ORIGINAL RESEARCH ARTICLE

Retrospective Comparison of Survival Projections for CAR T-Cell Therapies in Large B-Cell Lymphoma

Elisabeth F. P. Peterse[1] · Elisabeth J. M. Verburg-Baltussen[1] · Alexa Stewart[2] · Fei Fei Liu[3] · Christopher Parker[4] · Maarten Treur[1] · Bill Malcolm[4] · Sven L. Klijn[3]

Abstract

Background Durable remission has been observed in patients with relapsed or refractory (R/R) large B-cell lymphoma (LBCL) treated with chimeric antigen receptor (CAR) T-cell therapy. Consequently, hazard functions for overall survival (OS) are often complex, requiring the use of flexible methods for extrapolations.

Objectives We aimed to retrospectively compare the predictive accuracy of different survival extrapolation methods and evaluate the validity of goodness-of-fit (GOF) criteria-based model selection for CAR T-cell therapies in R/R LBCL.

Methods OS data were sourced from JULIET, ZUMA-1, and TRANSCEND NHL 001. Standard parametric, mixture cure, cubic spline, and mixture models were fit to multiple database locks (DBLs), with varying follow-up durations. GOF was assessed using the Akaike information criterion and Bayesian information criterion. Predictive accuracy was calculated as the mean absolute error (MAE) relative to OS observed in the most mature DBL.

Results For all studies, mixture cure and cubic spline models provided the best predictive accuracy for the least mature DBL (MAE 0.013−0.085 and 0.014−0.128, respectively). The predictive accuracy of the standard parametric and mixture models showed larger variation (MAE 0.024−0.162 and 0.013−0.176, respectively). With increasing data maturity, the predictive accuracy of standard parametric models remained poor. Correlation between GOF criteria and predictive accuracy was low, particularly for the least mature DBL.

Conclusions Our analyses demonstrated that mixture cure and cubic spline models provide the most accurate survival extrapolations of CAR T-cell therapies in LBCL. Furthermore, GOF should not be the only criteria used when selecting the optimal survival model.

1 Introduction

Clinical trials assessing the efficacy of new oncology treatments often have limited follow-up at the time of submission to health technology assessment (HTA) agencies. An understanding of patient outcomes beyond the maximum follow-up can have salient implications for future treatment options, as well as a profound impact on cost of care. Survival modeling can be a useful tool in extrapolating future outcomes based on available data, and has been used to inform cost-effectiveness analyses [1]. To determine the full clinical and economic value of a new treatment,

extrapolation of observed survival is routinely required, and different extrapolation methods may lead to substantive variation in total survival gain and consequent value estimation. To effectively compare the cost effectiveness of available treatment options, it is critical to determine the predictive accuracy of the statistical models used to measure and compare survival. The most common extrapolation approach is to use standard parametric survival models (e.g., exponential, Weibull, gamma) [2]; however, these models might be unsuitable to capture survival after chimeric antigen receptor (CAR) T-cell therapy, given the potential for durable remission (e.g., 5-year survival) [3–5]. Durable remission has been observed in patients with relapsed or refractory (R/R) large B-cell lymphoma (LBCL), after treatment with salvage chemotherapy followed by high-dose chemotherapy and autologous stem cell transplantation [6] or CAR T-cell therapies [7–9]. Durable remission can be observed in the Kaplan–Meier (KM) curve, where a plateau indicates that

Elisabeth F.P. Peterse, Elisabeth J.M. Verburg-Baltussen and Alexa Stewart are no longer employed by OPEN Health; and Christopher Parker is no longer employed by Bristol Myers Squibb. Affiliation at the time the research was conducted.

Extended author information available on the last page of the article

Key Points for Decision Makers

Standard parametric survival models may not be suitable for survival extrapolation in cases of complex hazard functions, such as those seen in patients treated with CAR T-cell therapies. Several alternative survival extrapolation methods have been developed, but current guidance on which alternative method to use under which circumstances is limited.

Our analyses demonstrated that mixture cure and cubic spline models generate more accurate survival predictions than standard parametric and mixture models for CAR T-cell therapies in large B-cell lymphoma, even with limited follow-up data.

Goodness-of-fit criteria demonstrated a low correlation with predictive accuracy, suggesting that limited weight should be assigned to these criteria when selecting the optimal survival model.

the hazard converges towards the hazard in the general population [10].

Three autologous CAR T-cell therapies have received regulatory approval for R/R LBCL: axicabtagene ciloleucel (axi-cel; 2-year survival rate of 51%) [11], tisagenlecleucel (2-year survival probability of 40%) [12], and lisocabtagene maraleucel (liso-cel; 2-year survival probability of 51%) [13]. To capture the plateau observed in the KM curve for overall survival (OS), a hybrid survival model using a 1-knot spline model with a cure point was used in HTA submissions of tisagenlecleucel, whereas a mixture cure model was used in HTA submissions of axi-cel [14, 15]. Although the National Institute for Health and Care Excellence technical support document (NICE TSD) 21 provides a clear description of these flexible methods [16], it does not provide sufficient guidance on which method to use under what circumstances [17]. Goodness-of-fit (GOF) criteria are frequently used for model selection [2]. These criteria assess how well a model captures the observed survival, but a good statistical fit to the data does not necessarily yield an accurate, or even plausible, estimate of long-term survival, especially if data are immature [18]. This problem is exacerbated when modeling more complex hazard functions, as seen in R/R LBCL [19]. Therefore, our study aimed to retrospectively compare the predictive accuracy of different survival extrapolation methods (i.e., standard parametric distributions, mixture cure, cubic spline, and mixture models) and evaluate the validity of GOF criteria-based model selection using data

from various follow-up time points from CAR T-cell therapies in R/R LBCL.

2 Methods

The JULIET (tisagenlecleucel, NCT02445248), ZUMA-1 (axi-cel, NCT02348216), and TRANSCEND NHL 001 (TRANSCEND; liso-cel, NCT02631044) trials evaluated the efficacy of CAR T-cell therapies in patients with R/R LBCL who had received two or more prior lines of systemic treatment. Various survival extrapolation approaches were applied to the three least mature database locks (DBLs) of JULIET and ZUMA-1 and the two least mature DBLs of TRANSCEND (Table 1), in which GOF criteria were evaluated to represent the anticipated data available at the time of HTA filing. Subsequently, predictive accuracy was assessed by comparing survival predictions based on the least mature DBLs to observed survival from the most mature DBLs (internal model validation) and by comparing the extrapolated survival curves to external data (external model validation). Lastly, mean lifetime survival (evaluated as survival until 100 years of age) was quantified for all models.

2.1 Data

Details of the JULIET, ZUMA-1, and TRANSCEND trials have been described elsewhere [7–9]. An overview of these trials and the included DBLs are provided in Table 1. For JULIET [8, 12, 19, 20] and ZUMA-1 [9, 11, 21, 22], DBLs were based on publicly available data. KM curves from the different DBLs were digitized using WebPlotDigitizer [23]. The widely accepted and validated algorithm developed by Guyot et al. [24] was used to reconstruct patient-level data. It is an iterative numerical method solving the inverted KM equations, using the digitized KM curve coordinates, the numbers at risk, and total number of events. Potential reliability issues of the algorithm, related to low resolution of KM curves and a lack of reported numbers of events and numbers at risk, did not apply to either trial. The associated KM curves were visually compared with the original KM curves. Furthermore, summary statistics such as survival probabilities and median survival were compared with the original KM data to ensure accuracy of this process. For TRANSCEND, the DBLs used in this study matched the DBLs used (internally) by Bristol Myers Squibb, and patient-level data of these DBLs were available.

2.2 Survival Extrapolation Methods

Four different extrapolation methods were used: standard parametric distributions, mixture cure models, cubic spline models, and mixture models. Standard parametric

Table 1 Studies and DBLs included in the study

Study	Intervention	Patients receiving CAR T, n	Recruitment period	First DBL, date (min. FU)	Second DBL, date (min. FU)	Third DBL, date (min. FU)	Most mature DBL, date (min. FU)
JULIET	Tisagenlecleucel	115[a]	Jul 2015–Dec 2017[a]	8 Dec 2017[a] (0 months)	21 May 2018 (5 months)	1 Jul 2019 (19 months)	20 Feb 2020 (26 months)
ZUMA-1	Axi-cel	108[b]	May 2015–Sep 2016	11 Aug 2017[b] (11 months)	11 Aug 2018 (23 months)	11 Aug 2019 (35 months)	11 Aug 2020 (47 months)
TRAN- SCEND NHL 001	Liso-cel	256	Jan 2016–Jul 2019	12 Aug 2019 (1 month)	19 Jun 2020 (11 months)	–	4 Jan 2021 (18 months)

Axi-cel axicabtagene ciloleucel, *CAR T* chimeric antigen receptor T-cell therapy, *DBLs* database locks, *liso-cel* lisocabtagene maraleucel, *min. FU* minimum follow-up

[a]At the 2017 JULIET DBL, 111 patients had received tisagenlecleucel. This differs from the 2018 data-cut and onward, wherein 115 patients had received tisagenlecleucel

[b]In the 2017 ZUMA-1 DBL, data were presented on 108 patients who received axi-cel across phase I (7 patients) and phase II (101 patients). From the 2018 data-cut and onward, data were available from 101 patients. Phase II ran from November 2015 to September 2016

distributions were originally introduced for extrapolating survival data. Cubic spline models were introduced to overcome the inability of standard parametric models to adequately capture more complicated hazard functions seen in applied studies [15]. Mixture models allow for two or more latent clusters of patients with different survival profiles. Cure models are used in a situation where some individuals will never experience cause-specific death (i.e., their disease-related hazard rate will be zero and the OS curve plateaus at a non-zero value). The evaluated standard parametric distributions were exponential, Gompertz, Weibull, log-normal, log-logistic, gamma, and generalized gamma. Cubic spline models can accommodate more complex hazard patterns, with two or more turning points that can capture multiple important changes in the slope. Evaluated splines had one or two intermediate knots, equally distributed over uncensored log times, using a hazard, odds, or probit model for the transformed survival function [25]. The mixture cure and the mixture models used each represent the population as a mixture of two subpopulations. In the mixture cure models, those cured of the disease were assumed to have survival equal to that of the general population, and survival for those not cured was modeled using one of the standard parametric distributions, resulting in seven different mixture cure models [16]. In the mixture models, survival of both subpopulations was modeled using one of the standard parametric distributions, resulting in a total of 28 different mixture models. Mixture cure, cubic spline, and mixture models were included as per the NICE TSD 21 [26] and the OPTIOM guideline [17]. Standard parametric models were included for reference, as they are the most commonly used extrapolation method [2]. Piecewise models and response-based landmark models were also considered, but these were excluded because of the lack of a clinical basis for selecting

a cut-off point and incomplete response-stratified survival data, respectively. Analyses were conducted in R [27], using the flexsurv and muhaz packages [26, 28]. All nonconverging models and models with a nonpositive Hessian matrix were excluded. For the application of background mortality, a relative survival framework was used for all models in line with NICE TSD 21 [16], based on United States 2018 lifetables [29] adjusted to the corresponding trial for age and sex.

2.3 Internal Model Validation

GOF criteria (Akaike information criterion [AIC] [30] and Bayesian information criterion [BIC] [31]) were assessed for all models. Subsequently, the most mature DBLs of the three trials were used for internal validation of the survival extrapolations of all fitted models, using the mean absolute error (MAE) and root mean squared error (RMSE). MAE and RMSE were evaluated per observed event from the date of CAR T-cell infusion ($t = 0$) to the last event of the most mature DBL.

2.4 External Model Validation

External validation of the survival extrapolations was performed by comparing the extrapolated survival curves to external data from SCHOLAR-1, a retrospective study evaluating outcomes in patients with refractory diffuse LBCL treated with salvage chemotherapy [6]. It was assumed that survival and cure rates with CAR T-cell therapy would be higher than with salvage chemotherapy. Therefore, two validation checks were performed based on SCHOLAR-1 data. First, models with a 5-year survival estimate lower than that observed in SCHOLAR-1 (5-year survival estimate of 16%), were deemed implausible. Second, mixture cure models with

an estimated cure fraction <16% were also deemed implausible. In addition to the validation using the SCHOLAR-1 data, the extrapolations were validated by checking whether the disease-specific hazard converged to 0 after 5 years.

3 Results

The most mature survival data and the survival data used to fit survival models are visualized in electronic supplementary material (ESM) Figs. S1 and S2 and Fig. 1 for JULIET, ZUMA-1, and TRANSCEND, respectively. The importance of selecting an appropriate survival model was demonstrated by the large variation seen in the 5-year survival estimates of the different extrapolation models, ranging from 0 to 45%, 10 to 55%, and 0 to 47% over all DBLs of JULIET, ZUMA-1, and TRANSCEND, respectively (Table 2 and ESM Tables S1−S7). Survival and hazard plots of all models fitted to the different DBLs of the different trials are reported in the Appendix (ESM Figs. S3−S34). All models converged, except for 36% of the mixture models. The nonconvergence of some mixture models was not limited to specific distributions, trials, or DBLs, although nonconvergence was the highest in the earliest DBL (39%).

3.1 Performance of Standard Parametric Models

Standard parametric models had poor predictive accuracy, as measured using the MAE. Among these, the generalized gamma distribution performed best and the exponential distribution performed worst in all trials and most DBLs. Notably, the generalized gamma distribution is the most flexible model due to its three parameters and resulting ability to capture turning points in the hazard in contrast to the exponential distribution, which only has one parameter and assumes a constant hazard. The MAE results from the JULIET trial showed poor performance of the standard parametric models compared with the other extrapolation methods (Fig. 2). MAE results from the ZUMA-1 and TRANSCEND trials are shown in ESM Figs. S35 and S36, respectively. With increasing data maturity, the improvement in performance of the standard parametric models was relatively minor compared with the more flexible models. For example, in JULIET, the MAE of standard parametric models was between 0.027 and 0.132 for the first DBL and between 0.023 and 0.114 for the third DBL (Fig. 2, ESM Tables S1 and S3). A similar pattern was observed in ZUMA-1 and TRANSCEND (ESM Figs. S35 and S36). The external validation showed that 48%, 24%, and 7% of all standard parametric models fitted to the first, second, and third DBL, respectively, across all trials, projected a 5-year survival rate lower than that observed in SCHOLAR-1, indicating that these models did not pass the external validation checks (Table 2 and ESM Tables S1−S7). In particular, the exponential, Weibull, and gamma distributions performed poorly in the external validation. For example, in the first, second, and third DBLs of the JULIET trial, the exponential distribution estimated a 5-year survival rate of 3%, 4%, and 12%, respectively, all of which fall below the 16% 5-year

Fig. 1 Kaplan–Meier curves of the different DBLs of TRANSCEND NHL 001. Shading indicates 95% confidence intervals, and the table below the Kaplan–Meier curves indicates the number of patients at risk for each time point. *DBL* database lock
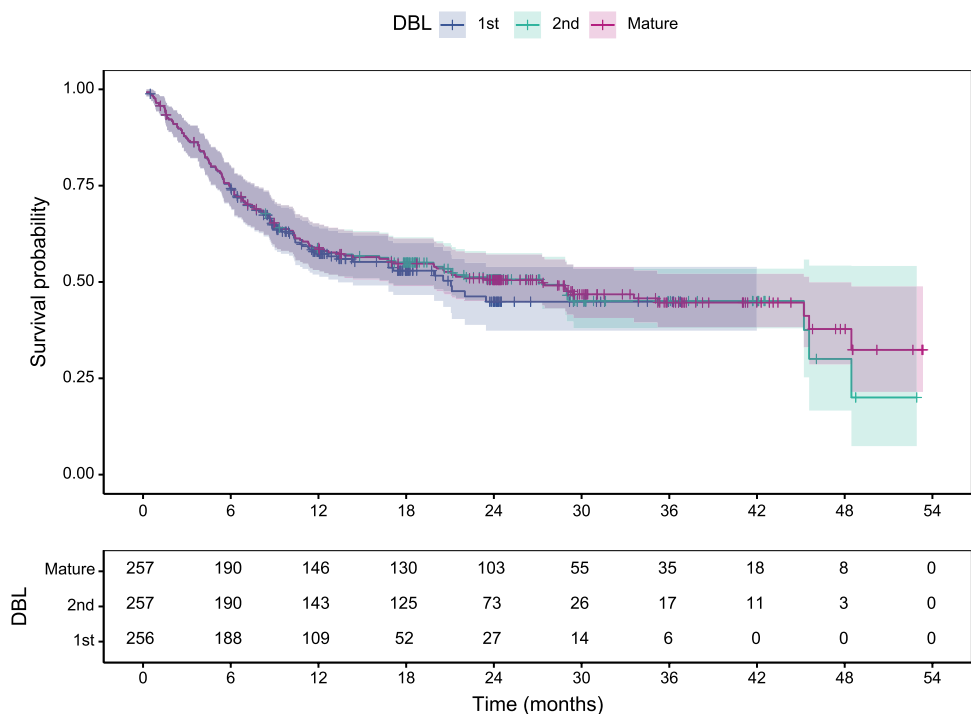
Table 2 Characteristics of models fitted to the first database lock of TRANSCEND NHL 001

Distribution	Group 1, %	Group 2, %[a]	AIC	BIC	MAE	RMSE	5-year survival, %[a]	Mean survival, years
Parametric								
Exponential			986.6	990.2	0.091	0.110	**11.3**	2.25
Gompertz			973.6	980.7	0.024	0.027	38.7	15.57
Weibull			985.3	992.4	0.067	0.083	16.4	2.67
Log-normal			972.3	979.3	0.043	0.055	25.3	5.01
Log-logistic			976.8	983.9	0.055	0.069	22.9	4.81
Gamma			986.8	993.9	0.074	0.092	**14.4**	2.47
Generalized gamma			972.6	983.2	0.030	0.037	30.0	7.08
Mixture cure								
Weibull/cure	53.8	46.2	969.8	980.4	0.019	0.032	46.2	18.70
Exponential/cure	58.3	41.7	972.2	979.2	0.022	0.027	41.9	17.03
Log-logistic/cure	61.2	38.8	969.0	979.6	0.015	0.024	41.6	16.23
Gompertz/cure	54.1	45.9	972.7	983.3	0.021	0.033	45.9	18.57
Gamma/cure	54.2	45.8	969.1	979.8	0.018	0.031	45.9	18.56
Generalized gamma/cure	55.9	44.1	970.7	984.9	0.016	0.028	44.3	17.93
Log-normal/cure	65.9	34.1	970.3	981.0	0.019	0.023	38.4	14.56
Cubic spline								
Spline hazard 1			972.0	982.6	0.023	0.029	30.3	5.72
Spline hazard 2			971.1	985.3	0.014	0.018	35.5	8.33
Spline probit 1			972.8	983.5	0.030	0.038	29.1	6.23
Spline probit 2			970.9	985.1	0.015	0.019	35.6	9.20
Spline odds 1			971.6	982.3	0.025	0.031	31.3	7.48
Spline odds 2			971.3	985.4	0.015	0.019	35.9	9.64
Mixture								
Gompertz/Weibull	75.1	24.9	973.2	990.9	0.013	0.020	39.8	15.48
Gompertz/log-normal	56.0	44.0	973.8	991.6	0.016	0.024	41.9	16.70
Gompertz/log-logistic	39.1	60.9	973.0	990.7	0.015	0.024	41.7	16.26
Gompertz/exponential	43.3	56.7	976.2	990.3	0.022	0.027	41.9	17.03
Generalized gamma/Weibull	55.9	44.1	974.7	996.0	0.016	0.028	44.3	17.92
Generalized gamma/log-logistic	84.7	15.3	973.5	994.7	0.013	0.019	38.0	11.19
Generalized gamma/exponential	55.8	44.2	972.7	990.5	0.016	0.028	44.2	17.81
Generalized gamma/gamma	87.2	12.8	973.3	994.6	0.013	0.018	37.3	10.67
Log-logistic/Weibull	61.2	38.8	973.0	990.7	0.015	0.024	41.6	15.85
Log-logistic/log-normal	34.0	66.0	974.3	992.1	0.019	0.023	38.4	14.54
Log-logistic/gamma	50.2	49.8	973.1	990.9	0.015	0.025	42.0	11.50
Gamma/Weibull	53.0	47.0	973.1	990.9	0.017	0.029	44.6	14.71
Gamma/log-normal	37.5	62.5	973.0	990.7	0.014	0.021	39.1	10.83
Gamma/exponential	53.6	46.4	971.1	985.3	0.018	0.030	45.4	17.21
Exponential/Weibull	58.0	42.0	976.2	990.4	0.022	0.027	41.5	15.81
Exponential/log-normal	34.1	65.9	972.3	986.5	0.019	0.023	38.4	14.49

AIC Akaike information criterion, *BIC* Bayesian information criterion, *MAE* mean absolute error, *RMSE* root mean squared error

[a]Models that did not pass the validation checks are indicated in bold (no bold indicates all models passed validation checks). Two validation checks were performed based on SCHOLAR-1 data. First, extrapolated 5-year survival for the chimeric antigen receptor T-cell therapies could not be lower than 16.15%. Second, for the mixture cure model, the estimated cure fraction could not be lower than 16.15%

survival observed in SCHOLAR-1 and making their use clinically and scientifically implausible.

3.2 Performance of Mixture Cure, Cubic Spline, and Mixture Models

The mixture cure models had MAEs across all DBLs of 0.010−0.050, 0.013−0.085, and 0.015−0.024 for JULIET, ZUMA-1, and TRANSCEND, respectively. The cubic spline models had lower MAEs compared with the standard parametric models, including ranges across all DBLs of 0.013−0.035, 0.023−0.128, and 0.014−0.030 for JULIET, ZUMA-1, and TRANSCEND, respectively. Among the mixture models that converged, there was a large variation in the MAE in the first DBLs of JULIET and ZUMA-1. Even with this variation, especially in the more mature data, mixture models performed better than the standard parametric models. The MAE of the mixture models was 0.100−0.123, 0.014−0.176, and 0.012−0.022 for JULIET, ZUMA-1, and TRANSCEND, respectively (Table 2 and ESM Tables S1−S7).

With increasing data maturity, mixture cure and cubic spline models continued to have high predictive accuracy (ESM Tables S1−S8). Mixture models showed an increase in predictive accuracy with increasing length of follow-up. For example, the mixture models resulted in an MAE of 0.025−0.176 in the first ZUMA-1 DBL, which decreased

(i.e., improved) to 0.014−0.050 and 0.029−0.044 in the second and third DBLs, respectively (ESM Fig. S35 and ESM Tables S4−S6). All mixture cure and cubic spline models generated 5-year survival estimates that exceeded SCHOLAR-1. For the mixture models, 24%, 2%, and 6% of the models did not meet this requirement for JULIET, ZUMA-1, and TRANSCEND, respectively.

Spline models fitted to the earliest DBLs estimated lower 5-year survival than mixture cure models (ESM Figs. S37−S39). However, the slope in the tail of the spline models became very shallow, particularly with more mature data, nearly converging to mixture cure models for the period for which data were available (Fig. 3 and ESM Figs. S37−S39). This trend was also observed in the minimum clinical validation threshold 5-year survival estimates of the mixture cure and cubic spline models. The mixture cure models showed little change in mean 5-year survival estimates with increasing data maturity, but the differences between the 5-year survival estimates of the different mixture cure models decreased (Table 2 and ESM Tables S1−S7). In contrast, the 5-year survival estimates of the cubic spline models gradually increased as data maturity increased. For example, in ZUMA-1, 5-year survival estimates for the first, second, and third DBL were 25−55%, 47−51%, and 47−51%, respectively, for mixture cure models, and 16−34%, 34−45%, and 43−47%, respectively, for cubic spline models (ESM Tables S4−S6). Despite this convergence, the mean lifetime survival estimated by mixture cure and cubic spline models differed substantially. For example, for models fitted to the third DBL of ZUMA-1 (Fig. 3), mean lifetime survival estimates ranged from 19.7 to 21.3 years for mixture cure models and from 12.0 to 18.2 years for cubic spline models (ESM Table S6).

3.3 Correlation Between Goodness of Fit and Predictive Accuracy

Model selection using GOF criteria did not maximize predictive accuracy, especially for early DBLs. This was demonstrated by the limited correlation observed between AIC/BIC and MAE (Fig. 4 and ESM Figs. S40−S47). Similar results were obtained when comparing AIC/BIC values with RMSE (Table 2 and ESM Tables S1−S7). For example, for the first DBL of TRANSCEND, the three models with the lowest BIC that passed external validation checks were the exponential distribution, exponential mixture cure model, and log-logistic mixture cure model. In contrast, the models with the lowest MAE were the generalized gamma/log-logistic, generalized gamma/gamma, and Gompertz/Weibull mixture models (Table 2). Among the models that passed
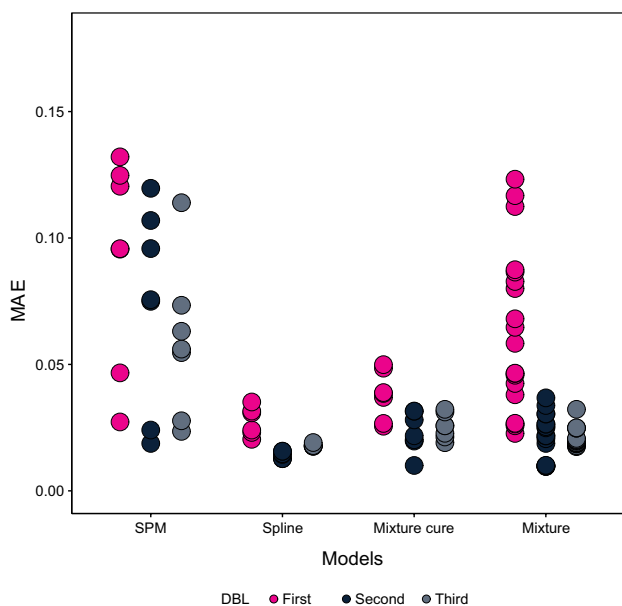

Fig. 2 MAE of the different models fitted to the first, second, or third DBL of JULIET. Lower MAE scores indicate better predictive accuracy. *DBL* database lock, *MAE* mean absolute error, *SPM* standard parametric models
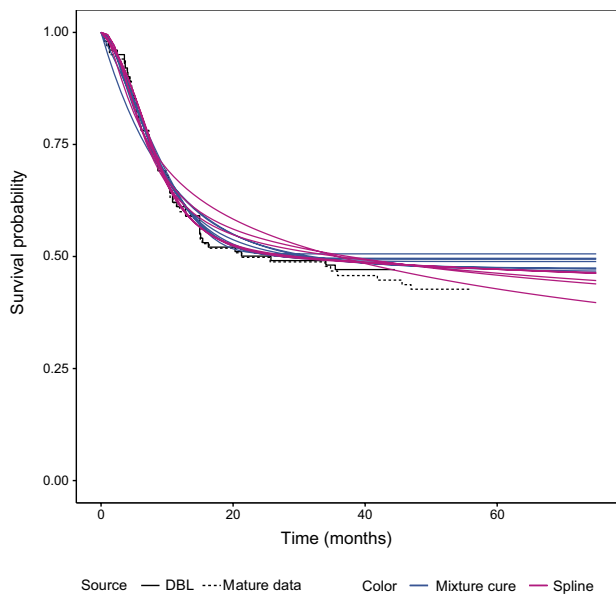
Fig. 3 Comparison of the cubic spline models and the mixture cure models fitted to the third DBL of ZUMA-1. *DBL* database lock
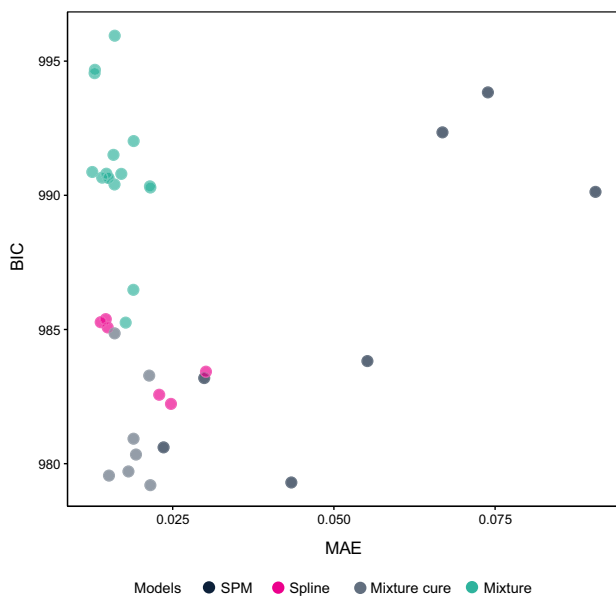

Fig. 4 Correlation between MAE and BIC in the models fitted to the first DBL of TRANSCEND NHL 001. *BIC* Bayesian information criterion, *DBL* database lock, *MAE* mean absolute error, *SPM* standard parametric models

external validation checks, mean lifetime survival estimates for the models with the lowest BIC were 7.6, 4.2, and 17.0 years, and 8.7, 16.7, and 15.5 years for the models with the lowest MAE, for the first DBLs of JULIET, ZUMA-1, and TRANSCEND, respectively (ESM Tables S1 and S4, and Table 2). This demonstrates that GOF criteria-based model

selection can result in selecting models that substantially overestimate or underestimate mean lifetime survival.

4 Discussion

Our analyses demonstrated that mixture cure and cubic spline models can achieve a higher predictive accuracy for OS of CAR T-cell therapies in R/R LBCL compared with standard parametric and mixture models, even when data are relatively less mature. Furthermore, GOF criteria demonstrated a low correlation with predictive accuracy, suggesting that other criteria should be prioritized for model selection. Approximately one-third of the mixture models did not converge, and among the models that did converge, there was much variation in predictive accuracy for early DBLs. Both the nonconvergence and the variance in predictive accuracy are likely a result of the low number of patients and the low number of events, suggesting that mixture models may not be suitable for survival extrapolations of early DBLs. Cubic spline models, especially when fitted to more mature data, had a hazard almost identical to that of the general population for the period beyond the trial follow-up, behavior typically seen in cure models. One limitation of cubic spline models may be the subjectivity of the number and placement of the knots. In this study, we evaluated cubic spline models with one or two intermediate knots, where the knots were equally distributed over uncensored log time. As durable remission can be achieved in R/R LBCL, using mixture cure models is clinically justified, and therefore may be a more appropriate survival extrapolation method for CAR T-cell therapies in R/R LBCL. It has been suggested that long-term survivors in LBCL have minor excess mortality [32]. In the NICE submissions of tisagenlecleucel, a standardized mortality ratio of 1.09 was applied to background mortality to model survival after the cure point to reflect this [15, 20]. Although a relative survival framework was adopted, our model estimations did not account for excess mortality for long-term survivors, which is in line with the updated submission of axi-cel [14]. Incorporating a standardized mortality ratio of 1.09 would only have a minor impact on the overall lifetime survival estimates and it would not affect relative model performance, as the adjustments would apply equally to all models.

One major strength of our study is that we compared several extrapolation methods for multiple DBLs from three different trials of CAR T-cell therapies in R/R LBCL, allowing us to study to what extent the optimal extrapolation method differs between trials and DBLs in a similar clinical setting (i.e., cellular therapies in the same indication). However, there are two limitations of note in our study. First, we did not consider progression-free survival (PFS). Survival

extrapolations on OS may underestimate the cure fraction owing to the lack of maturity of the OS data. In practice, a plateau in PFS is reached earlier than for OS, which can provide information about the cure fraction. However, in JULIET and TRANSCEND, disease progression was determined by an independent review committee, whereas in ZUMA-1, disease progression was determined by investigator assessment. This difference in disease progression assessment creates bias when analyzing PFS data across the three trials, which is why we did not consider PFS in this study. Second, even though we used the most mature data available at the time these analyses were conducted, data maturity was still limited, with minimum follow-up times of 26, 47, and 18 months for JULIET, ZUMA-1, and TRANSCEND, respectively. Since the time of these data analyses, longer follow-up data for ZUMA-1 have been presented [33]. When more mature data for all three trials become available, stronger conclusions about predictive accuracy may be drawn than were drawn based on the current study.

In a recent study, Vadgama et al. [34] compared the performance of standard parametric models, mixture cure models, cubic spline models, and nonmixture cure models fitted to the first, second, and third DBLs of ZUMA-1. Our study is more comprehensive, as it uses data from JULIET, ZUMA-1, and TRANSCEND and also includes mixture models. Furthermore, Vadgama et al. [34] used a different measure to evaluate the performance of the survival extrapolation methods, as they compared predicted survival at 48 months with 48-month survival observed in the most mature data. This method can be used to evaluate the difference between the predicted and observed survival at one specific time point, whereas the MAE and RMSE used in this study account for the difference in the predicted and observed survival over the entire period for which data are available (e.g., from receiving CAR T-cell therapy to the date of the last event in the most mature data). Despite the differences between this study and the study by Vadgama et al. [34], both studies concluded that cure-based models provided the most accurate extrapolations of long-term survival for patients with R/R LBCL treated with CAR T-cell therapies, which is also in line with studies of other immuno-oncology therapies [35, 36]. Additionally, in the final NICE appraisal of axi-cel, the independent academic reviewer group agreed that mixture cure models estimated on the 60-month OS data of ZUMA-1 provided good fits and were rightfully selected to model OS.

5 Conclusions

Standard parametric models are inappropriate for extrapolation of survival data for CAR T-cell therapies in LBCL. Mixture cure and spline models provided the best predictive accuracy, but mixture cure models may be more appropriate given the clinical relevance of their underlying assumptions to LBCL. Furthermore, due to the reduced correlation between GOF criteria and predictive accuracy observed in our analysis, limited weight should be assigned to GOF criteria when selecting the optimal survival model for endpoints with complex hazard functions, particularly when this is due to a subset of patients achieving long-term remission and data are immature. Although this study focused on CAR T-cell therapies in patients with R/R LBCL who received two or more prior systemic treatments, we postulate that the results of the current study may also be applicable to earlier treatment lines and other oncology indications in which lasting remission is observed, as the hazard functions are expected to be similar.

Declarations

References

1. Sussman M, Crivera C, Benner J, Adair N. Applying state-of-the-art survival extrapolation techniques to the evaluation of CAR-T therapies: evidence from a systematic literature review. Adv Ther. 2021;38(8):4178–94.

2. Latimer N. NICE DSU technical support document 14: survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data. 2011. https://www.sheffield.ac.uk/nice-dsu/tsds/flexible-methods-survival-analysis. Accessed 10 Oct 2021.

3. 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.

4. Maruvka YE, Tang M, Michor F. On the validity of using increases in 5-year survival rates to measure success in the fight against cancer. PLoS ONE. 2014;9(7): e83100.

5. Welch HG, Schwartz LM, Woloshin S. Are increasing 5-year survival rates evidence of success against cancer? JAMA. 2000;283(22):2975–8.

6. Crump M, Neelapu SS, Farooq U, et al. Outcomes in refractory diffuse large B-cell lymphoma: results from the international SCHOLAR-1 study. Blood. 2017;130(16):1800–8.

7. Abramson JS, Palomba ML, Gordon LI, et al. Lisocabtagene maraleucel for patients with relapsed or refractory large B-cell lymphomas (TRANSCEND NHL 001): a multicentre seamless design study. Lancet. 2020;396(10254):839–52.

8. Neelapu SS, Locke FL, Bartlett NL, et al. Axicabtagene ciloleucel CAR T-cell therapy in refractory large B-cell lymphoma. N Engl J Med. 2017;377(26):2531–44.

9. Schuster SJ, Bishop MR, Tam CS, et al. Tisagenlecleucel in adult relapsed or refractory diffuse large B-cell lymphoma. N Engl J Med. 2019;380(1):45–56.

10. Lambert PC. Modeling of the cure fraction in survival studies. Stata J. 2007;7(3):351–75.

11. Locke FL, Ghobadi A, Jacobson CA, et al. Long-term safety and activity of axicabtagene ciloleucel in refractory large B-cell lymphoma (ZUMA-1): a single-arm, multicentre, phase 1–2 trial. Lancet Oncol. 2019;20(1):31–42.

12. Jaeger U, Bishop MR, Salles G, et al. Myc expression and tumor-infiltrating T cells are associated with response in patients (Pts) with relapsed/refractory diffuse large B-cell lymphoma (r/r DLBCL) treated with tisagenlecleucel in the JULIET trial [poster 1194]. In: Presented at the 62nd ASH Annual Meeting (virtual), 5–8 Dec 2020.

13. Abramson JS, Palomba ML, Gordon LI, et al. Two-year follow-up of TRANSCEND NHL 001, a multicenter phase 1 study of lisocabtagene maraleucel (liso-cel) in relapsed or refractory (R/R) large B-cell lymphomas (LBCL) [poster 2840]. In: Presented at the 63rd ASH Annual Meeting, Atlanta, GA, 11–14 Dec 2021.

14. National Institute for Health and Care Excellence. Final appraisal document - axicabtagene ciloleucel for treating diffuse large B-cell lymphoma and primary mediastinal B-cell lymphoma after 2 or more systemic therapies. 2018. https://www.nice.org.uk/guidance/ta872/documents/final-appraisal-determination-document. Accessed 1 Nov 2021.

15. National Institute for Health and Care Excellence. Final appraisal document - tisagenlecleucel for treating relapsed or refractory diffuse large B-cell lymphoma after 2 or more systemic therapies. 2019. https://www.nice.org.uk/guidance/ta567/documents/final-appraisal-determination-document. Accessed 1 Nov 2021.

16. Rutherford MJ, Lambert PC, Sweeting MJ, et al. NICE DSU technical support document 21. Flexible methods for survival analysis. 2020. https://nicedsu.sites.sheffield.ac.uk/tsds/flexible-methods-for-survival-analysis-tsd. Accessed 10 Oct 2021.

17. Palmer S, Borget I, Friede T, et al. A guide to selecting flexible survival models to inform economic evaluations of cancer immunotherapies. Value Health. 2023;26(2):185–92.

18. Kearns B, Stevenson MD, Triantafyllopoulos K, Manca A. Comparing current and emerging practice models for the extrapolation of survival data: a simulation study and case-study. BMC Med Res Methodol. 2021;21(1):263.

19. Norwegian Medicines Agency. Tisagenlecleucel (Kymriah) for the treatment of second or later relapsed/refractory diffuse large B cell lymphoma (DLBCL). https://legemiddelverket.no/Documents/Offentlig%20finansiering%20og%20pris/Metodevurderinger/K/Kymriah_DLBCL_2019.pdf. Accessed 1 Nov 2021.

20. Australian Government Medical Services Advisory Committee. Application No. 1519.1- Tisagenlecleucel (CTL019) for treatment of relapsed or refractory diffuse large B-cell lymphoma (DLBCL). http://www.msac.gov.au/internet/msac/publishing.nsf/Content/A2B10F9A03293BC8CA2583CF001C7A4D/$File/1519.1%20Final%20updated%20PSD%20Nov%2019_redacted.pdf. Accessed 1 Nov 2021.

21. Jacobson C, Locke FL, Ghobadi A, et al. Long-term survival and gradual recovery of B cells in patients with refractory large B cell lymphoma treated with axicabtagene ciloleucel [poster 1187]. In: Presented at the 62nd ASH Annual Meeting (virtual), 5–8 Dec 2020.

22. Norwegian Medicines Agency. Axicabtagene ciloleucel (Yescarta) for the treatment of second or later relapsed/refractory diffuse large B cell lymphoma (DLBCL) and primary mediastinal large B-cell lymphoma (PMBCL). https://legemiddelverket.no/Documents/Offentlig%20finansiering%20og%20pris/Metodevurderinger/Y/Yescarta_%20DLBCL-%20PMBLC_2020.pdf. Accessed 1 Nov 2021.

23. Rohatgi A. WebPlotDigitizer. July, 2020. https://automeris.io/WebPlotDigitizer/. Accessed 1 Nov 2021.

24. 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).

25. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med. 2002;21(15):2175–97.

26. Hess K, Gentleman R. muhaz: hazard function estimation in survival analysis. 2021. https://CRAN.R-project.org/package=muhaz. Accessed 18 Jan 2021.

27. R Core Team. R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. 2021. https://www.R-project.org. Accessed 18 Jan 2021.

28. Jackson CH. flexsurv: a platform for parametric survival modeling in R. J Stat Softw. 2016;70:8.

29. Centers for Disease Control and Prevention, National Center for Health Statistics. Life tables 2018. https://www.cdc.gov/nchs/products/life_tables.htm. Accessed 15 Nov 2021.

30. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Control. 1974;19(6):716–23.

31. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6(2):461–4.

32. Maurer MJ, Ghesquieres H, Jais JP, et al. Event-free survival at 24 months is a robust end point for disease-related outcome in diffuse large B-cell lymphoma treated with immunochemotherapy. J Clin Oncol. 2014;32(10):1066–73.

33. Jacobson CA, Locke FL, Ghobadi A, et al. Long-term (4- and 5-year) overall survival in ZUMA-1, the pivotal study of axicabtagene ciloleucel in patients with refractory large B-cell lymphoma [poster 1746]. In: Presented at the 63rd ASH Annual Meeting, Atlanta, GA, 11–14 Dec 2021.

34. Vadgama S, Mann J, Bashir Z, Spooner C, Collins GP, Bullement A. Predicting survival for chimeric antigen receptor T-cell therapy: a validation of survival models using follow-up data from ZUMA-1. Value Health. 2022;25(6):1010–7.

35. Ouwens M, 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):1129–38.

36. Bullement A, Latimer NR, Bell GH. Survival extrapolation in cancer immunotherapy: a validation-based case study. Value Health. 2019;22(3):276–83.

Authors and Affiliations

Elisabeth F. P. Peterse[1] · Elisabeth J. M. Verburg-Baltussen[1] · Alexa Stewart[2] · Fei Fei Liu[3] · Christopher Parker[4] · Maarten Treur[1] · Bill Malcolm[4] · Sven L. Klijn[3]

✉ Sven L. Klijn
  sven.klijn@bms.com

[1] Present Address: OPEN Health, Rotterdam, The Netherlands

[2] Present Address: OPEN Health, Oxford, UK

[3] Bristol Myers Squibb, Princeton, NJ, USA

[4] Present Address: Bristol Myers Squibb, Uxbridge, UK