hazards approach[5,6] (approach 1), which decomposes ACM hazards into DSM and GPM hazards within the log-likelihood function of the parametric model, is compared with the use of parametric distributions; (2) without GPM adjustment; (3) adjusted for GPM once the fitted parametric ACM hazards drop below the GPM hazards (converging hazards);[7-18] (4) where GPM hazards are added to the fitted parametric ACM hazards (external additive hazards);[7,19-22] and (5) assuming proportional hazards for ACM versus GPM (proportional hazards).[7,23-25]

Table 1 presents the underlying hazard functions of the 5 approaches with and without GPM adjustment. The no GPM adjustment approach does not consider GPM mortality, and exclusively relies on the parameters of the standard parametric distributions fitted to the patient-level trial data. The converging hazards approach[7-18] has the following interpretation: patients are modeled as "cured," subject only to GPM, from the point at which the fitted parametric ACM hazards drop below the GPM hazards. The external additive hazards approach[7,19-22] assumes that the GPM hazards are negligible during the trial follow-up period. In the extrapolation period, GPM hazards increase and therefore only the extrapolation is based on the sum of GPM and the fitted parametric ACM hazards. When the patient population is older and GPM hazards are not negligible, this approach will overestimate the probability of an event during the trial follow-up

Figure 1. (A) Kaplan–Meier curves of the multiple myeloma data set (30-month data cut, overall survival). (B) Kaplan–Meier curves of the multiple myeloma data set (75-month data cut, overall survival). (C) Kaplan–Meier curves of the breast cancer data set (progression-free survival).



period. In the proportional hazards approach,[7,23-25] a hazard ratio is estimated by comparing the ACM hazards of patients versus the GPM hazards. In this approach, survival is extrapolated over a lifetime horizon by applying the estimated hazard ratio to the GPM hazards. This proportional hazards assumption requires validation in different cancers and might not always be appropriate.

Analyses

For each of the tested approaches, 5 parametric distributions are considered: the exponential, Weibull, log-logistic, lognormal, and Gompertz distributions. These are widely used in survival analysis and health economic evaluations and present diverse hazard profiles. Specifically, they capture constant (exponential), monotonic (Weibull and Gompertz), and unimodal (log-logistic and lognormal) hazards over time.[1] The different hazard profiles are expected to predict a diverse range of (incremental) mean life-years.

The underlying hazard and survival functions of the parametric distributions are specified in Table 2.

All survival analyses have been conducted from a Bayesian perspective, in line with the NICE TSD21 recommendations for future research.[5] All models were fitted using Hamiltonian Monte Carlo in rstan[32] (an R package that interfaces with the software Stan[33]), because Hamiltonian Monte Carlo is generally more efficient for Bayesian modeling of time to event in the presence of censoring.[34] Specifically, we used 3 Markov chains, each with 5000 warm-up iterations and 10 000 total iterations for posterior inference. On all modeled parameters, treatment coefficients were added in the hazard functions to avoid proportionality restrictions. For all parameters, a priori noninformative normal distributions with a mean of 0 and a standard deviation of 5 on the natural logarithm of the parameter were applied (see Table 2).

Compared Outcomes

The tested approaches were compared based on predicted (incremental) mean survival and corresponding uncertainty measures over a lifetime horizon (corresponding to 35 and 50 years for the multiple myeloma and breast cancer data sets, respectively), statistical fit, face validity of the predictions, and uncertainty over different parametric distributions. Additionally, the survival extrapolations based on the 30-month multiple myeloma data cut were superposed on the observed Kaplan-Meier curve for the 75-month cut to evaluate the accuracy of each extrapolation method.

The statistical fits to the observed data were compared based on the Watanabe-Akaike information criterion (WAIC).[35] The WAIC can be interpreted in the same way as the standard Akaike

Table 1. Hazard functions of the compared approaches.

Approach	Hazard functions	Explanation
1. Internal additive hazards[5]	$h_1(t) = h_{GPM}(t)+h_{DSM}(t)$ t = time $h_{DSM}(t)$ = parametric hazard (DSM) $h_{GPM}(t)$ = GPM hazard	The hazards ($h_1(t)$) are the sum of the GPM hazards ($h_{GPM}(t)$) and DSM hazards estimated by a parametric distribution ($h_{DSM}(t)$).
2. No GPM adjustment parametric hazards[2]	$h_2(t) = h_{ACM}(t)$ t = time $h_{ACM}(t)$ = parametric hazard (ACM)	The survival extrapolations rely purely on the hazards ($h_2(t)$) derived from the parametric distribution fitted on the trial data.
3. Converging hazards[6]	$h_3(t) = IF[(h_{GPM}(t) > h_2(t)), h_{GPM}(t), h_2(t)]$ t = time $h_2(t)$ = parametric hazard rates (ACM) $h_{GPM}(t)$ = GPM hazard	The hazards ($h_2(t)$) derived from the parametric distribution fitted on the trial data are applied until the GPM hazards ($h_{GPM}(t)$) become higher than the estimated hazards ($h_2(t)$). From that point in time onward, the GPM hazards are applied for the survival extrapolations.
4. External additive hazards[6]	$h_4(t) = h_{GPM}(t)+h_2(t)$ t = time $h_2(t)$ = parametric hazard rates (ACM) $h_{GPM}(t)$ = GPM hazard	The hazards predicted by this approach ($h_4(t)$) consist of GPM hazards ($h_{GPM}(t)$) added to the hazards ($h_2(t)$) derived from the parametric distribution fitted on the trial data.
5. Proportional hazards[6]	$h_5(t) = h_{GPM}(t) \times HR$ t = time HR = hazard ratio $h_{GPM}(t)$ = GPM hazard	The hazards predicted by this approach ($h_5(t)$) are the GPM hazards ($h_{GPM}(t)$) multiplied by a hazard ratio that represents the excess mortality in the RCT versus the general population.

ACM indicates all-cause mortality; DSM, disease-specific mortality; GPM, general population mortality; HR, hazard ratio; RCT, randomized controlled trial.

information criterion and the deviance information criterion (ie, lower scores represent better goodness of fit to the data).[35] All of these criteria assess the fit over the observed period of the trial, and do not help in establishing the accuracy of the extrapolation. Because the WAIC is regarded as an improvement and preferred alternative versus the deviance information criterion,[36-39] this article is restricted to WAIC.

Face validity is defined in 2 ways: (1) in terms of the visual fit of the survival models to the observed data; and (2) whether the hazards predicted over time are in line with the expected long-term survival of the disease under study, given background clinical knowledge. Uncertainty is quantified by the 95% credible interval of the predicted incremental life-years. Finally, the variation in incremental survival over the different parametric distributions (exponential, Weibull, lognormal, log-logistic, and Gompertz) is assessed.

Results

For each data set, the mean incremental life-years, the corresponding interval estimates, and the WAIC values of the tested approaches are presented in Table 3. In the supplementary materials, we present the visual fit (see Appendix Figs. 2-4 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008), long-term extrapolations (see Appendix Figs. 5-7 in

Table 2. Parameterization of the exponential, Weibull, log-logistic, lognormal and Gompertz distributions.

Distribution	Parameterization	Hazard function	Parameters	Noninformative priors
Exponential	$S(t) = \exp(-\lambda t)$	λ	λ: rate t: time	λ: Mean = 0, SD = 5
Weibull	$S(t) = \exp\left(-\left(\frac{t}{\beta}\right)^{\alpha}\right)$	$\frac{\alpha}{\beta^{\alpha}}t^{\alpha-1}$	α: shape β: scale t: time	α: Mean = 0, SD = 5 β: Mean = 0, SD = 5
Log-logistic	$S(t) = \frac{1}{1+\left(\frac{t}{\alpha}\right)^{\beta}}$	$\frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{1+(t/\alpha)^{\beta}}$	α: scale β: shape t: time	α: Mean = 0, SD = 5 β: Mean = 0, SD = 5
Lognormal	$S(t) = 1 - \Phi\left(\frac{\log(t)-\mu}{\sigma}\right)$	$\frac{\phi\left(\frac{\log(t)-\mu}{\sigma}\right)}{\sigma t\left[1-\Phi\left(\frac{\log(t)-\mu}{\sigma}\right)\right]}$	Φ: standard normal distribution μ: mean σ: standard deviation t: time	μ: Mean = 0, SD = 5 σ: Mean = 0, SD = 5
Gompertz	$S(t) = exp\left(-\frac{\beta}{\alpha}(\exp(\alpha t) - 1)\right)$	$\beta \exp(\alpha t)$	α: shape β: rate t: time	α: Mean = 0, SD = 5 β: Mean = 0, SD = 5

Table 3. Mean incremental life-years of the compared approaches.

Data set	Approach	Exponential		Weibull		Log-logistic		Lognormal		Gompertz	
		WAIC	Incremental survival in years (95% CrI)	WAIC	Incremental survival in years (95% CrI)	WAIC	Incremental survival in years (95% CrI)	WAIC	Incremental survival in years (95% CrI)	WAIC	Incremental survival in years (95% CrI)
Multiple myeloma (30 months)	1 Internal additive hazards	1399.6	2.2 (0.5-3.8)	1399.6	4.3 (1.4-6.5)	1398.4	3.6 (1.3-5.5)	**1395.0**	3.5 (1.4-5.3)	1401.7	7.3 (0.5-9.4)
	2 No GPM adjustment	1403.0	2.9 (0.6-5.6)	1403.1	7.6 (1.9-13.5)	1401.5	7.8 (2.7-12.5)	**1397.6**	8.5 (3.1-13.0)	1404.8	20.3 (−3.1-24.9)
	3 Converging hazards		2.6 (0.6-4.6)		5.4 (1.6-7.7)		4.0 (1.6-6.0)		3.5 (1.3-5.4)		7.1 (−0.2 to 9.3)
	4 External additive hazards		1.7 (0.4-3.0)		3.4 (1.0-5.4)		3.0 (1.1-4.6)		2.9 (1.1-4.5)		6.2 (−0.4 to 8.4)
	5 Proportional hazards	No distribution: 1387.7 / 1.6 (0.4-2.8)									
Multiple myeloma (75 months)	1 Internal additive hazards	4029.6	1.7 (0.8-2.6)	4029.8	1.6 (0.7-2.6)	4044.6	1.8 (0.8-2.7)	4068.6	2.1 (1.0-3.2)	**4020.3**	0.9 (0.1-1.4)
	2 No GPM adjustment	4038.5	1.8 (0.7-3.1)	4038.0	1.7 (0.6-3.1)	4061.5	2.6 (1.0-4.4)	4094.7	3.4 (1.5-5.5)	**4028.6**	0.8 (−0.1 to 1.8)
	3 Converging hazards		1.7 (0.6-2.8)		1.7 (0.6-2.9)		1.9 (0.8-3.0)		2.1 (0.9-3.3)		0.8 (−0.1 to 1.8)
	4 External additive hazards		1.1 (0.4-1.9)		1.1 (0.4-1.9)		1.3 (0.5-2.1)		1.5 (0.6-2.3)		0.7 (0.1-1.4)
	5 Proportional hazards	No distribution: 4013.4 / 1.2 (0.5-1.9)									
Breast cancer (64 months)	1 Internal additive hazards	5155.6	0.9 (0.5-1.2)	5159.1	0.9 (0.5-1.3)	**5095.6**	1.3 (0.7-2.0)	5120.0	1.3 (0.7-2.0)	5141.1	1.2 (−0.3 to 3.0)
	2 No GPM adjustment	5155.9	0.9 (0.5-1.3)	5159.4	0.9 (0.5-1.3)	**5095.4**	1.6 (0.9-2.4)	5121.4	1.6 (0.8-2.4)	5141.5	1.4 (−1.4 to 5.0)
	3 Converging hazards		0.9 (0.5-1.3)		0.9 (0.5-1.3)		1.5 (0.8-2.2)		1.5 (0.8-2.3)		1.3 (−0.4 to 3.3)
	4 External additive hazards		0.9 (0.5-1.2)		0.9 (0.5-1.3)		1.3 (0.7-1.9)		1.3 (0.7-2.0)		1.2 (−0.3 to 2.9)
	5 Proportional hazards	No distribution: 5139.3 / 0.6 (0.4-0.9)									

Bold values represent the best-fitting distributions (lowest WAIC scores) for each data set.
CrI indicates credible interval; GPM, general population mortality; WAIC, Watanabe–Akaike information criterion.

Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008), and predicted hazards over time (Appendix Figs. 8 and 9 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008) for each data set (see Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008). Appendix Figure 10 displays extrapolations of the 30-month multiple myeloma survival data superposed on the Kaplan-Meier curve of the 75-month data cut (see Appendix Fig. 10 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008).

When comparing the 5 different parametric distributions, the immature (30 month) multiple myeloma data set is best fitted by the lognormal distribution in terms of WAIC. For the mature (75 months) multiple myeloma and breast cancer (64 months) data sets, the best-fitting distributions are the Gompertz and log-logistic, respectively.

Table 3 shows that, overall and for each of the survival distributions individually, the incremental mean survival varies the most for the immature multiple myeloma data set. For this data set, the best-fitting lognormal distribution has an incremental survival of 3.5 years (95% credible interval: 1.4-5.3) for the internal additive hazards approach, 8.5 (3.1-13.0) for the no GPM adjustment approach, 3.5 (1.3-5.4) for the converging hazards approach, 2.9 (1.1-4.5) for the external additive hazards approach, and 1.6

(0.4-2.8) for the proportional hazards approach. Absolute life-years are the highest for the no GPM adjustment approach (18.2 and 9.8 years for the active and control arms, respectively), followed by the converging hazards approach (11.3 and 7.8 years), internal additive hazards approach (9.9 and 6.5 years), external additive hazards approach (8.9 and 6.0 years), and proportional hazards approach (5.5 and 3.9 years) (see Appendix Fig. 5 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008).

Regarding face validity in terms of the fit to the observed data, most approaches and distributions fit the data well. One exception is the external additive hazards approach, which assumes GPM-related mortality during the trial is negligible. As this is unlikely in the 71-year old multiple myeloma data set, the approach underestimates the observed Kaplan-Meier survival data of the 2 data cuts (see Appendix Figs. 2 and 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008). In the breast cancer data set, the underestimation is less appreciable, as the patient population is younger (see Appendix Fig. 4 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008). The other exception is the visual fit of the proportional hazards approach to the breast cancer data set (see Appendix Fig. 4 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008), which suggests that the proportional hazards assumption is not valid in this scenario.

Regarding the face validity of the predicted hazards beyond the trial period, all predicted hazards are higher than those of the general population for the internal additive hazards, the external additive hazards, and the proportional hazards approaches (see Appendix Figs. 8 and 9 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008). The no GPM adjustment approach is only considered face valid if the predicted hazards remain above those of the general population. This is only the case for the Gompertz distribution, when it models increasing hazards for the control arm in both multiple myeloma data cuts and the active arm in the 75-month data cut (see Appendix Figs. 8 and 9 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008). When the converging hazards approach was applied to the immature 30-month multiple myeloma data set, the lognormal hazards were set equal to the GPM hazard from a certain point (6.0 and 11.0 years in the active and control arms, respectively) (see Appendix Fig. 4 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008). This induces a sudden change from decreasing to increasing hazards (see Appendix Figs. 8 and 9 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008). This approach is not considered face valid in the multiple myeloma data set, because it implies that extrapolated trial hazards fall below GPM hazards and that patients could be considered cured. However, multiple myeloma is still considered an incurable disease.[40]

The predictions of the internal additive hazards approach are face valid for each data set, both in terms of fit to the observed data and predicted hazards beyond the trial period.

As expected, the uncertainty over the 5 different modeling approaches and the 5 parametric distributions varies the most in the immature multiple myeloma data set (Table 3). For the lognormal distribution, the incremental life-years' credible interval has ranges of 3.9, 9.9, 4.1, 3.4, and 2.4 years with the internal additive hazards, no GPM adjustment, converging hazards, external additive hazards, and proportional hazards approaches, respectively.

The uncertainty over different distributions is assessed by the range of mean incremental life-years of the exponential, Weibull, log-logistic, lognormal, and Gompertz distributions in Table 3. The incremental life-years over the immature multiple myeloma data set range from 2.2 to 7.3 for the internal additive hazards approach, from 2.9 to 20.3 for the no GPM adjustment approach, from 2.6 to 7.1 for the converging hazards approach, and from 1.7 to 6.2 years for the external additive hazards approach. As the proportional hazards approach is exclusively based on semi-parametric Cox modeling, uncertainty over the different parametric distributions (eg, exponential, Weibull, etc.) is not assessed.

In Appendix Figure 10 (see Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.03.008), we superpose the 30-month multiple myeloma survival extrapolations on the Kaplan-Meier curve corresponding to the 75-month data cut. Visually, the extrapolation produced by the proportional hazards approach provides the best fit.

Discussion

The results for the 5 tested approaches show that it is important to adjust for GPM. The no GPM adjustment approach often predicts survival exceeding that of the general population. As expected, GPM adjustment is more crucial for extrapolations of immature data than for those of more mature data. For the less mature multiple myeloma data cut, the internal additive hazards, no GPM adjustment, converging hazards, external additive hazards, and proportional hazards approaches largely vary in outcomes. As data are often immature at the time of health economic evaluations, the variation in outcomes is likely to lead to large variations in ICER and may influence reimbursement decisions. Therefore, it is crucial to evaluate the robustness of outcomes under different GPM modeling approaches.

This study assesses the outcomes, face validity, and uncertainty quantification of different approaches, with and without GPM adjustment, to decide which methodology to apply. No GPM adjustment is only plausible when the extrapolated trial mortality hazards are above the GPM hazards at all times. This may be the case for cancers with increasing mortality hazards over time, in which case there will be less fluctuation in the results of the tested approaches.

On the contrary, survival outcomes will be more dissimilar for indications with decreasing hazards over time. The converging hazards approach is only appropriate if cure is clinically plausible within an indication. If cure is plausible, mixture cure or non-mixture cure models should be preferred.[5] The converging hazards approach was applied in the health technology assessment submission of liposomal cytarabine and daunorubicin for untreated acute myeloid leukemia, but the NICE committee recommended the use of cure models instead.[8]

The external additive hazards approach is applicable in younger populations, where the GPM hazard is low. Otherwise, the predicted survival of this approach will likely lie below the observed trial overall survival. The methodology has been previously used and accepted in the NICE health technology assessment submission of blinatumomab in relapsed/refractory acute lymphoblastic leukemia.[19] The patient population for this submission was relatively young, 41 years old on average. Therefore, the external additive hazards approach did not underestimate the observed trial mortality and, simultaneously, corrected the long-term mortality extrapolations for GPM. Alternatively, GPM hazards could also be added beyond the trial period, but in older patient populations this will then result in a sudden increase in the hazard and therefore an instantaneous drop in the overall survival after the trial period.

Finally, the proportional hazards approach is only valid if the proportional hazards assumption for ACM versus GPM holds and/or is clinically plausible. This appears to be the case for the multiple myeloma data but not for the breast cancer data set, where the provided extrapolations are not accurate. Therefore, it is

crucial to validate whether the proportional hazards assumption is clinically plausible. The proportional hazards approach may be valid outside oncology, for instance in cardiovascular diseases.

The internal additive hazards approach is considered valid for all data sets. In terms of the uncertainty quantification of predicted life-years and the variation of results under different parametric distributions, the performance of this methodology is comparable to that of the other tested approaches.

In future research, the 4 approaches that adjust for GPM can be applied to additional data sets to further investigate the generalizability of our findings. The internal additive hazards approach can also be extended to parametric survival network analyses[41]; other parametric distributions such as the generalized gamma; and other, more advanced survival modeling methods such as fractional polynomials, mixture models, piecewise models, and spline-based models. Another potential extension of the internal additive hazards approach involves applying stronger priors informed by external information to improve the face validity of long-term survival extrapolations.[42] This is especially of interest when the RCT data are very immature, for example, the median survival time has not been reached for the standard-of-care arm. Finally, in future research, the 5 approaches can be tested using a simulation study with known disease-specific and general population mortality.

In conclusion, different approaches for the inclusion of GPM result in different estimates of (incremental) survival and uncertainty. These findings are important as (incremental) survival is a major driver of the ICER, which is the basis of many reimbursement decisions. The internal additive hazards approach is considered face valid for each data set and a potential new standard for survival extrapolations, albeit it may provide similar incremental survival results than other GPM adjustment approaches in some cases.

Supplemental Materials

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2021.03.008.

Article and Author Information

Accepted for Publication: March 1, 2021

Published Online: Month xx, xxxx

doi: https://doi.org/10.1016/j.jval.2021.03.008

Author Affiliations: Ingress Health, Rotterdam, The Netherlands (Van Oostrum, Heeg); Department of Health Sciences, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands (Van Oostrum, Postma); AstraZeneca, Mölndal, Sweden (Ouwens); Department of Statistical Science, University College London, London, UK (Remiro-Azócar, Baio); Department of Economics, Econometrics & Finance, University of Groningen, Faculty of Economics & Business, Groningen, The Netherlands (Postma, Buskens); Department of Epidemiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands (Buskens).

Correspondence: Ilse van Oostrum, MSc, Ingress-Health, Weena 316-318, 3012 NJ, Rotterdam, The Netherlands. Email: ivoostrum92@gmail.com

Author Contributions: *Concept and design*: van Oostrum, Ouwens, Buskens, Heeg.
Acquisition of data: van Oostrum, Heeg
Analysis and interpretation of data: van Oostrum, Ouwens, Remiro-Azócar, Baio, Heeg
Drafting of the manuscript: van Oostrum, Ouwens, Remiro-Azócar, Baio, Postma, Buskens, Heeg
Critical revision of the paper for important intellectual content: van Oostrum, Ouwens, Remiro-Azócar, Baio, Postma, Buskens, Heeg
Statistical analysis: van Oostrum, Heeg
Supervision: Ouwens, Postma, Buskens

Conflict of Interest Disclosures: Drs van Oostrum and Heeg reported working as an external consultant for numerous pharmaceutical companies outside the submitted work. Dr Ouwens reported receiving personal fees from AstraZeneca outside the submitted work. Dr Remiro-Azócar reported receiving personal fees from IQVIA and ICON plc outside the submitted work. Dr Postma reported receiving grants and personal fees from Merck Sharp & Dohme, GlaxoSmithKline, Pfizer, Boehringer Ingelheim, Novavax, Bristol-Myers Squibb, AstraZeneca, Sanofi, IQVIA, and Seqirus outside the submitted work; personal fees from Quintiles, Novartis, and Pharmerit outside the submitted work; and grants from Bayer, BioMerieux, the World Health Organization, the EU, FIND, Antilope, DIKTI, LPDP, and Budi outside the submitted work. He is also a stockholder in Health-Ecore and PAG Ltd, an advisor to Asc Academics, and an editor for *Value in Health* but had no role in the peer review process of this article. No other disclosures were reported.

Funding/Support: The authors received no financial support for this research.

Acknowledgments: We thank Andrea Garcia for editorial support.

REFERENCES

1. Latimer NR. *NICE DSU Technical Support Document 14: Undertaking Survival Analysis for Economic Evaluations Alongside Clinical Trials - Extrapolation with Patient-Level Data*. London. London, UK: National Institute for Health and Care Excellence (NICE); 2013.
2. Latimer NR. Survival analysis for economic evaluations alongside clinical trials–extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. *Med Decis Making*. 2013;33(6):743–754.
3. Department of Health. Guidelines for preparing a submission to the Pharmaceutical Benefits Advisory Committee: version 5, September 2016. Canberra. https://pbac.pbs.gov.au/content/information/files/pbac-guidelines-version-5.pdf; 2016. Accessed February 20, 2020.
4. Canadian Agency for Drugs and Technologies in Health. *Guidelines for the Economic Evaluation of Health Technologies: Canada* — 4th Edition. Ottawa. https://www.cadth.ca/sites/default/files/pdf/guidelines_for_the_economic_evaluation_of_health_technologies_canada_4th_ed.pdf; 2017. Accessed February 19, 2020.
5. Rutherford MJ, Lambert PC, Sweeting MJ, Pennington R, Crowther MJ, Abrams KR. NICE DSU Technical Support Document 21. Flexible Methods for Survival Analysis. http://nicedsu.org.uk/; 2020. Accessed December 3, 2020.
6. Andersson TM-L, Dickman PW, Eloranta S, Lambe M, Lambert PC. Estimating the loss in expectation of life due to cancer using flexible parametric survival models. *Stat Med*. 2013;32(30):5286–5300.
7. Jackson C, Stevens J, Ren S, et al. Extrapolating survival from randomized trials using external data: A review of methods. *Med Decis Making*. 2016;37(4):377–390.
8. National Institute for Health and Care Excellence (NICE). Liposomal cytarabine–daunorubicin for untreated acute myeloid leukaemia: technology appraisal guidance [TA552]; Committee papers and final appraisal determination. https://www.nice.org.uk/guidance/ta552; 2018. Accessed February 10, 2019.
9. Grant A, Wileman S, Ramsay C, et al. The effectiveness and cost-effectiveness of minimal access surgery amongst people with gastro-oesophageal reflux disease – a UK collaborative study. The REFLUX trial. *Health Technol Assess*. 2008;12(31), 1-181:iii-iv.
10. Peek GJ, Elbourne D, Mugford M, et al. Randomised controlled trial and parallel economic evaluation of conventional ventilatory support versus extracorporeal membrane oxygenation for severe adult respiratory failure (CESAR). *Health Technol Assess*. 2010;14(35):1–46.
11. Chambers D, Paulden M, Paton F, et al. Sugammadex for the reversal of muscle relaxation in general anaesthesia: a systematic review and economic assessment. *Health Technol Assess*. 2010;14(39):1–211.
12. Hind D, Ward S, De NE, Simpson E, Carroll C, Wyld L. Hormonal therapies for early breast cancer: systematic review and economic evaluation. *Health Technol Assess*. 2007;11(26):iii-iv, ix-xi, 1-134.
13. Main C, Pitt M, Moxham T, Stein K. The clinical effectiveness and cost-effectiveness of rituximab for the first-line treatment of chronic lymphocytic leukaemia: an evidence review of the submission from Roche. *Health Technol Assess*. 2010;14(Suppl 2):27–32.
14. Aballéa S, Chancellor JVM, Raikou M, et al. Cost-effectiveness analysis of oxaliplatin compared with 5-fluorouracil/leucovorin in adjuvant treatment of stage III colon cancer in the US. *Cancer*. 2007;109(6):1082–1089.

15. Hisashige A, Yoshida S, Kodaira S. Cost-effectiveness of adjuvant chemotherapy with uracil-tegafur for curatively resected stage III rectal cancer. *Br J Cancer*. 2008;99(8):1232–1238.
16. Messori A, Trippoli S. A new method for expressing survival and life expectancy in lifetime cost-effectiveness studies that evaluate cancer patients (review). *Oncol Rep*. 1999;6(5):1135–1141.
17. Viscomi S, Pastore G, Dama E, et al. Life expectancy as an indicator of outcome in follow-up of population-based cancer registries: the example of childhood leukemia. *Ann Oncol*. 2006;17(1):167–171.
18. Howard DH, Tangka FK, Seeff LC, Richardson LC, Ekwueme DU. The impact of detection and treatment on lifetime medical costs for patients with precancerous polyps and colorectal cancer. *Health Econ*. 2009;18(12):1381–1393.
19. National Institute for Health and Care Excellence (NICE). Blinatumomab for previously treated Philadelphia-chromosome-negative acute lymphoblastic leukaemia: Technology appraisal guidance [TA450]; Committee papers and Final appraisal determination. https://www.nice.org.uk/guidance/ta450. Accessed February 10, 2019.
20. Fang CT, Chang YY, Hsu HM, et al. Life expectancy of patients with newly-diagnosed HIV infection in the era of highly active antiretroviral therapy. *QJM*. 2007;100(2):97–105.
21. Chu P-C, Wang J-D, Hwang J-S, Chang Y-Y. Estimation of life expectancy and the expected years of life lost in patients with major cancers: extrapolation of survival curves under high-censored rates. *Value Health*. 2008;11(7):1102–1109.
22. Hwang JS, Wang JD. Monte Carlo estimation of extrapolation of quality-adjusted survival for follow-up studies. *Stat Med*. 1999;18(13):1627–1640.
23. Barton P, Jobanputra P, Wilson J, Bryan S, Burls A. The use of modelling to evaluate new drugs for patients with a chronic condition: the case of antibodies against tumour necrosis factor in rheumatoid arthritis. *Health Technol Assess*. 2004;8(11):iii, 1-91.
24. Rodgers M, Hodges R, Hawkins J, et al. Colour vision testing for diabetic retinopathy: a systematic review of diagnostic accuracy and economic evaluation. *Health Technol Assess*. 2009;13(60):1–160.
25. Chen Y-F, Jobanputra P, Barton P, et al. A systematic review of the effectiveness of adalimumab, etanercept and infliximab for the treatment of rheumatoid arthritis in adults and an economic evaluation of their cost-effectiveness. *Health Technol Assess*. 2006;10(42):iii-iv, xi-xiii, 1-229.
26. San Miguel JF, Schlag R, Khuageva NK, et al. Bortezomib plus melphalan and prednisone for initial treatment of multiple myeloma. *N Engl J Med*. 2008;359(9):906–917.
27. San Miguel JF, Schlag R, Khuageva NK, et al. Persistent overall survival benefit and no increased risk of second malignancies with bortezomib-melphalan-prednisone versus melphalan-prednisone in patients with previously untreated multiple myeloma. *J Clin Oncol*. 2013;31(4):448–455.
28. Baselga J, Cortés J, Kim S-B, et al. Pertuzumab plus trastuzumab plus docetaxel for metastatic breast cancer. *N Engl J Med*. 2012;366(2):109–119.
29. Mitchell M, Muftakhidinov B, Winchen T, et al. Engauge Digitizer software. http://markummitchell.github.io/engauge-digitizer. Accessed January 27, 2019.
30. Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12:9.
31. Bell FC, Miller ML. Life tables for the United States social security area 1900-2100. *SSA Pub. No. 11-11536*. [Actuarial Study No. 120.]. 2005.
32. Stan Development Team. RStan: the R interface to Stan. R package version 2.18.2. http://mc-stan.org/. Accessed January 27, 2019.
33. Carpenter B, Gelman A, Hoffman MD, et al. Stan: a probabilistic programming language. *J Stat Softw*. 2017;76(1).
34. Baio G. survHE: survival analysis for health economic evaluation and cost-effectiveness modeling. *J Stat Softw*. 2020;95(14).
35. Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J Mach Learn Res*. 2010;11:3571–3594.
36. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput*. 2017;27(5):1413–1432.
37. Plummer M. Penalized loss functions for Bayesian model comparison. *Biostatistics*. 2008;9(3):523–539.
38. Linde A. DIC in variable selection. *Stat Neerland*. 2005;59(1):45–56.
39. Bürkner P-C. brms: an R package for Bayesian multilevel models using Stan. *J Stat Softw*. 2017;80(1):1–28.
40. Ravi P, Kumar SK, Cerhan JR, et al. Defining cure in multiple myeloma: a comparative study of outcomes of young individuals with myeloma and curable hematologic malignancies. *Blood Cancer J*. 2018;8(3):26.
41. Ouwens MJNM, Philips Z, Jansen JP. Network meta-analysis of parametric survival curves. *Res Synth Methods*. 2010;1(3-4):258–271.
42. Soikkeli F, Hashim M, Ouwens M, Postma M, Heeg B. Extrapolating survival data using historical trial-based a priori distributions. *Value Health*. 2019;22(9):1012–1017.