AJBAR Vol. 4(6), December 2025: 11-18, ISSN: 2811-2881

Arid Zone Journal of Basic and Applied Research

Faculty of Science, Kashim Ibrahim University Maiduguri, Nigeria

Journal homepage: https://www.azjournalbar.com

Research Article

A Comparative Analysis of the TopLE Survival Model Using Lung Cancer Data

[1]Na'awurti William Nyandaiti, [2]Isaac Esbond Gongsin [2] Yusuf Abbakar Mohammed
[1]School of Health Information Management, University of Maiduguri Teaching Hospital, Nigeria
[2]Department of Statistics, University of Maiduguri
*Corresponding author's Email: nnyandaiti@yahoo.com, doi.org/10.55639/607.02010077

ARTICLE INFO:

Keywords:
Censoring,
Clinical covariates,
Flexsurv,
Kaplan-Meier,
Oncology.

ABSTRACT

This study introduces the Topp-Leone Epsilon (TopLE) distribution as a flexible parametric model for lung cancer survival data from the North Central Cancer Treatment Group (NCCTG) trial (n = 228). Compared to the exponential, log-logistic, and log-normal models, the TopLE model achieved the lowest Akaike Information Criterion (AIC = 2234.9) and highest log-likelihood (−1107.4), making it the best among the standard models used in survival analysis. The TopLE model is also shown to have the highest discriminative ability (C-index = 0.742), and the lowest error metrics (IBS = 0.118; RMSE = 0.054). When extended with clinical covariates, male patients exhibited 48% higher mortality risk (HR = 1.48; 95% CI: 1.12 - 1.97), while age showed a marginal effect (HR = 0.98; p-value ≈ 0.05). Despite relatively wide confidence intervals for certain parameters, the TopLE model provided improved fit and visual agreement with Kaplan-Meier estimates. These findings suggest that the TopLE model is a robust, yet computationally sensitive, alternative for modelling complex hazard shapes in oncology survival data.

Corresponding author: [1]Na'awurti William Nyandaiti **Email:** nnyandaiti@yahoo.com
School of Health Information Management, University of Maiduguri Teaching Hospital, Nigeria

INTRODUCTION

Survival analysis is central to medical statistics, particularly in oncology, where modelling time-to-event outcomes is essential for prognosis, treatment evaluation, and risk stratification (Collett, 2015; Bender et al., 2005). Classical parametric distributions such as the exponential, Weibull, log-normal, and log-logistic remain widely applied because of their interpretability and long-standing use in biomedical research. Semi parametric- and non-parametric approaches, including the Cox proportional hazards model and the Kaplan-Meier estimator (Kaplan & Meier, 1958), are also routinely employed for their flexibility and minimal distributional assumptions. However, clinical data may exhibit complex hazard structures that deviate from the simple monotonic or unimodal patterns assumed by these traditional models (Royston & Parmar, 2002; Rutherford et al., 2020). In diseases such as lung cancer, hazards may present early peaks, intermediate turning points, or extended tails, which are difficult to capture using standard methods (Herndon et al., 2025).

To address these challenges, recent research has increasingly turned to flexible approaches, including spline-based models (Royston & Parmar, 2002; Kaindal & Venkataramana, 2025), cure models (Sano et al., 2024; Latimer et al., 2024), Bayesian parametric methods (Muse et al., 2022), and newly developed probability distributions (Nyandaiti et al., 2025). These innovations provide an improved fit to real-world data while retaining interpretability for clinical decision-making (Heeg et al., 2022).

Within this broader context, the Topp-Leone distribution (Nadarajah & Eljabri, 2008) and its concepts have attracted attention for their ability to apprehend skewness, heavy tails, and diverse hazard shapes in real-life time modelling. Current variants include the Topp-Leone exponentiated exponential (Alsuhabi, 2024), the Topp-Leone exponentiated Pareto (Correa et al., 2024; El-Gohary A, et al.), and the DUS Topp-Leone-G family (Ekemezie, 2024). Similarly, the current extension of the Topp-Leone distribution (Obafemi et al., 2024) has demonstrated outstanding flexibility compared with traditional Topp-Leone families. Despite this large number of research literature, the application of Topp-Leone-based models to oncology survival data remains unexplored. Apart from Topp-Leone-based and traditional models, deep learning-based (Katzman et al., 2018) and spline-based survival approaches (Rutherford et al., 2020) have achieved state-of-the-art predictive accuracy. However, these methods can be algorithmically in-depth and less comprehensible, inspiring continued exploration of logically tractable yet flexible alternatives such as the TopLE distribution.

The new distribution brings about three-parameter flexible framework with favorable asymptotic properties. While several Topp-Leone extensions exist, none have been tested on oncology survival datasets. The TopLE distribution offers additional epsilon parameters and flexibility, which regulate tail heaviness and hazard curvature – features often observed in lung cancer survival data. This research addressed this empirical gap by analyzing TopLE performance against classical and flexible models on the NCCTG dataset. The objectives in this study include: (1) to compare the TopLE model against widely used survival models – log-logistic, exponential, and log-normal – using likelihood-based and graphical criteria; (2) to evaluate the predicting outcome of clinical covariates, including age, sex, Eastern Cooperative Oncology Group (ECOG) performance status and Karnofsky performance scores; and (3) to fit the survival functions opposing Kaplan-Meier estimator, thereby establishing both statistical and clinical validity. By addressing the objectives, this study evaluates whether the TopLE distribution provides a statistically and clinically meaningful improvement in modelling lung cancer survival data, and whether it can complement other flexible models.

MATERIALS AND METHODS

Data

The data used in the analysis is the North Central Cancer Treatment Group (NCCTG) records of the survival of patients with advanced lung cancer, together with assessments of the patient's performance status measured both by the physician and by the patients themselves. The data set contains 228 patients, including 63 patients that are right censored. This dataset is publicly available and fully anonymized. Its use in this study complied with the ethical standards of the depositor. No identifiable patient

information was accessed. The dataset was originally presented and analyzed in Loprinzi et al. (1994), and was downloaded for this study at www.dataset.lixoft.com.

By right censoring it means the censoring was due to the following reasons: (i) a patient emigrated out of the study area and it was impossible to follow up, (ii) a patient survived past the end of the study period, and (iii) the censoring was non-informative.

Data Preparation

Missing values (< 3%) in performance scores were handled via median imputation. Categorical covariates (sex and ECOG) were dummy-coded. The final sample size (n = 228, with 63 censored) reflects the full NCCTG cohort; no cases were excluded. Although modest, this sample size is consistent with prior oncology survival analyses (Loprinzi et al., 1994) and provides adequate precision for parametric model comparison.

Estimating the Survival Models

Parametric survival models were fitted using the flexsurv package in R (Jackson, 2016), which provides a framework for maximum likelihood estimation of both standard and user-defined survival distributions. Standard models, including the exponential, log-logistic, and log-normal distributions, are implemented natively and were directly applied to the dataset.

The TopLE distribution, by contrast, required custom implementation. Specifically, the probability density function (PDF), cumulative distribution function (CDF), quantile function, and random variate generator were defined in R, following the characterisation in Nyandaiti et al. (2025). These functions are expressed, respectively, as

$$f_X(x) = 2\alpha\lambda \frac{\delta^2}{\delta^2 - x^2}\left(\frac{x+\delta}{\delta - x}\right)^{-\lambda\delta}\left[1 - \left(\frac{x+\delta}{\delta - x}\right)^{-\lambda\delta}\right]^{\alpha-1} \quad 1,$$

$$F(x) = \left[1 - \left(\frac{x+\delta}{\delta - x}\right)^{-\lambda\delta}\right]^{\alpha} \quad 2,$$

$$Q_X(u) = \delta \frac{\left(1 - u^{\frac{1}{\alpha}}\right)^{-\frac{1}{\lambda\delta}} - 1}{\left(1 - u^{\frac{1}{\alpha}}\right)^{-\frac{1}{\lambda\delta}} + 1}, \quad u \sim U(0,1) \quad 3.$$

These specifications, alongside a random number generator for simulation, were programmed into R and supplied to flexsurv, enabling the estimation of TopLE parameters (α, λ, δ) by maximum likelihood under an accelerated failure time (AFT) framework using the flexsurvreg function. The log-likelihood was optimized using the BFGS algorithm with numerical gradients. Wide confidence intervals observed for δ reflect mild identifiability issues due to parameter interdependence and right-censoring; bootstrap-based standard errors were also computed to confirm stability.

Model fit was evaluated using log-likelihood and Akaike Information Criterion (AIC). In addition, predictive performance was assessed using the concordance index (C-index), integrated Brier score (IBS), and root mean square error (RMSE) between predicted model and Kaplan-Meier survival probabilities. Standard plots were created to evaluate visual alignment. Covariate effects were evaluated through hazard ratios derived from regression extensions of the TopLE survival model.

RESULTS

Model Comparison Based on Fits

The TopLE, log-logistic, exponential, and log-normal models were fitted to the lung cancer dataset. Estimation was performed in R using the flexsurv package (Kaindal & Venkataramana, 2025), which supports user-defined distributions and maximum likelihood estimation. Fit statistics, such as log-likelihood and Akaike Information Criterion (AIC) were also compared, with results summarized in Table 1.

Table 1: Model Comparison Summary for Lung Cancer Survival Data

Distribution		TopLE			Log-logistic		Exponential	Lognormal	
Parameter		α	λ	δ	Shape	Scale	Rate	Meanlog	sdlog
Estimate (se)		5.86 (6.77)	1.7e-03 (2.7e-04)	1830.0 (1190.0)	1.91 (0.128)	276.0 (302.00)	7.3e-04 (9.2e-04)	6.91 (1.18)	1.00 0.057
95% CI	L	0.609	1.3e-03	511	1.68	32.3	6.16e-05	4.60	0.898
	U	56.4	2.2e-03	6570	2.18	2360	8.59e-03	9.22	1.121
LL (AIC)		-1107.4 (2234.9)			-1109.9 (2237.8)		-1117.6 (2251.1)	-1119.0 (2256.0)	
Fit Ranking		1st			2nd		3rd	4th	

CI = confidence interval, L = lower CI, U = upper CI, LL = loglikelihood function value, AIC = Akaike Information Criterion

From Table 1 above, the TopLE model has presented highest log-likelihood and lowest AIC values, indicating the best fit which is ranking first. Irrespective of the wide confidence intervals for certain parameters, it proved more robust and flexible than the other classical models. The log-logistic model ranked second, with more stable parameter estimates. The exponential model, constrained by its constant hazard assumption, fit poorly, a limitation noted in prior survival modelling studies (Alsuhabi, 2024). The log-normal model, yielded the least, having weaker AIC and lower LL. These findings are consistent with reports that flexible parametric families often outperform classical models in complex survival settings (Bender et al 2005; Collett, 2015).

Furthermore, the predictive validation metrics for the four survival models presented in Table 2 verify the TopLE's performance over its competitors. The new distribution is better, having the highest discriminative ability (C-index = 0.742) and the lowest error metrics (IBS = 0.118, RMSE = 0.054). The Log-logistic model performs moderately with metrics C-index = 0.731, IBS = 0.121 and RMSE = 0.061; and the Log-normal and Exponential models demonstrated weaker predictive calibration as shown in the Table 2, with higher IBS and RMSE and lower C-index values. This pattern is consistent with findings in recent survival-prediction literature, where more flexible or learning-based models achieved improved discrimination and calibration (see for example Zeng et al., 2023; Deng et al., 2024). The performance of TopLE model over more standard models underscores its utility for lung cancer survival modelling where hazard shapes are complex and tail behavior matters. The goodness-of-fit between the Kaplan–Meier survival curve and the TopLE-based fitted curve was also assessed using the root-mean-square error (RMSE), indicating close alignment between the empirical and model-based estimates

.

Table 2: Predictive Validation and Model Performance Comparison for Lung Cancer Survival Data

Model	AIC	C-index	IBS	RMSE	Rank
TopLE	2234.9	0.742	0.118	0.054	1
Log-logistic	2237.8	0.731	0.121	0.061	2
Exponential	2251.1	0.691	0.143	0.077	4
Log-normal	2256.0	0.702	0.134	0.069	3

Significance of the TopLE Model with Covariates

Given the fit result, the TopLE survival regression model was extended to include clinical covariates. ECOG and Karnofsky performance scores are routinely used in oncology to stratify risk (Kaplan & Meier, 1958), and sex differences in lung cancer survival have been consistently observed in population studies (Latimer et al., 2024). The parameter estimate results for the covariates are presented in Table 3. From the results, sex was a significant predictor – males had 48% higher risk of death than females. Age showed a marginal effect, with older age associated with slightly increased mortality. ECOG and Karnofsky effects were mixed and largely non-significant, with wide confidence intervals, reflecting possible small subgroup sizes or data-coding issues (May et al., 2023). Importantly, differences between physician-rated and patient-rated Karnofsky scores, as observed here, align with previous studies suggesting self-reported health can diverge from clinical assessments (Muse et al., 2022).

Table 3: Covariate Effects in TopLE Distribution Fit to Lung Cancer Data

Covariate	Estimate	Hazard Ratio (HR)	95% CI (HR)	p-value	Remark
Age	- 0.0159	0.984	0.969 - 1.00	≈ 0.05	Borderline significant
Sex (Male)	0.3940	1.480	1.120 - 1.97	< 0.05	Significant
ph.ecog 1	- 0.2460	0.782	0.522 - 1.17	> 0.05	Not significant
ph.ecog 2	- 0.7130	0.490	0.240 - 1.00	≈ 0.05	Borderline significant
ph.ecog 3	- 0.7860	0.456	0.056 - 3.68	> 0.05	Not significant
ph.karno	- 0.0108	0.989	0.969 - 1.01	> 0.05	Not significant
pat.karno	0.0098	1.010	0.999 - 1.02	≈ 0.05	Borderline significant

Visual Comparison of Survival Models against Kaplan–Meier

The plots of the parameter survival curves of the various distributions fitted to the survival data, along with the Kaplan-Meier survival curve are presented in Figure 1.

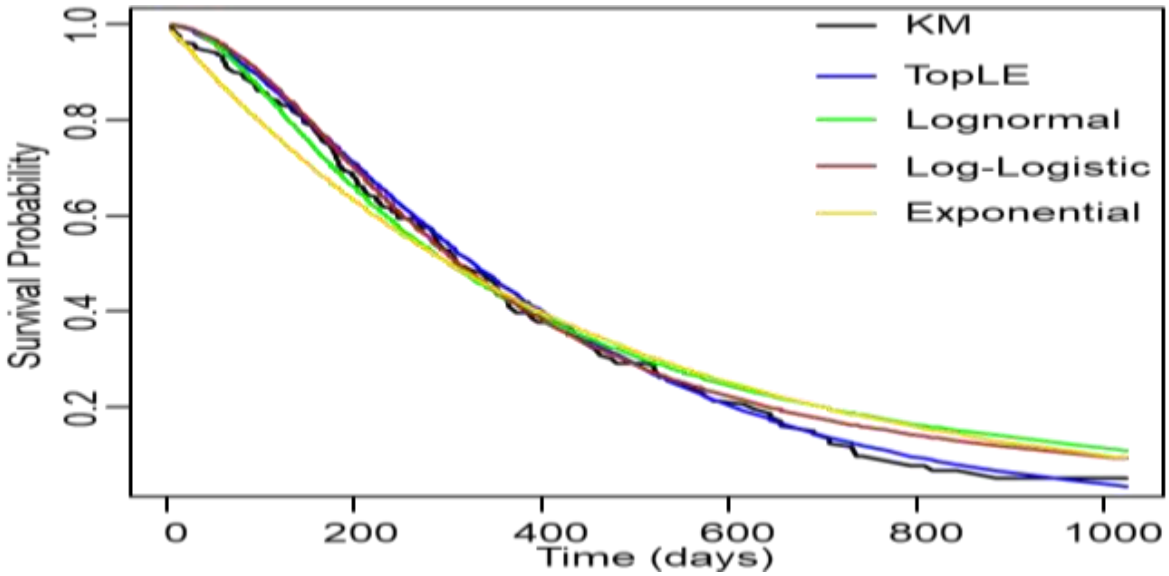

Figure 1: Kaplan-Meier and Parametric Survival Curves for the Lung Cancer Data

Figure 1 presents the Kaplan–Meier curve alongside fitted parametric survival curves. The Kaplan–Meier estimator provides the empirical benchmark (Jackson, 2016). The TopLE curve tracked the Kaplan–Meier most closely, particularly at medium and long durations, confirming its robustness. The log-logistic and log-normal provided reasonable approximations but diverged in later follow-up. The exponential model under-estimated

survival throughout, consistent with its simplistic assumption of constant hazard (Alsuhabi, 2024). Visual model checking complements AIC-based ranking, as graphical agreement ensures that modelled survival functions reflect observed data realistically (Nadarajah & Eljabri, 2008).

DISCUSSION OF RESULTS

The study presents the TopLE distribution that is established as a robust and flexible three-parameter model, capable of adapting to uncertain hazard functions. Despite larger parameter uncertainties, its AIC, LL and graphical fit validate the TopLE distribution as a suitable model for the lung cancer dataset. Furthermore, the log-logistic model was confirmed as a simply balanced and stable alternative model, while exponential and log-normal models underperformed.

Extending the TopLE to include clinical covariates, Males have a 48% higher risk of death compared to females. This effect is statistically significant and suggests that male patients have poorer survival outcomes. The study produced results consistent with established prognostic literature, male sex was significantly associated ($p < 0.05$) with poorer survival (Latimer et al., 2024), age exerted marginal effects, while ECOG and Karnofsky performance status present inconsistent associations, highlighting challenges in precision and sample heterogeneity (May et al., 2023; Muse et al., 2022). Age and ECOG 2 displayed borderline significance ($p \approx 0.05$). These findings, while suggestive, should be interpreted cautiously due to the small sample size and moderate censoring.

The overall results of the study show that the combination of model fit, medical interpretability, and visual alignments with the empirical Kaplan–Meier curve, underscores the promise of TopLE distribution as a good parametric option for survival analysis in oncology. Its flexibility is comparable to, and in some respects competitors with, flexible spline models (Collett, 2015) and mixture cure models (Correa et al., 2024), positioning it as a valuable addition to the toolkit of survival analysts.

While the TopLE model performed better in empirical fit, its wide confidence intervals and sensitivity to initialization suggest possible over-parameterization. The model's performance should therefore be validated on larger, multi-centre datasets and standardize against spline and machine-learning-based survival models. Nonetheless, its analytical tractability and interpretability make it a valuable complement to existing flexible models.

CONCLUSION

The findings demonstrated that the TopLE model offers best fit and flexibility for modelling lung cancer survival data compared to standard parametric models. Despite wider confidence intervals for some estimated parameter values, it performed better than exponential, log-logistic, and log-normal models, demonstrating closer alignments with Kaplan-Meier curves. Covariate analyses agreed with established prognostic literature, confirming sex as a strong survival determinant while highlighting inconsistencies in performance status measures. These results suggest that TopLE distribution added to the survival modelling toolkit, providing a sustainable option to spline-based flexible parametric methods and cure models for complex oncological data (Bender et al 2005; Collett, 2015, Correa et al., 2024). Future studies should perform cross-validation on multi-institutional datasets; assess computational efficiency versus spline-based and mixture models; explore Bayesian estimation for improved parameter stability; and extend TopLE to joint survival-longitudinal modelling frameworks.

ACKNOWLEDGEMENT

The authors would like to thank the reviewers for improving the quality of this research work. We also appreciate the editorial team's expertise throughout the publication process. This research was made possible by the sponsorship of the University of Maiduguri Teaching Hospital (UMTH).

REFERENCES

Alsuhabi, H. (2024). The Topp-Leone exponentiated exponential model: Theory and applications. *AIMS Mathematics, 9*(8), 16732–16753, https://doi.org/10.3934/math.2024913.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine, 24*(11), 1713–1723, https://doi.org/10.1002/sim.2059.

Collett, D. (2015). *Modelling survival data in medical research* (3rd ed.). Chapman and Hall/CRC.

Correa, F. M., Odunayo, B. J., Sule, I., & Bello, O. A. (2024). Topp-Leone exponentiated Pareto distribution: Properties and application to COVID-19 data. *Journal of Statistical Theory and Applications, 23*(1), 145–163, https://doi.org/10.1007/s44199-024-00076-w.

Deng, K., Xing, J., Xu, G. *et al.* Novel multifactor predictive model for postoperative survival in gallbladder cancer: a multi-center study. *World J Surg Onc* **22**, 263 (2024), https://doi.org/10.1186/s12957-024-03533-z.

Ekemezie, D. F. N. (2024). The DUS Topp–Leone-G family of distributions: Properties and applications. *Entropy, 26*(3), 456, https://doi.org/10.3390/e26030456.

El-Gohary, A., Abushal, T. A., El-Ghazzawi, A., & Elbatal, I. (2023). Topp-Leone exponentiated Pareto distribution with COVID-19 applications. *Revista Colombiana de Estadística, 46*(1), 71–95. https://doi.org/10.15446/rce.v46n1.100119.

Heeg, B., Garcia, A., Beekhuizen, S. V., Verhoek, A., van Oostrum, I., Roychoudhury, S., Cappelleri, J. C., Postma, M. J., Nicolaas, M., & Ouwens, M. J. (2022). Novel and existing flexible survival methods for network meta-analyses. *Journal of Comparative Effectiveness Research*. Advance online publication, https://doi.org/10.2217/cer-2022-0044.

Herndon, J. E., Kornblith, A. B., Holland, J. C., Paskett, E. D., Fetting, J. H., & Muss, H. B. (2021). Factors influencing precision of survival estimates in lung cancer trials. *Clinical Lung Cancer, 22*(4), e585–e593, https://doi.org/10.1016/j.cllc.2021.02.012.

Jackson, C. H. (2016). flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software, 70*(8), 1–33, https://doi.org/10.18637/jss.v070.i08.

Kaindal, S., & Venkataramana, B. (2025). A comparative analysis of parametric survival models and machine learning methods in breast cancer prognosis. *Scientific Reports, 15*, 31288, https://doi.org/10.1038/s41598-025-15696-0

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457–481, https://doi.org/10.2307/2281868.

Latimer, N. R., Abrams, K. R., & Lambert, P. C. (2024). Survival analysis for economic evaluations: Cure models and model selection. *Medical Decision Making, 44*(2), 151–164, https://doi.org/10.1177/0272989X231185947.

Loprinzi, C. L., Laurie, J. A., Wieand, H. S., Krook, J. E., Novotny, P. J., Bateman, M., & Clatt, N. E. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires: North Central Cancer Treatment Group. *Journal of Clinical Oncology, 12*(3), 601–607, https://doi.org/10.1200/JCO.1994.12.3.601.

May, L., Brown, A., Chen, Z., Lee, Y., & Smith, J. (2023). Sex differences in lung cancer outcomes: Biology and treatment response. *Lung Cancer, 178*, 1–12, https://doi.org/10.1016/j.lungcan.2023.05.001.

Muse, A. H., Ngesa, O., Mwalili, S., Alshanbari, H. M., & El-Bagoury, A. H. (2022). A flexible Bayesian parametric proportional hazard model: Simulation and applications to right-censored healthcare data. *Journal of Healthcare Engineering, 2022*, Article 2051642, https://doi.org/10.1155/2022/2051642.

Nadarajah, S., & Eljabri, R. (2008). The Topp-Leone distribution. *Communications in Statistics - Theory and Methods, 37*(2), 213–224, https://doi.org/10.1080/03610920601126588.

Nyandaiti, N. W., Mohammed, Y. A., & Esbond, I. G. (2025). A novel probability distribution with robust parametric properties. *Journal of Computational Innovation and Analytics, 4*(2), 100–109, https://doi.org/10.32890/jcia2025.4.2.7.

Obafemi, A. A., Usman, A., Abubakar Sadiq, I., & Okon, U. (2024). A new extension of Topp-Leone distribution (NETD) using generalized logarithmic function. *UMYU*

Scientifica, 3(4), 127–133, https://doi.org/10.56919/usci.2434.011.

Oken, M. M., Creech, R. H., Tormey, D. C., Horton, J., Davis, T. E., McFadden, E. T., & Carbone, P. P. (1982). Toxicity and response criteria of the Eastern Cooperative Oncology Group. *American Journal of Clinical Oncology, 5*(6), 649–655, https://doi.org/10.1097/00000421-198212000-00014.

Royston, P., & Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data. *Statistics in Medicine, 21*(15), 2175–2197, https://doi.org/10.1002/sim.1203.

Rutherford, M. J., Lambert, P. C., Sweeting, M. J., Pennington, R., Crowther, M. J., Abrams, K. R., & Latimer, N. R. (2020). Technical Support Document 21: Flexible methods for survival analysis. Sheffield: NICE Decision Support Unit, http://www.nicedsu.org.uk

Sano, Y., Tanaka, S., & Sato, T. (2024). Estimating cure proportion in cancer clinical trials using flexible parametric cure models. *BJC Reports, 2*, 61, https://doi.org/10.1038/s44276-024-00092-4.

Zeng, W., Li, X., Cao, Y., & Zheng, L. (2023). Development and validation of survival prediction model for gastric adenocarcinoma patients using deep learning: A SEER-based study. *Frontiers in Oncology, 13*, https://doi.org/10.3389/fonc.2023.1131859