Article:
Kearns, B. orcid.org/0000-0001-7730-668X, Stevenson, M., Triantafyllopoulos, K. et al. (1 more author) (2021) The extrapolation performance of survival models for data with a cure fraction: a simulation study. Value in Health, 24 (11). pp. 1634-1642. ISSN 1098-3015

https://doi.org/10.1016/j.jval.2021.05.009

ScienceDirect

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

Methodology

The Extrapolation Performance of Survival Models for Data With a Cure Fraction: A Simulation Study

Benjamin Kearns, MSc, PhD, Matt D. Stevenson, BSc, PhD, Kostas Triantafyllopoulos, MSc, PhD, Andrea Manca, MSc, PhD

ABSTRACT

Objectives: Curative treatments can result in complex hazard functions. The use of standard survival models may result in poor extrapolations. Several models for data which may have a cure fraction are available, but comparisons of their extrapolation performance are lacking. A simulation study was performed to assess the performance of models with and without a cure fraction when fit to data with a cure fraction.

Methods: Data were simulated from a Weibull cure model, with 9 scenarios corresponding to different lengths of follow-up and sample sizes. Cure and noncure versions of standard parametric, Royston-Parmar, and dynamic survival models were considered along with noncure fractional polynomial and generalized additive models. The mean-squared error and bias in estimates of the hazard function were estimated.

Results: With the shortest follow-up, none of the cure models provided good extrapolations. Performance improved with increasing follow-up, except for the misspecified standard parametric cure model (lognormal). The performance of the flexible cure models was similar to that of the correctly specified cure model. Accurate estimates of the cured fraction were not necessary for accurate hazard estimates. Models without a cure fraction provided markedly worse extrapolations.

Conclusions: For curative treatments, failure to model the cured fraction can lead to very poor extrapolations. Cure models provide improved extrapolations, but with immature data there may be insufficient evidence to choose between cure and noncure models, emphasizing the importance of clinical knowledge for model choice. Dynamic cure fraction models were robust to model misspecification, but standard parametric cure models were not.

Keywords: cure models, flexible survival models, forecasting, survival extrapolation.

VALUE HEALTH. 2021; ■(■):■–■

Introduction

Estimates of future survival are frequently a key component of technology appraisals (TAs). Decisions about whether a technology should be funded are often sensitive to the extrapolation approach used.[1] Hence, it is important that appropriate methods are used when generating extrapolations. In the last few years, there has been a growing interest in generating extrapolations for curative treatments, such as immuno-oncology drugs.[2–5] These present unique challenges, because the presence of a cured fraction creates heterogeneity in the survivor (and hazard) functions, resulting in complex hazard shapes. Standard parametric survival models, as typically employed in TAs, are usually insufficiently flexible to adequately describe these shapes.[4,6]

The development of potentially curative treatments has led to an increased use of cure fraction models in TAs.[7] During the recent update of the methods used in National Institute for Health and Care Excellence (NICE) TAs, cure fraction models were explicitly identified as an option for consideration.[8] Cure models were also included in the recently published NICE decision support unit technical support document on flexible methods for survival analysis.[9] Based on the results of simulation studies, the authors concluded that when the truth contained a cure fraction, cure models had lower bias than alternative methods. Nevertheless, there are limitations with cure models. Reliable estimates of the cure fraction require long follow-up; in practice estimates of the cure fraction are very sensitive to model specification.[4,10,11] Although there is a growing awareness of the importance of using a cure fraction model in TAs, there are limited examples of their use; a recent review of NICE TAs did not find any reimbursement decisions that had been made based on a nonzero cured fraction.[7]

A proper assessment of the performance of cure models requires multiple data sets with full follow-up, so that models may be fit to an interim data cut. This allows an evaluation of both within-sample fit and extrapolation performance. In the absence of suitable data, simulation studies may be used; to-date 2 simulations of cure models for extrapolation in TAs have been

considered.[7,9] These are limited because between them only 2 cure fraction models have been considered: a Weibull cure model and a Royston-Parmar (RP) cure model with 2 internal knots. There is a need for a comprehensive assessment of multiple cure fraction models for extrapolating data with a cured fraction. The primary aim of this study was to address this evidence gap by comparing the overall goodness of fit, for both within-sample estimates and extrapolations, of cure models when fit to data containing a cure fraction. There were 2 secondary aims:

1. To identify the accuracy with which the true cure fraction was estimated
2. To assess the impact of not incorporating external data, by comparing extrapolations from cure models with extrapolations from models that do not assume a cure fraction

Methods

To assess the extrapolation performance of models with a cure fraction, a simulation study was used. The reporting of the simulation study follows published guidance.[12] Components of the simulation study are reported based on their aims (provided in the previous section), data-generating mechanism, methods, estimands, and performance measures. The code used is provided in Appendix 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.05.009.

Data-Generating Mechanism

The "true" survival and hazard functions were simulated from a Weibull cure data-generating mechanism. Hazard and survival functions are represented by a Weibull model for individuals who will die of the disease. General population English life tables are used for cured individuals.[13] A Weibull model with shape (γ) and scale (Λ) values of 1.6 and 2.6, respectively, was used for uncured individuals, and it was assumed that one-quarter of the sample would be cured (ρ = 0.25). The shape and scale were arbitrarily chosen, with the aim of providing disease-specific death within a moderately short timeframe (mean disease-specific survival 2.33 years) and with few survivors beyond 8 years. The cure fraction was arbitrarily chosen with the intention that it was not the majority of patients, but large enough to notably influence survival.

The equations for obtaining the observed survivor and hazard functions are:

$$S_{t_i} = \rho S^P_{t_i} + (1-\rho) S^U_{t_i}$$

$$\lambda_{t_i} = \frac{\rho \lambda^P_{t_i} S^P_{t_i} + (1-\rho)\lambda^U_{t_i} S^U_{t_i}}{S_{t_i}}$$

where $S^P_{t_i}$, $S^U_{t_i}$ are the survival for the cured and uncured populations at time t_i, with $\lambda^P_{t_i}$, $\lambda^U_{t_i}$ the corresponding hazard values (both of which are monotone increasing for this study). The individual components are provided in Appendix Figure 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.05.009. A total of 9 scenarios were considered, with 200 data sets simulated for each scenario. For each individual, survival times were randomly sampled from $S^P_{t_i}$ or $S^U_{t_i}$ with probabilities 0.25 and 0.75, respectively. That is, uncertainty in survival (and hence hazard) functions was simulated but it was assumed that there was no uncertainty in the cure fraction. The 9 scenarios corresponded to 3 different sample sizes (small = 100, medium = 300, large = 600) and 3 different lengths of follow-up (short = 2 years, medium = 4 years, long = 8 years). The longest follow-up was chosen so that there were almost no uncured individuals still alive. The shortest follow-up was chosen to be representative of the lengths of follow-up often seen in cancer trials.[14] Sample sizes were chosen to represent typical sample sizes of cancer treatments seen in TAs. Details on the 9 scenarios are provided in Table 1. Figure 1 shows the simulated hazards for each scenario (in gray) along with the true hazard (in black, which is the same for each scenario). The short-term increase in the hazard function is driven by deaths among the uncured population. The first turning point occurs when the contribution of the cured population outweighs that of the uncured population, with the overall hazard decreasing to that of the cured population. This is followed by a long-term increase in the hazard function because of aging.

Estimand and Performance Measures

The estimand was the natural logarithm of the time-varying hazard function λ_{t_i}. The use of the hazard function is preferable to the survivor function because the latter is a cumulative measure, so estimates will not be independent. The primary and secondary performance measures were mean-squared error (MSE) and bias. Both time-varying and summary performance measures were considered. A time horizon of 40 years was used (at which overall survival was 0.1%), with time-steps of 0.05 years. Appendix 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.05.009 provides further details and justification.

Methods

For all the methods (models) discussed in this section, a brief overview is provided. Further details on model specification (for the noncured component) are provided in a previous publication.[15]

A total of 4 types of cure fraction model were included. For these, the correct background mortality was provided for the cured group. Hence, the models estimated the cure proportion and the hazard function for the uncured group.

1. Weibull cure model: this is the same as the data-generating mechanism, so the model structure is correctly specified.
2. Lognormal cure model: the model structure is incorrectly specified (misspecified) with regard to the true functional form for the uncured group.
3. RP cure model. These introduce additional flexibility through the use of piecewise cubic polynomials. Between 0 and 4 internal knots were considered for these spline-based models. Models were fit on the hazard scale, so include as a special case the Weibull cure model (zero knots). Hence, the model is overspecified because it includes the correctly specified model as a special case but allows for a range of more flexible models. For each data set, the model with the lowest value of Akaike information criterion (AIC) was used to generate extrapolations.
4. Dynamic cured fraction models (DCFMs). These are models for which parameters evolve dynamically over time, as modeled by a time series (such as a random walk). A total of 2 models were considered: a local trend model (a Weibull model with dynamic parameters) and a damped trend model (as before, but the trend in extrapolations is dampened over time until it becomes 0). The local trend model is overspecified (it is the same as the Weibull cure model if there is no parameter evolution), while the damped trend model is incorrectly specified.

Table 1. Details of the 9 scenarios simulated.

Scenario	Follow-up (survival %)	Sample size
Short follow-up, small sample size	2 years (63.9%)	100
Short follow-up, medium sample size	Cured = 97.7%	300
Short follow-up, large sample size	Uncured = 52.6%	600
Medium follow-up, small sample size	4 years (34.8%)	100
Medium follow-up, medium sample size	Cured = 95.5%	300
Medium follow-up, large sample size	Uncured = 14.5%	600
Long follow-up, small sample size	8 years (22.7%)	100
Long follow-up, medium sample size	Cured = 90.1%	300
Long follow-up, large sample size	Uncured = 0.3%	600

Hence, in total 5 cure models were included. R version 3.5.3 (R Core Team, Austria, Vienna) was used. All dynamic models were fit using RStan (Stan Development Team, Columbia); the remaining cure fraction models used the cuRe package (Jakobsen, Aalborg, Denmark).[16]

A total of 4 classes of model without a cured fraction were considered. Where multiple model specifications were possible, model choice was based on minimizing AIC to provide an automated method that reflects current approaches to model choice.[17] One exception was the Gompertz, which was excluded from the main results because the majority of extrapolations lacked face validity.

1. Current practice. These models were designed to reflect the models currently used in TAs.[6] A total of 6 models were evaluated: exponential, Weibull, lognormal, log-logistic, gamma, and generalized gamma. One model was retained. Results including the Gompertz are provided in Appendix Figure 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.05.009. Fitting used the flexsurv package (Jackson, Cambridge, England).[18]
2. Fractional polynomials (FPs). These represent the outcome as a sum of polynomial terms. First-order FP and second-order FP models were considered, where the order denotes the number of polynomial terms, fit using the stats package. A total of 8

Figure 1. Simulated hazards (gray lines) for the 9 scenarios, along with the truth (black line).

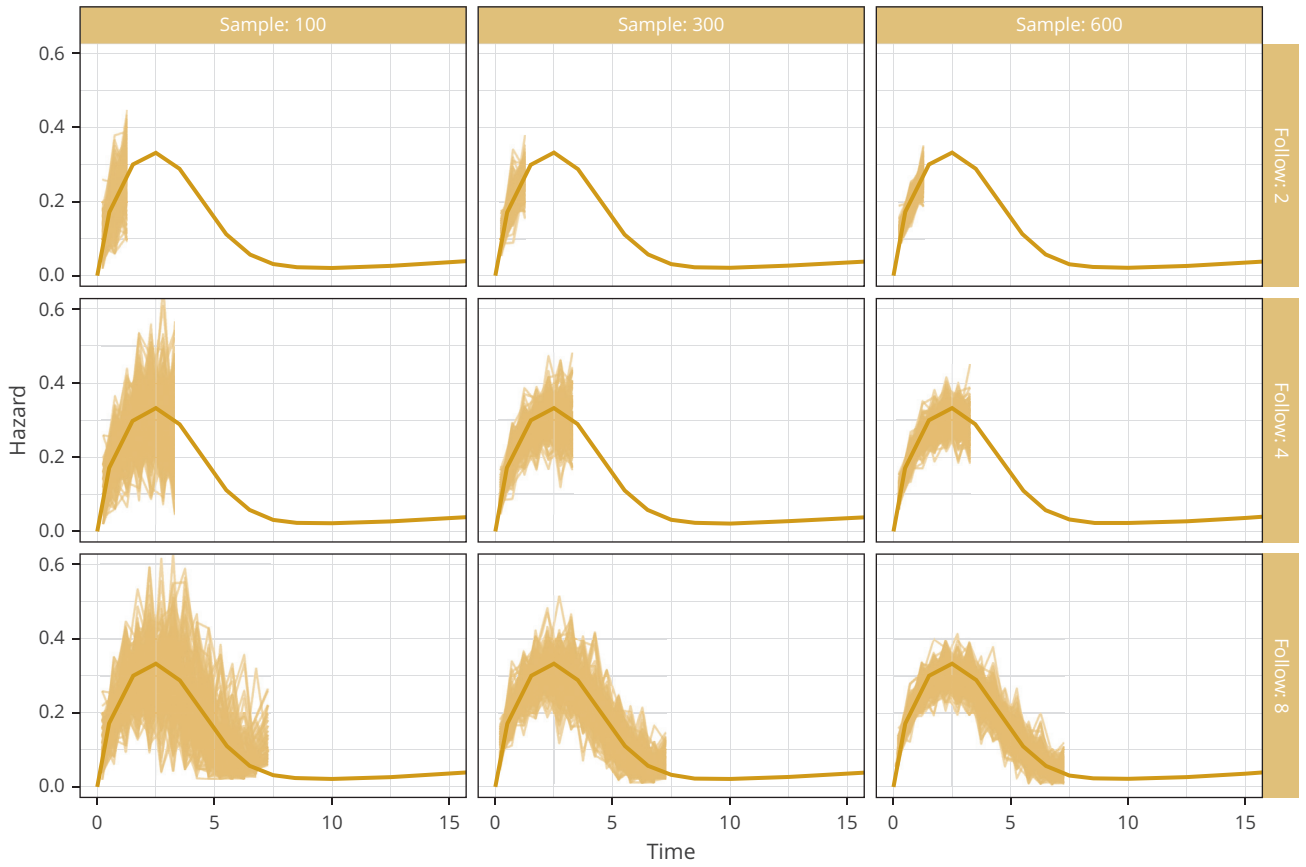
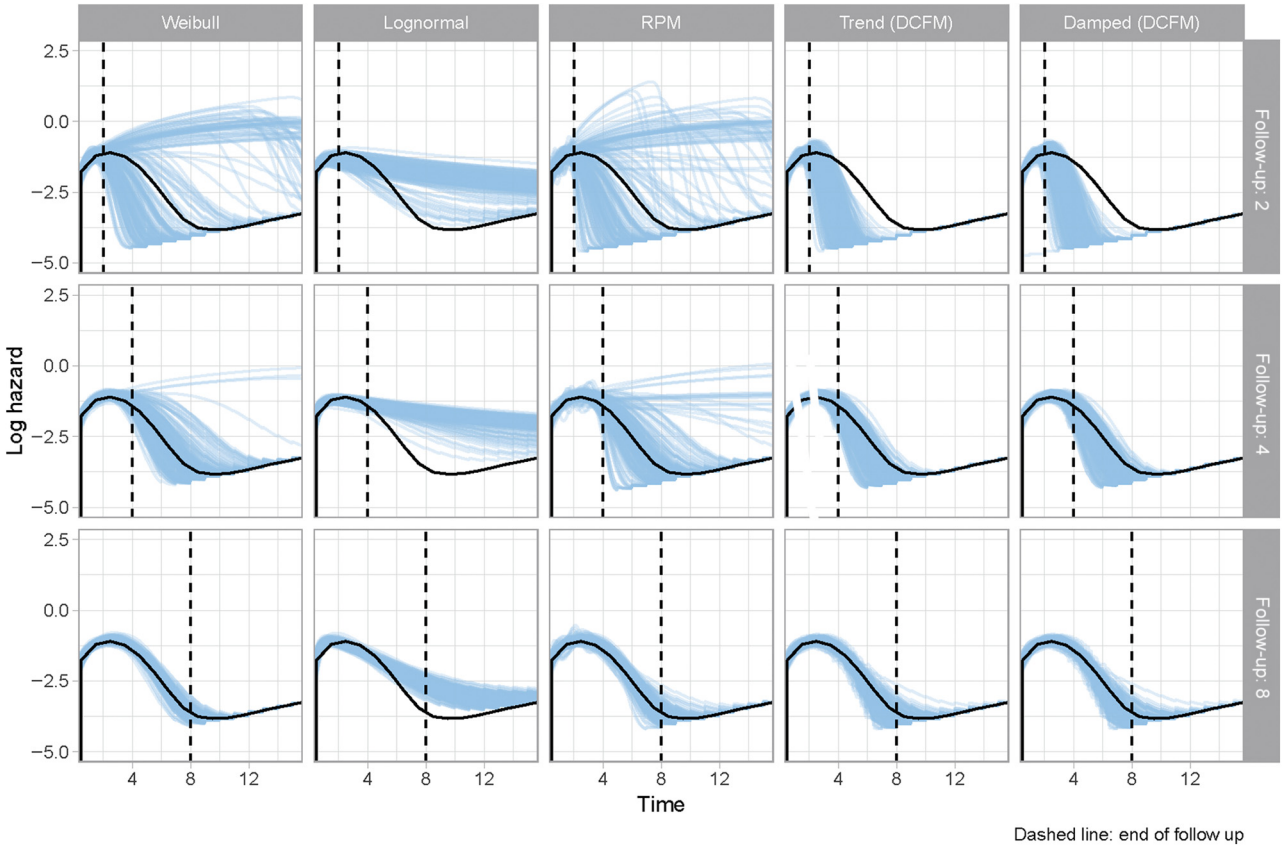


Figure 2. Estimates of the log-hazard compared with the truth: cure fraction models.



DCFM indicates dynamic cured fraction model; RPM, Royston-Parmar model.

first-order FP and 36 second-order FP models were considered; 1 model was retained for each.

3. Spline-based models. A total of 2 implementations were evaluated. One used the generalized additive models implemented in the mgcv package (Wood, Edinburgh, Scotland),[19] which incorporates penalization (via shrinkage of parameter estimates) during model fitting, removing the decision of how many knots to use. The second was the RP model (RPM), with up to 5 internal knots (no internal knots being the same as the Weibull model), as estimated by the flexsurv package, with 1 RPM retained.[18]
4. Dynamic survival models (DSMs). A total of 2 DSMs were used: a local trend DSM and a damped trend DSM.

This provided 7 models for which goodness of fit was examined.

Results

For the 9 scenarios considered, varying the length of follow-up had a larger impact on results than varying the sample size. Hence detailed results are presented for the 3 follow-ups (2, 4, 8 years), with a sample size of 300. Summary results are provided for the remaining 6 scenarios.

Cure Models

A visual comparison of the model estimates and the true log-hazards is provided in Figure 2. With the shortest follow-up, all the considered models provided poor predictions. The Weibull cure and RP cure models both resulted in very similar extrapolations; exhibiting the largest variation of the 5 models considered with both overestimates and underestimates of the true hazard. This similarity was because for most simulations (78%) the RP cure model chose zero internal knots, resulting in a Weibull cure model. The mean number of internal knots ranged from 0.32 to 0.55 across the 9 scenarios. Extrapolations from the DCFMs underestimated the true hazards while the lognormal cure model provided extrapolations that nearly always overestimated the truth. For the shortest follow-up, increasing the sample size led to less variation in extrapolations from the lognormal cure model (but not improvement in fit) and had a negligible impact on extrapolations from the remaining 4 cure models. For all 5 cure models visual goodness of fit improved as the length of follow-up increased; with the longest follow-up, all the cure models provided very good fits except for the misspecified lognormal cure model, which continued to systematically overestimate the truth.

Summary measures of MSE and bias averaged over the entire time horizon are provided for all 9 scenarios in Table 2, along with a rankogram in Appendix Figure 3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.05.009. A rankogram displays the average number of times each model achieved each rank based on its goodness of fit (where 1 is the best and 5 the worst). Time-varying performance measures are provided in Appendix Figures 4, 5 and Appendix Table 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.05.009. The within-sample 95% confidence intervals all included zero, with point estimates very close to zero, indicating that within-sample

Table 2. Goodness of fit over the entire time horizon: cure fraction models.

Measure and model	Sample size: 100. Follow-up:			Sample size: 300. Follow-up:			Sample size: 600. Follow-up:		
	2 years	3 years	4 years	2 years	3 years	4 years	2 years	3 years	4 years
Overall mean squared error									
Weibull cure model	1.73	0.34	0.04	1.35	0.14	0.01	0.84	0.04	0.01
Lognormal cure model	0.85	0.87	0.23	0.87	1.04	0.17	0.92	1.06	0.17
Royston-Parmar cure model	1.77	0.34	0.06	1.36	0.31	0.02	0.91	0.22	0.01
Local trend cure model	0.90	0.13	0.07	0.70	0.08	0.02	0.66	0.06	0.01
Damped trend cure model	0.87	0.13	0.07	0.68	0.08	0.02	0.64	0.06	0.01
Overall bias									
Weibull cure model	0.20	0.16	0.08	0.07	0.04	0.01	0.01	0.02	0.02
Lognormal cure model	0.33	0.45	0.48	0.49	0.59	0.60	0.18	0.20	0.21
Royston-Parmar cure model	0.08	0.12	0.08	0.05	0.10	0.07	0.01	0.02	0.02
Local trend cure model	−0.42	−0.37	−0.37	−0.10	−0.08	−0.07	−0.04	−0.03	−0.02
Damped trend cure model	−0.42	−0.37	−0.36	−0.10	−0.08	−0.06	−0.04	−0.03	−0.02

bias was small and within an acceptable range. Similar within-sample results were obtained for MSE values which, when compared with the range of MSE values for the extrapolated period, were very minor. In the scenarios with the longest follow-up, all the models apart from the lognormal had overall (within-sample and extrapolated period) MSE and bias values that were very close to zero. Furthermore, except for the lognormal, extrapolation performance and overall goodness of fit of the models generally increased as the length of follow-up or the sample size increased. The lognormal cure model had similar within-sample fit to the other models; the poor overall performance of this model was due to it providing poor extrapolations, which remained biased even with 8 years follow-up.

The results of the rankogram further illustrate the poor performance of the lognormal model. The remaining 4 models typically had similar MSE values, while the 2 DCFMs usually had the lowest bias values. Hence, for the nondynamic cure models, model misspecification (using a lognormal cure model instead of a Weibull cure model) led to reduced extrapolation performance, while using an overspecified model (RP cure model) had a negligible impact on bias. The overall goodness of fit of the DCFMs was similar to that of the correctly specified Weibull cure model, despite being overspecified (local trend) or misspecified (damped trend). Across the 9 scenarios, the Weibull cure model and damped trend cure models had the lowest MSE in 4 scenarios each. The remaining scenario had the least mature data (follow-up 2 years and sample size = 100). For this, the lognormal cure model provided the best MSE as its extrapolations had very low variability. The second lowest MSE was from a DCFM in 6 scenarios and the RP cure model in the remaining 3.

Estimates of the Cure Fraction

Estimates of the cure fraction for each model and each scenario are provided in Figure 3. For the shortest follow-up, none of the models provided accurate estimates of the cure fraction; on average, the lognormal cure model underestimated the true value, while the remaining models overestimated it. Increasing length of follow-up led to more accurate and less variable estimates, although, even at the longest follow-up (when virtually all the uncured patients had died), the lognormal provided an underestimate of the cure fraction.

In general, the Weibull cure and RP cure models provided slightly more accurate estimates of the true cure fraction than the 2 DCFMs. Nevertheless, this did not correspond to improved goodness of fit. This was most notable for the scenarios with shortest follow-up, for which both DCFMs had better overall MSE despite substantially overestimating the cured fraction (range for mean estimates: DCFMs 44.5%-45.3%, Weibull cure and RP cure models 30.3%-34.7%).

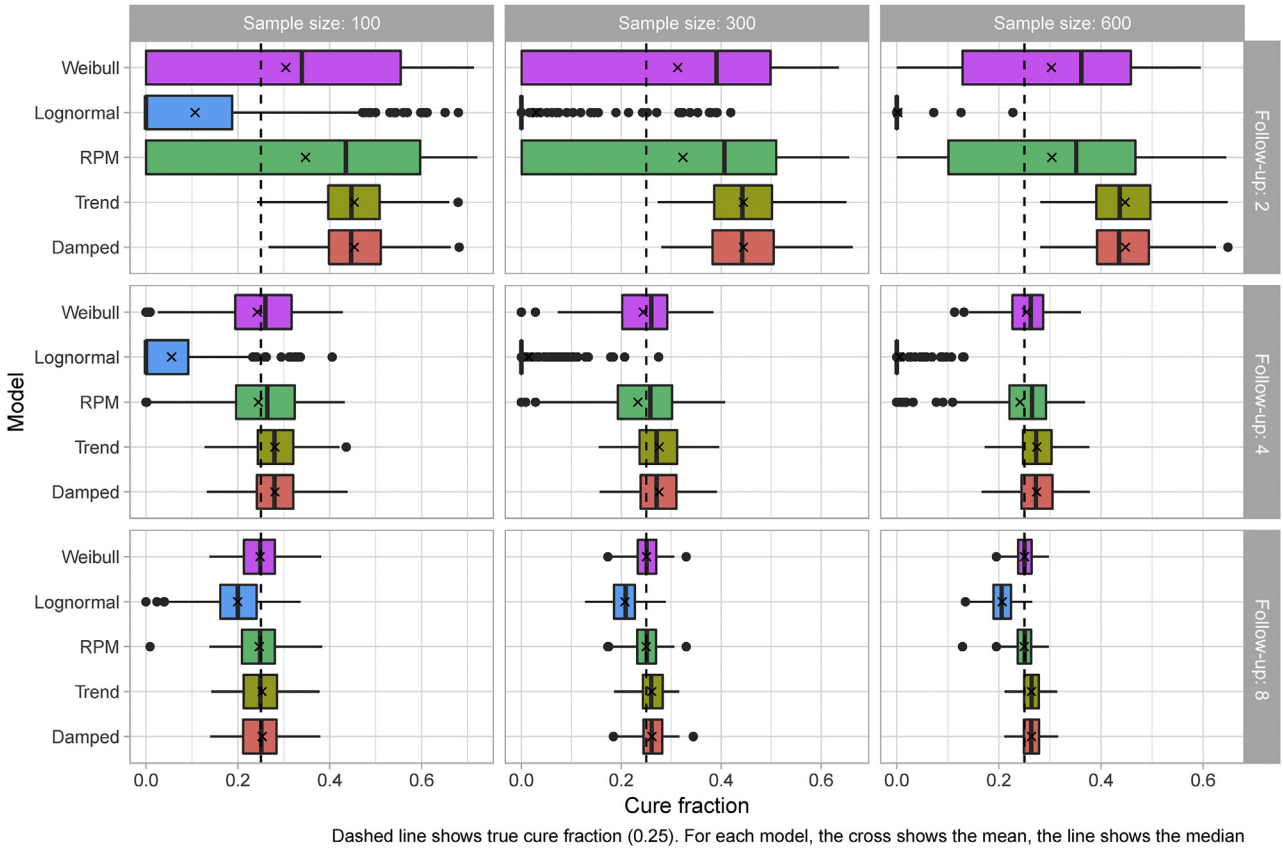
Models Without a Cure Fraction

Visual extrapolations from the models that do not include a cured fraction are provided in Figure 4. Without external evidence, none of the models provided accurate extrapolations for any of the scenarios considered. As such, estimates of bias and MSE are not quantified because there is little merit in identifying the model with the best goodness of fit when none of the models are useful. The poor performance of these models is because, without external evidence, they are unable to describe the unobserved long-term increase in hazards beyond the follow-up period.

Distinct extrapolation patterns may be seen for the current practice models. These patterns arise from the selection of different parametric models; results for the individual models (including the Gompertz) are shown in Appendix Figure 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.05.009. For the scenarios with short-to-medium follow-up, despite all the models (excluding the exponential) having very similar within-sample fit years (range in mean MSE values: 0.01-0.04), extrapolations varied markedly; extrapolations from the Gompertz, Weibull, and gamma all increased, and those from the log-logistic and lognormal decreased, while the generalized gamma provided extrapolations that could increase or decrease. This highlights the danger in basing model choice on within-sample goodness of fit, because models with near-identical within-sample fit could provide qualitatively discrepant extrapolations (with no models providing accurate extrapolations).

The impact of within-sample goodness of fit on model choice was also explored for the choice of using a cure model versus a current practice or RPM. Appendix Table 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.05.009 provides the average absolute improvement in AIC by scenario when using a cure model in preference to its noncure alternative (the focus is on AIC because in practice MSE and bias are unknown, and AIC cannot be calculated for the dynamic models). On average, the use of a cure model led to an improved within-sample fit for all 9 scenarios, with larger improvements in more data-rich scenarios (longer follow-up or increased sample size). Nevertheless, in the scenarios with the least data, there was very little difference between cure fraction models and their corresponding noncure models, suggesting that it would be difficult to choose between the models in these situations. Given that extrapolations varied

Figure 3. Estimates of the cure fraction.



RPM indicates Royston-Parmar model.

markedly between cure and noncure models, this highlights the difficulties with model specification in data-poor scenarios.

Practical Recommendations When Choosing a Survival Model for Extrapolation

A key first step is to assess the plausibility of assuming that a health technology will result in a fraction of patients being "cured" and so only experiencing general population mortality. This is a clinical question, emphasizing the importance of including subject-matter specialists.

If there are clinical reasons to believe that there may be a cure fraction, the results of this study suggest that models that make weak structural assumptions should be preferred. These models typically performed and the correctly specified model while avoiding the sensitivity to model misspecification. This includes the RP cure model and both DCFMs. The sensitivity of extrapolations to the choice of survival model should be assessed in sensitivity analyses. The choice of base-case model should be guided by the plausibility of extrapolations, because good within-sample fit does not guarantee accurate extrapolations. If multiple models provide plausible extrapolations, the use of a damped trend DCFM may be preferable. This is because this model dampens any extrapolated trend, so partly mitigates the danger of extrapolating an incorrect trend. Care should be taken to avoid overinterpreting the estimated cure fraction. Owing to a lack of identifiability, cure fraction models can provide accurate extrapolations even if the estimated cure fraction is wrong. If there is uncertainty about the plausibility of a cure, this should be assessed by using noncure models in sensitivity analyses.

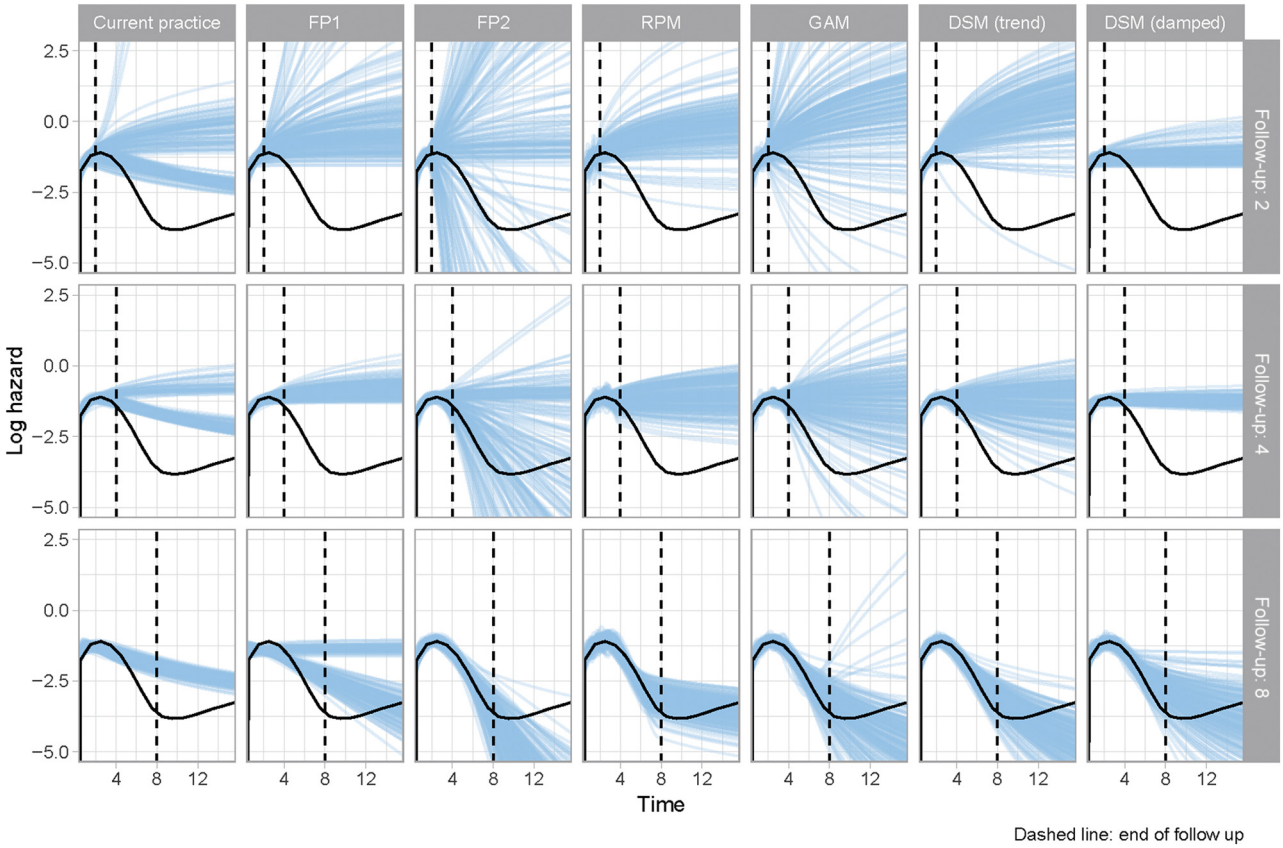
By definition, the accuracy of extrapolations is not known in practice and waiting for longer follow-up is typically not an option. The results of this article emphasize the importance of follow-up on extrapolation accuracy; with short follow-up, there is a danger that no model will provide useful predictions of the future.

Discussion

Recent innovations in health technologies have led to some patients experiencing long-term survival. Generating accurate extrapolations in the presence of a cured fraction is important but challenging. This article compared the within-sample and extrapolation performance of models with and without a cure fraction when fit to data with a cured fraction. The use of models with a cured fraction led to generally acceptable extrapolations. In contrast, noncure models failed to provide plausible extrapolations even at 8 years of follow-up, when over three-quarters of the sample had died. Nevertheless, at the shortest follow-up considered (2 years, with over a third of the sample dead), even the correctly specified Weibull cure model provided poor extrapolations.

Of the cure models considered, the Weibull and lognormal models both make strong structural assumptions about the shape of the disease-specific hazard: that it is either monotonic or has a single turning point, respectively. The results shown here suggest

Figure 4. Estimates of the log-hazard compared with the truth: models without a cure fraction.



DSM indicates dynamic survival model; FP1, first-order fractional polynomial; FP2, second-order fractional polynomial; GAM, generalized additive model; RPM, Royston-Parmar model.

that when these strong structural assumptions are incorrect, resulting extrapolations can be poor. In contrast, the cure RPM and DCFMs make very weak structural assumptions and provided extrapolations that were similar to the correctly specified model across the scenarios considered. In particular, the misspecified damped trend cure fraction model had the lowest overall MSE in 4 of the 9 scenarios and the second lowest in a further 2. As anticipated, the correctly specified Weibull cure model led to the most accurate estimates of the cure fraction for the cure models considered, on average. Nevertheless, this did not result in improved goodness of fit, which is influenced by estimates of both the cure fraction and the hazard function for the uncured fraction. Furthermore, with short follow-up, estimates from the Weibull cure model were also highly variable while within-sample fit was no better than that from a noncure Weibull model. Together these highlight the lack of identifiability for data with a cure fraction; where the cure fraction is unknown, the overall observed hazard function may be described equally well by different combinations of disease-specific hazard functions and cure fractions.

There is limited assessment of the performance of cure fraction models in the literature. An analysis of long-term ovarian cancer registry data, using Weibull cure and lognormal cure models, showed both that estimate of the cure fraction was sensitive to model choice and that the Weibull cure model provided more a plausible estimate despite having a worse within-sample fit. The authors also cautioned that "The estimate of the cure fraction can be unstable when there are a small number at risk toward the end of follow-up."[11] Stedman et al[20] also used registry data and demonstrated that estimates of the cure fraction were sensitive to both the length of follow-up and model choice and to cancer site and stage. The results of this study confirm the findings of these studies, by showing that estimates of the cure fraction are strongly influenced by the length of follow-up and can vary by model. Grant et al[7] performed a case study to assess the performance of Weibull cure models when applied to data representative of a NICE appraisal, with an overall cure fraction of 40% and the majority of uncured people dead by 150 months. Models fit to 40 months of follow-up were found to fit the observed data well but underestimate the overall cure fraction and provide visually poor extrapolations. Their findings support the findings of this study in demonstrating that extrapolation with short follow-up (relative to lifetime follow-up) can provide poor extrapolations, even if within-sample fit is good. A simulation study performed by Rutherford et al[9] demonstrated that, if the truth includes a cure fraction, cure models will provide better extrapolations than noncure models; similar results were observed in this study.

This is the first time that the within-sample and extrapolation performance of DCFMs has been assessed. In total, 69 models were considered (9 cure models, 60 noncure), with 11 models retained for estimating extrapolation performance (4 cure models, 7 noncure). This allows for an assessment of the impact that model misspecification has on both goodness of fit and model selection in practice. This includes both misspecifying the model for disease-specific mortality in a cure fraction model and the

misspecification of not using a cure fraction model for data with a cure fraction. For the former, the use of a misspecified current practice model did not affect within-sample fit but led to very poor and consistently biased extrapolations. In contrast, the use of a misspecified DCFM had little effect. For all model classes, the use of a model without a cured fraction provided extremely poor extrapolations.

A potential limitation of this study is that the results only represent an upper bound on the performance of the cure models in practice. This is because for this study it has been assumed that the survival of the cured patients is known with certainty (the same life tables are used in the data-generating mechanism and the models). In practice, there will be some misspecification; individual patient characteristics and local geographical factors may lead to survival that is different to national life tables. This would affect absolute goodness of fit but is unlikely to affect the relative performance of the models assessed. The hazard function of the uncured population is also relatively simple, arising from a monotonic Weibull model. In reality the hazard function may not be monotonic (eg, because of patient heterogeneity in survival), which again may hamper extrapolation performance. This is likely to most affect the Weibull cure and lognormal cure models, because this study has demonstrated that extensions to current practice models are sensitive to model misspecification. This simulation study also only assessed 1 set of parameters for 1 data-generating mechanism. Future research could continue to assess the goodness of fit of cure models with different data-generating mechanisms or in situations with real data with long follow-up where a proportion of the sample are known to be cured. This could include situations where the "cured" fraction have a mortality that is persistently elevated compared with the general population. In addition, it would be beneficial to know whether cure versions of RPMs are affected by misspecification (do not include the true model as a special case). This article has illustrated the impact of length of follow-up on extrapolation performance. Further research to identify the situations when there is sufficient follow-up to provide reliable extrapolations would be valuable, particularly for situations such as rare diseases when it may be difficult to obtain large sample sizes or long follow-up. Future studies could also expand the data-generating mechanism to consider the impact of disease progression on both survival and censoring, as was modeled in the study of Grant et al.[7]

Conclusions

The presence of a cure fraction creates complex hazard patterns that can pose a challenge for extrapolation. Extensions of current practice models to incorporate a cure fraction work well if they match the true data-generating mechanism but can provide poor results otherwise. Dynamic models with a cure fraction generally performed and the correctly specified model, while avoiding the sensitivity to model misspecification. For all the models evaluated, incorrectly omitting the cure fraction led to very poor extrapolations. When the truth did include a cure fraction, all the cure models provided poor extrapolations at the shortest follow-up considered, and in the data-poor scenarios, there was little difference in the within-sample fit of cure and noncure models. Hence, incorporating external data, in the form of general population hazards, can in some situations improve extrapolation performance but it is not guaranteed to do so. It is not a substitute for having both adequate follow-up and subject-matter input into the plausibility of assuming a cure fraction.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2021.05.009.

Article and Author Information

Accepted for Publication: May 25, 2021

Published Online: xxxx

doi: https://doi.org/10.1016/j.jval.2021.05.009

Author Affiliations: School of Health and Related Research, The University of Sheffield, Sheffield, England, UK (Kearns, Stevenson, Triantafyllopoulos); Centre for Health Economics, The University of York, York, England, UK (Manca).

Correspondence: Benjamin Kearns, MSc, PhD, School of Health and Related Research, The University of Sheffield, Regent Court (ScHARR), 30 Regent St, Sheffield, England, United Kingdom S1 4DA. Email: b.kearns@sheffield.ac.uk

Author Contributions: *Concept and design*: Kearns, Stevenson, Manca
Acquisition of Data: Kearns
Analysis and interpretation of data: Kearns, Stevenson, Triantafyllopoulos, Manca
Drafting of the manuscript: Kearns
Critical revision of paper for important intellectual content: Stevenson, Manca
Statistical analysis: Kearns, Triantafyllopoulos
Obtaining funding: Kearns
Supervision: Stevenson, Triantafyllopoulos, Manca

Conflict of Interest Disclosures: This work reports the results of a PhD. Dr Kearns reported receiving funding for salary paid whilst undertaking the PhD by the National Institute for Health Research Doctoral Research Fellowship (DRF-2016-09-119) during the conduct of this study. Dr Manca is an editor for Value in Health and had no role in the peer review process of this article. No other disclosures were reported.

Funding/Support: The views expressed are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research, or the Department of Health and Social Care.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES

1. Kearns B, Stevens J, Ren S, Brennan A. How uncertain is the survival extrapolation? A study of the impact of different parametric survival models on extrapolated uncertainty about hazard functions, lifetime mean survival and cost effectiveness. *Pharmacoeconomics*. 2020;38(2):193–204.
2. Bullement A, Meng Y, Cooper M, et al. A review and validation of overall survival extrapolation in health technology assessments of cancer immunotherapy by the National Institute for Health and Care Excellence: how did the initial best estimate compare to trial data subsequently made available? *J Med Econ*. 2019;22(3):205–214.
3. Bullement A, Willis A, Amin A, Schlichting M, Hatswell AJ, Bharmal M. Evaluation of survival extrapolation in immuno-oncology using multiple pre-planned data cuts: learnings to aid in model selection. *BMC Med Res Methodol*. 2020;20(1):1–14.
4. Ouwens MJNM, Mukhopadhyay P, Zhang Y, Huang M, Latimer N, Briggs A. Estimating lifetime benefits associated with immuno-oncology therapies: challenges and approaches for overall survival extrapolations. *Pharmacoeconomics*. 2019;37(9):1129–1138.
5. Gibson E, Koblbauer I, Begum N, et al. Modelling the survival outcomes of immuno-oncology drugs in economic evaluations: a systematic approach to data analysis and extrapolation. *Pharmacoeconomics*. 2017;35(12):1257–1270.
6. Bell Gorrod H, Kearns B, Thokala P, et al. A review of survival analysis methods used in NICE technology appraisals of cancer treatments: consistency, limitations, and areas for improvement. *Med Decis Mak*. 2019;39(8):899–909. In press.

7. Grant TS, Burns D, Kiff C, et al. A case study examining the usefulness of cure modelling for the prediction of survival based on data maturity. *Pharmacoeconomics*. 2020;38(4):385–395.
8. The NICE methods of health technology evaluation: the case for change. National Institute for Health and Care Excellence. https://www.bioindustry.org/uploads/assets/e9c9093d-8a4b-4e1e-b01590df056d7f6a/BIA-response-to-the-NICE-methods-of-health-technology-evaluation-the-case-for-change.pdf. Accessed December 18, 2020.
9. Rutherford MJ, Lambert PC, Sweeting MJ, et al. NICE DSU technical support document 21: flexible methods for survival analysis. Decision Support Unit, University of Sheffield. http://nicedsu.org.uk/wp-content/uploads/2020/11/NICE-DSU-Flex-Surv-TSD-21_Final_alt_text.pdf. Accessed December 18, 2020.
10. Tisagenlecleucel for treating relapsed or refractory B-cell acute lymphoblastic leukaemia in people aged up to 25 years. National Institute for Health and Care Excellence. https://www.nice.org.uk/guidance/ta554. Accessed December 18, 2020.
11. Lambert PC, Thompson JR, Weston CL, Dickman PW. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*. 2006;8(3):576–594.
12. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;48(11):2074–2102.
13. National life tables, United Kingdom statistical bulletins. Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/nationallifetablesunitedkingdom/previousReleases. Accessed December 18, 2020.
14. Gallacher D, Auguste P, Connock M. How do pharmaceutical companies model survival of cancer patients? A review of NICE single technology appraisals in 2017. *Int J Technol Assess Health Care*. 2019;35(2):160–167.
15. Kearns B, Stevenson M, Triantafyllopoulos K, Manca A. Generalized linear models for flexible parametric modeling of the hazard function. *Med Decis Making*. 2019;39(7):867–878.
16. Jakobsen LH, Andersson TM, Biccler JL, El-Galaly TC, Bøgsted M. Estimating the loss of lifetime function using flexible parametric relative survival models. *BMC Med Res Methodol*. 2019;19(1):23.
17. 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.
18. Jackson CH. flexsurv: a platform for parametric survival modeling in R. *J Stat Softw*. 2016;70:i08.
19. Wood SN. *Generalized Additive Models: An Introduction With R*. 2nd ed. Boca Raton, FL: CRC press; 2017.
20. Stedman MR, Feuer EJ, Mariotto AB. Current estimates of the cure fraction: a feasibility study of statistical cure for breast and colorectal cancer. *J Natl Cancer Inst Monogr*. 2014;2014(49):244–254.