Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Advanced Survival Model Selection Techniques for Health Economic Evaluations.

Fidelsia Fri Esibe

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Data Science

SUPERVISOR :
Prof. dr. Inigo BERMEJO DELGADO

2024
2025

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Advanced Survival Model Selection Techniques for Health Economic Evaluations.

Fidelsia Fri Esibe

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Data Science

SUPERVISOR :

Prof. dr. Inigo BERMEJO DELGADO

Abstract

Accurate long-term survival extrapolation is essential for health economic evaluations, particularly in oncology, where treatment benefits may extend well beyond clinical trial follow-up periods. This study examines two methodological approaches to enhance survival model selection: k-fold cross-validation and inverse probability of censoring weighting (IPCW).

In this thesis, we investigated whether k-fold cross-validation could improve model selection and extrapolation accuracy in survival analysis, paying particular attention to datasets showing long-term survival plateaus. Traditional model selection based on information criteria like AIC and BIC evaluates models using the entire dataset, which can lead to overfitting and may not optimize extrapolation performance when projecting beyond observed data. In contrast, cross-validation assesses model fit on held-out data folds, providing more generalizability. Additionally, we examined whether inverse probability of censoring weighting could improve extrapolation accuracy in highly censored datasets, where we applied IPCW across simulated samples with censoring levels ranging from around 40% to 80%.

We conducted simulation studies using eight diverse cancer datasets from clinical trials and population registries to evaluate the performance of cross-validation compared to conventional approaches across various parametric and spline-based models. We conducted simulation studies using eight diverse cancer datasets from clinical trials and population registries to evaluate the performance of cross-validation compared to conventional approaches across various parametric and spline-based models.

Cross-validation provided modest improvements in extrapolation accuracy (measured using restricted mean survival time) in approximately half of the datasets examined, though benefits varied considerably by context. Despite clinical evidence supporting survival plateaus, flexible spline models were consistently selected over mixture cure models across all datasets. IPCW improved prediction accuracy in about 91% of comparisons, with peak benefits observed at approximately 60% censoring.

These findings suggest that while cross-validation offers selective advantages for model selection, inverse probability of censoring weighting consistently improves prediction accuracy under high censoring conditions. The results have practical implications for enhancing survival analysis in health technology assessments, particularly when dealing with immature data or populations where a cure is clinically plausible.

Contents

1 Introduction

Survival analysis makes an extremely important contribution to health economic assessments, particularly in the assessment of the economic value of new healthcare interventions. In most clinical trials, due to time and resource constraints, the follow-up of patients is limited. This results in incomplete knowledge of the long-term outcomes of the treatment, which are important for economic evaluation of a drug (N. R. Latimer 2013). Health economic models tend to require extrapolation from observed trial data to predict lifetime costs and impacts of interventions (NICE Decision Support Unit 2013). It is therefore important that the extrapolations are accurate, since they directly affect healthcare resource allocation decisions. Therefore, the selection of appropriate survival models is a critical methodological concern in health technology assessment (Jackson et al. 2017).

The challenge of selecting the most appropriate survival model for extrapolation has long been recognized in health economics literature. The traditional approaches to select models have revolved around goodness-of-fit criteria such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). The above measures balance fit by penalizing additional parameters, thereby attempting to select models that generalize rather than overfit the data available. While these metrics assess fit on observed data, they do not necessarily identify the model that best predicts survival in the unobserved tail (NICE Decision Support Unit 2013).

A key issue with these conventional methods is that model fit is assessed using the full observed dataset, which can give rise to models that fit the data too closely and then perform very poorly when extrapolating beyond the observed time horizon. When selecting a model on the basis of how well it fits the complete dataset, it will fit noise or patterns that exist only in the dataset itself and not the underlying survival pattern itself (Harrell 2015). This is particularly problematic in health economic evaluations, where extrapolations can extend decades beyond the available trial data (N. R. Latimer 2013).

Machine learning techniques in recent years have strengthened model selection approaches in a broad range of fields, with cross-validation being a robust method to estimate predictive performance (Hastie, Tibshirani, and Friedman 2009). K-fold cross-validation is the dividing data into k folds and systematically using k-1 folds for model training and reserving one fold for testing, hence enabling assessment of model performance on new unseen data (Arlot and Celisse 2010). Despite it's use and application in predictive modeling, cross-validation is still partially underutilized in the selection of survival models for health economic evaluation.

There are several advantages of Cross-validation over traditional model selection methods. Firstly, it clearly evaluates out-of-sample predictive accuracy, reflecting the real-world scenarios where models must extrapolate beyond the available data (Arlot and Celisse 2010). Second, it provides a more stable model generalizability estimate by taking a mean performance for a number of train-test splits in order to reduce the effects of idiosyncrasies present in a single partition of data (Molinaro, Simon, and Pfeiffer 2005). Third, cross-validation is less susceptible to sample size effects compared to information criteria such as AIC and BIC, which could be particularly important where smaller clinical datasets are frequent, as is often the case in health economic evaluations (Browne 2000). In the survival modeling scenario, cross-validation can be used to identify models that generalize well to new data, rather than those that best fit the observed data.

Despite these established benefits in predictive modeling, cross-validation remains underutilized in the context of survival model selection for health economic evaluation, where standard information criteria remain dominant in practice (Gallacher, Kimani, and Stallard 2021). This methodological deficit represents an opportunity to potentially increase the validity of survival extrapolations that inform economic models and healthcare resource allocation decisions.

A previous study has established that k-fold cross-validation can improve the selection of traditional parametric and flexible survival models for extrapolation in health economic evaluations (Bermejo and Grimm 2024). The study using seven datasets showed that the models selected using cross-validation had significantly lower errors in restricted mean survival time (RMST) compared to the models selected using classical AIC or BIC methods across the entire datasets. A finding worth noting was that cross-validation tended to favor simpler models with better generalizability and avoided the overfitting that occurred from more complex models.

However, a limitation identified in this previous work was the inability to adequately deal with datasets with prolonged plateaus in their survival curves. These plateaus are increasingly common in survival data, particularly in new modern treatments like immunotherapy and some advanced oncology treatments. These developments in oncology treatment commonly yield a subgroup of patients with prolonged survival or a ”cured” fraction, generating survival curve plateaus that challenge classical parametric assumptions (Grant et al., 2019). There has been an increasing focus on cure models, which account for a cured subpopulation. Mixture cure models (MCMs), for instance, separate the population into cured and uncured components, yielding a more theoretical model for such data.

This thesis extends the previous methodology to address the limitation by incorporating cure models into the analysis. Specifically, we investigate whether k-fold cross-validation leads to better extrapolation performance than the usual model selection methods for datasets with long-term survival plateaus. The investigation ranges from standard parametric survival models to cure models that directly model the presence of a cure fraction (Lambert et al. 2007).

2 Research Questions

This study examines whether the benefits of cross-validation observed for selecting standard survival models for health economic evaluations extend to the specialized context of cure models.

Primary Research Questions:

- How does k-fold cross-validation improve model selection and extrapolation accuracy in mixture cure models, particularly in datasets exhibiting long-term survival plateaus?

In addition to this primary research question, the thesis explores another secondary research question aimed at improving the accuracy of extrapolation in survival data.

Secondary Research Question:

- How does inverse probability of censoring weighting (IPCW) affect extrapolation accuracy in datasets with high censoring rates?

3 Description of the Datasets

To investigate the performance of cure models across diverse clinical contexts and keeping in mind that cure models rely on a structural assumption of a cured population fraction, this study analyzes eight survival datasets obtained from a mix of clinical trial sources and observational cancer registries. We selected these datasets based on the presence of long-term survival plateaus, their relevance to oncology (particularly immunotherapy), and sufficient follow-up duration. Furthermore, medical plausibility, whether a durable response or cure is biologically reasonable for the condition and treatment, was a key consideration in selecting datasets appropriate for cure modeling.

A particular focus was placed on datasets reflecting immunotherapy outcomes, as these treatments can achieve a prolonged response in some patients, which suggests the presence of a cured fraction within the population (Patel et al. 2016). However, due to the lack of publicly available IPD (individual patient-level data) from immunotherapy trials, due to proprietary and regulatory constraints, we used digitization techniques to reconstruct IPD from published Kaplan-Meier (KM) survival curves.

Digitization involves obtaining survival times and event indicators from published Kaplan-Meier plots through software tools such as *DigitizeIt*, *WebPlotDigitizer*, or the *IPDfromKM* algorithm (Guyot et al. 2012). This approach provides an estimate of individual patient data (IPD) in cases where the original datasets are not accessible. In this study, five datasets were reconstructed using this method based on their reporting in key immunotherapy trials.

The remaining three datasets were selected from publicly available sources, including the SEER registry (National Cancer Institute 2023) and curated clinical trial repositories, and cover a variety of cancer types and censoring levels. The Kaplan-Meier curves presented in Figure 1 demonstrate the survival patterns in the eight cancer datasets.

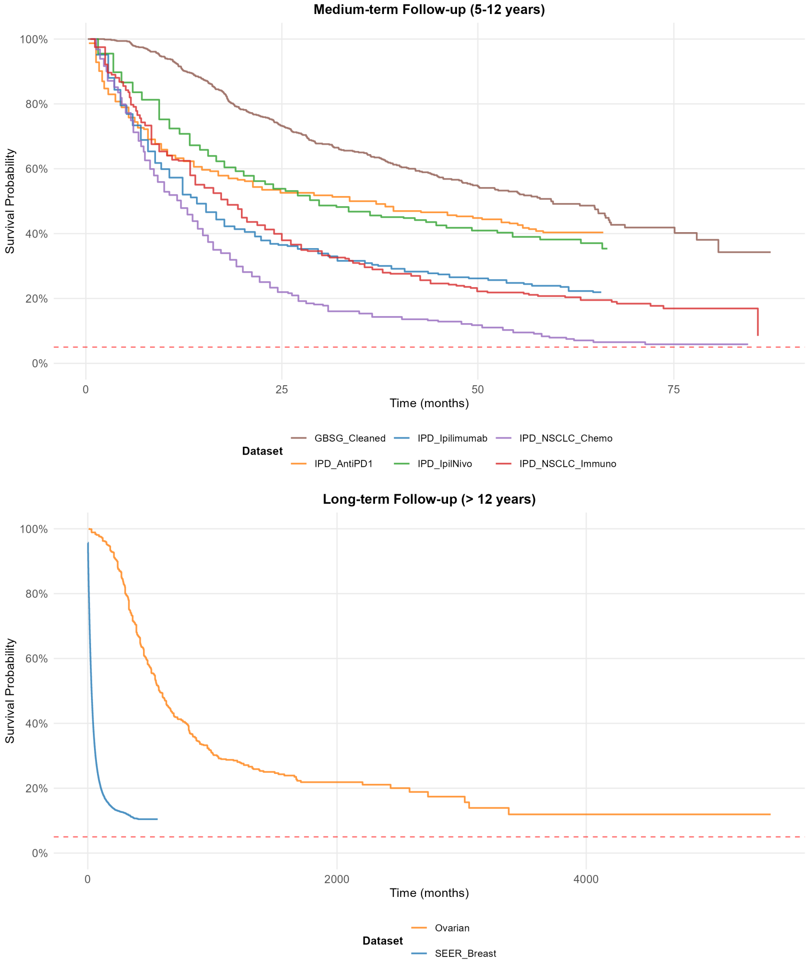

Figure 1: Kaplan-Meier survival curves for the eight cancer datasets grouped by follow-up duration. The top plot displays medium-term follow-up datasets (follow-up $<$ 5 years) which demonstrate varying plateau patterns. The plot below shows long-term follow-up datasets (follow-up $>$ 5 years) exhibiting more pronounced plateau regions. These plateau regions visible across multiple datasets, provide the rationale for using them in the analysis.

The datasets used include:

1. **Ipilimumab Monotherapy Dataset (OS_Ipilimumab):** This dataset represents the overall survival data for patients treated with ipilimumab monotherapy, reconstructed from published Kaplan-Meier curves using established digitization techniques (Guyot et al. 2012). The data originates from the Dutch Melanoma Treatment Registry, a prospective nationwide cohort study that demonstrated plateau formation in survival curves, indicating potential for long-term survival and cure fractions in real-world clinical practice (van Not et al. 2024).
2. **Anti-PD-1 Therapy Dataset (OS_AntiPD1):** Data from an anti-PD-1 immunotherapy trial in melanoma, with a substantial survival plateau. The data from the Dutch Melanoma

Treatment Registry study showed that anti-PD-1 antibodies achieve durable responses with characteristic plateau formation in survival curves outside clinical trial settings (van Not et al. 2024).

3. **Combination Immunotherapy Dataset (OS_IpiNivo):** This dataset represents patients with advanced melanoma treated with a first-line combination of ipilimumab and nivolumab therapy. The data, also derived from the Dutch Melanoma Treatment Registry, demonstrated that combination checkpoint inhibition achieves superior long-term survival outcomes compared to monotherapy, with enhanced plateau formation suggesting higher cure rate potential in real-world clinical practice (van Not et al. 2024).

4. **NSCLC_without_chemo:** This dataset was digitized from the Kaplan-Meier overall survival (OS) curve from Figure 1A of (Peters et al. 2025) for cure model analysis. This figure presents long-term OS outcomes for patients with metastatic NSCLC (non-small cell lung cancer) and tumor PD-L1 expression less than 1%, treated with first-line nivolumab plus ipilimumab–based regimens, pooled from the CheckMate 227 and CheckMate 9LA trials. This is the immunotherapy arm of the dataset with a sample size of n=322. It demonstrated significant survival improvements with plateau-forming survival curves.

5. **NSCLC Chemotherapy Dataset (NSCLC_chemo):** This dataset represents the control group from the CheckMate 227 trial, containing survival data from advanced NSCLC (non-small cell lung cancer). These patients received up to four cycles of platinum-based chemotherapy without any immunotherapy (Peters et al. 2025). The sample size was n=315. While traditional chemotherapy rarely achieves a cure in advanced NSCLC, some patients experience unexpectedly prolonged survival, making mixture cure model analysis relevant for understanding treatment heterogeneity and identifying potential long-term survivors.

6. **SEER Breast Cancer Dataset (SEER_Breast_Cancer):** This is observational data from the Surveillance, Epidemiology, and End Results (SEER) registry filtered for breast cancer (National Cancer Institute 2023). The SEER database provides high-quality, long-term follow-up data essential for cure model analysis. With over 30,000 records, this dataset exhibits a long follow-up time (over 500 months), a plateau after 31 months, and a good tail representation.

7. **Ovarian Cancer Dataset:** This dataset has survival information from ovarian cancer patients. The data has a long follow-up period of approximately 5480 months. Ovarian cancer presents unique survival characteristics with the potential for long-term disease-free survival in a subset of patients (Edmunson et al. 1979).

8. **German Breast Cancer Study Group Dataset (GBSG):** This dataset is from a prospective clinical trial led by the German Breast Cancer Study Group, which focuses on ”recurrence-free survival” among patients with node-positive breast cancer. Due to its comprehensive data collection and extended follow-up, the GBSG dataset is widely regarded as a benchmark for evaluating survival analysis methods (Royston and Altman 2013).

Each dataset was pre-processed to remove zero-time records, and we ensured the event indicators were valid. We also verified the time-to-event variables for positive values and the survival status coded as binary indicators (0 = censored, 1 = event).

4 Methodology

4.1 Mixture Cure Model Framework

This section addresses the research question: How does k-fold cross-validation impact the selection and extrapolation accuracy of mixture cure models (MCM) in datasets characterized by long-term survival plateaus? The goal is to evaluate whether cross-validation can improve model selection over traditional criteria (AIC/BIC) on the entire dataset in the context of cure models, leading to better extrapolation performance as measured by RMST accuracy and cure fraction estimation.

In this study, we utilized mixture cure models due to the clinical characteristics of the datasets, all of which show evidence of potential cure fractions based on the presence of plateaus and supported by clinical literature. Data maturity is crucial when using MCM, so the datasets have adequate follow-up durations.

The selection of mixture cure models is justified by the clinical characteristics of the included datasets, all of which exhibit evidence of potential cure fractions based on plateau formation in survival curves and clinical literature (Peng and Dear 2000). However, the application of cure models requires careful consideration of data maturity, as recent evidence suggests that immature data can lead to substantial overestimation of cure fractions and unreliable extrapolation (Grant et al. 2019). In order to address this concern, we are using datasets with sufficient follow-up duration and using artificial censoring to evaluate performance under conditions of limited data maturity.

4.1.1 Theoretical Foundation

Mixture cure models were initially introduced by (Boag 1949) and later formalized by (Berkson and Gage 1952). These models are designed for survival data in which a portion of the population is considered "cured," that is, they are no longer at risk of the event of interest. In such cases, the survival curve typically shows a plateau. Under a mixture cure model, the population survival function can be expressed as:

$$S(t) = \pi + (1 - \pi)S_u(t) \tag{1}$$

Where π represents the cure fraction (proportion of cured individuals), and $S_u(t)$ denotes the survival function for uncured individuals (Peng and Dear 2000).

Parameter estimation in mixture cure models is typically performed using maximum likelihood estimation (MLE). This approach involves constructing a likelihood function that accounts for both the cured and uncured components of the population. For uncensored observations, the contribution to the likelihood includes the density of the survival distribution among uncured individuals. In contrast, for censored observations, it includes a combination of the survival probability for uncured individuals and the cure fraction (Patilea and Keilegom 2017).

The cure fraction parameter captures the long-term survivors who are assumed to no longer be at risk, while the distribution describes the time-to-event pattern for those who remain susceptible. By estimating both components jointly, the model can accommodate survival curves that exhibit plateaus, a feature commonly seen in cancer immunotherapy trials (Othus et al. 2012).

This study uses the flexsurvcure package in R to implement the estimation(Amdahl 2022). The flexibility of this package allows for fitting a range of parametric forms to the uncured population while simultaneously modeling the cure fraction.

4.1.2 Parametric Distributions

In this study, we used five different parametric distributions to model the survival of patients who are not cured. The choice of parametric distributions for the baseline hazard follows established guidance for survival extrapolation studies, which recommends fitting multiple standard parametric models to evaluate how results may vary under different modeling assumptions (N. R. Latimer 2013). Standard parametric models are predominantly used in regulatory submissions and health technology assessments, particularly for mixture cure model applications where parameter interpretation is important. (Grant et al. 2019).

The following distributions were chosen because they are commonly used in cancer survival analysis and can represent different shapes of survival curves. Choosing the right distribution is important because it affects how well we can predict long-term survival beyond the time observed in the clinical data.

- **Weibull Distribution:** This distribution is among the most commonly used in survival analysis. It is flexible enough to model increasing or decreasing risk over time, which makes it a good default option for cure models (NICE Decision Support Unit 2013).
- **Log-normal Distribution:** This distribution is useful when the risk of the event (e.g., death or relapse) first increases and then decreases. It is beneficial in cancer studies where treatment effects take time to appear (NICE Decision Support Unit 2013).
- **Log-logistic Distribution:** Like the log-normal, this distribution can handle survival curves that rise and fall. It also has the advantage of being easy to interpret in clinical terms and is often used in comparisons of extrapolation methods (Gray, Hernandez, and N. Latimer 2020).
- **Exponential Distribution:** This is the simplest model and assumes that the risk of the event stays constant over time. Although it is not very flexible, it serves as a baseline and can

be useful when data are limited or follow-up is short (NICE Decision Support Unit 2013).

- **Gompertz Distribution:** This model is often used in cancer research because it can represent risks that increase over time, such as those related to aging. It has been found useful in cure models as well. (NICE Decision Support Unit 2013).

4.1.3 Flexible Parametric Survival Models (Spline-Based)

In situations where conventional parametric distributions may fail to capture complex hazard dynamics, we used spline-based survival models which gives a more flexible representation of the baseline hazard. The models were fitted using the `flexsurvspline` function in R, which allows adjusting both the number of spline knots (typically $k = 0$–4) and a chosen *scale* that transforms the survival function $S(t)$. This single choice of `scale` determines how the spline models the data and influences both interpretability and flexibility.

Specifically, the scale parameter defines the transformation $g(S(t))$ to which the spline is applied (Jackson 2025). The three available scales are:

- **Hazard scale** (`scale = "hazard"`): Models the *log cumulative hazard*, i.e., $g(S(t)) = \log(H(t))$ where $H(t) = -\log(S(t))$. When $k = 0$, the model simplifies to a Weibull distribution (Jackson 2025).

- **Odds scale** (`scale = "odds"`): Models the *log cumulative odds of failure*, i.e., $g(S(t)) = \log\big(F(t)/(1 - F(t))\big)$, where $F(t) = 1 - S(t)$. With $k = 0$, it reduces to the log-logistic distribution(Jackson 2025).

- **Normal (Probit) scale** (`scale = "normal"`): Models the inverse-normal transformation of survival, i.e., $g(S(t)) = -\Phi^{-1}(S(t))$, where $\Phi^{-1}(\cdot)$ is the standard normal inverse cumulative distribution function. At $k = 0$, the model simplifies to a log-normal distribution(Jackson 2025).

In each case, using multiple knots ($k > 0$) allows the model to move from the base parametric forms, accommodating more complex time-dependent hazard or survival patterns, while maintaining interpretability and statistical rigor.

Knot Placement

In spline-based models, knots are the points where separate polynomial pieces of the spline are joined. Between knots, the hazard function is modeled as a smooth curve, and the placement of knots determines the model's ability to capture changes in the shape of the hazard or survival curve. Few knots (e.g., $k = 1$) produce smoother curves that resemble simpler parametric models, while more knots allow greater flexibility to capture complex or non-monotonic hazard patterns. (Royston and Parmar 2002).

In this analysis, we used 1 to 4 internal knots to evaluate whether increased model flexibility improves predictive performance. The knots were positioned at equally spaced quantiles of the log survival times, with boundary knots placed at the minimum and maximum observed log survival times, following standard recommendations (Royston and Parmar 2002). This systematic evaluation from simple to complex models allows for assessment of the trade-off between model complexity and prediction accuracy.

Model Space

For our analysis, 17 models were evaluated, consisting of 5 parametric cure models and 12 spline configurations. The spline models combined 4 knot choices (k = 1, 2, 3, 4) with three scale types (hazard, odds, normal), which results in the following configurations:

- **k=1:** Spline k=1 hazard, Spline k=1 odds, Spline k=1 normal
- **k=2:** Spline k=2 hazard, Spline k=2 odds, Spline k=2 normal
- **k=3:** Spline k=3 hazard, Spline k=3 odds, Spline k=3 normal
- **k=4:** Spline k=4 hazard, Spline k=4 odds, Spline k=4 normal

This comprehensive ”model space” allowed us to compare simpler, interpretable parametric forms against more flexible spline-based alternatives, and assess how flexibility (via knot count) and choice of scale impacted model performance across datasets.

4.1.4 Model Selection Criteria

The model selection criteria used in this analysis are:

Akaike Information Criterion (AIC): Defined as $\text{AIC} = -2\ell + 2p$, where ℓ represents the log-likelihood and p denotes the number of parameters. The Akaike Information Criterion (AIC) evaluates model quality by balancing fit against complexity, prioritizing predictive accuracy over identification of the “true” underlying model (Burnham and Anderson 2002).

Bayesian Information Criterion (BIC): Defined as $\text{BIC} = -2\ell + p\log(n)$, where n represents the sample size. BIC applies a stronger penalty for model complexity than AIC, particularly in larger samples, and is designed to identify the “correct” model when it exists among the candidates

(Burnham and Anderson 2002). The distinction between AIC and BIC reflects different modeling philosophies: BIC is more conservative and tends to select simpler models, while AIC prioritizes predictive accuracy and is more tolerant of model complexity (Aho, Derryberry, and Peterson 2014).

Cross-Validation Approach

Cross-validation provides an alternative model selection strategy that estimates out-of-sample predictive performance (Arlot and Celisse 2010). This study implements k-fold cross-validation with the following procedure:

1. **Data Partitioning:** The dataset is randomly divided into $k = 10$ equal-sized folds
2. **Model Training:** For each fold i, mixture cure models are fitted using data from the remaining $k - 1$ folds(i.e the training set)
3. **Validation:** Model performance is evaluated on the held-out fold i using out-of-sample log-likelihood
4. **Per-Fold Information Criteria Calculation:** For each model and fold, AIC and BIC are computed using the validation log-likelihood and the number of estimated parameters.
5. **Aggregation Across Folds:** The per-fold AIC and BIC values are averaged across all k folds to obtain cross-validated AIC and BIC scores for each model.
6. **Model Selection:** The model with the lowest average cross-validated AIC (or BIC) is selected as the better model for extrapolation.

The choice of k = 10 folds balances computational efficiency with reliable performance estimation, following recommendations for moderate sample sizes (Hastie, Tibshirani, and Friedman 2009).

4.1.5 Experiments

The following steps outline the simulation-based procedure used to evaluate the model performance across multiple datasets. It describes the code execution sequence, from preparing the data to fitting models and evaluating outcomes.

1. **Sample Size Standardization:** A random sample of 250 observations is drawn from each full dataset to ensure consistent statistical power across simulations. This sample size approximates that of moderate-sized oncological clinical trials, which offers a realistic setting for model comparison (Grant et al. 2019).
2. **Artificial Censoring:** To mimic clinical trial follow-up limitations, we subject each dataset to artificial censoring at the 50th percentile of its empirical Kaplan–Meier survival distribution. This approach is consistent with methodological guidance for survival extrapolation studies, which emphasizes the importance of evaluating model performance under conditions

of limited data maturity. (Grant et al. 2019). The choice of 50% survival as the censoring point simulates a scenario of moderate data maturity, balancing the need for sufficient events with realistic clinical trial follow-up constraints.

3. **Replication Strategy:** Each analysis (sample-and-censor cycle is repeated 100 times per dataset to assess method stability and provide uncertainty quantification and ensure robustness. This repetition approach follows established practices in survival method validation studies and allows for assessment of the consistency of model selection performance across different random samples (Gray, Hernandez, and N. Latimer 2020).

4. **Model Fitting and Selection:** In each simulation, both the Parametric and spline-based cure models are fit to each sample. Two model selection strategies are evaluated:

   - **Traditional Information Criteria (AIC/BIC):** Each model is fit on the entire sample, and AIC/BIC are computed from that single fit.

   - **Cross-Validated AIC/BIC:** Each sample is split into 10 folds. Models are trained on nine folds and validated on the one remaining fold. This is repeated across all folds. The average validation log-likelihood across folds is used to compute cross-validated AIC and BIC. The model with the best cross-validated score is selected.

4.1.6 Evaluation Metrics

Restricted Mean Survival Time (RMST): RMST is the primary metric for assessing survival prediction accuracy in this study. RMST provides a robust and clinically interpretable summary of the survival curve, even when a cured subgroup exists. It integrates the area under the survival curve up to a specified time horizon, offering a more complete view of expected survival (Royston and Altman 2013) .
In this experiment, the reference standard RMST is calculated using the Kaplan-Meier estimator fitted to the complete original dataset. This benchmark represents the target value against which we compare the extrapolated RMST with the predictions from fitted models.

Absolute RMST Error: For each fitted model, the predicted RMST is computed up to the same time horizon (maximum observed time in the full dataset), and the absolute errors relative to the gold standard are calculated. This allows for comparison between models selected using traditional (AIC/BIC) and cross-validation-based approaches. The emphasis on RMST aligns with its growing use in health technology assessment, especially in contexts requiring survival extrapolation beyond clinical trial follow-up. (N. R. Latimer 2013).

Relative RMST Error: In addition to reporting the absolute RMST error, this study computes the relative RMST error to facilitate interpretation across datasets with different survival scales. Relative error expresses the deviation from the reference RMST as a proportion, making the model accuracy more interpretable, especially when RMST values vary substantially between populations.

It can be defined as the absolute difference between the predicted RMST and the reference RMST (estimated using the Kaplan–Meier curve from the uncensored dataset), divided by the reference RMST:

$$\text{Relative RMST Error (\%)} = \frac{|\text{Estimated RMST} - \text{Reference RMST}|}{\text{Reference RMST}} \times 100$$

This measure supports more balanced comparisons of extrapolation accuracy and is consistent with best practices in model evaluation, where scale-independent metrics are desirable (Royston and Parmar, 2013; Latimer, 2013).

By applying this evaluation framework across the repeated simulations the study examines how model selection strategies, particularly cross-validation versus traditional AIC/BIC, affect the reliability of cure fraction estimation under varying degrees of data maturity. The number of times each distribution was selected under AIC/BIC versus CV was recorded to identify selection trends.

4.1.7 Statistical Implementation

The analyses were conducted in R version 4.5.0 using the following packages:

- **flexsurvcure:** For mixture cure model fitting
- **survival and survminer:** Used for non-parametric survival analysis, Kaplan–Meier estimation, and visualization.
- **flexsurv:** For flexible parametric survival models

4.2 Inverse Probability of Censoring Weighting

This section answers the secondary research question, which investigates whether inverse probability of censoring weighting (IPCW) improves the accuracy of survival extrapolation from datasets with high levels of censoring.

4.2.1 Theoretical foundation

Inverse Probability of Censoring Weighting (IPCW) is a statistical correction technique used to address the bias introduced by right-censoring in survival data. In right-censored datasets, the survival time of some individuals is unknown beyond a certain point, either because they were lost to follow-up or the study ended before they experienced the event of interest. This can distort model estimation, especially when censoring is substantial.

The idea behind IPCW is to add weights to the observed (uncensored) events to make them more representative of the full population. Specifically, each event is weighted by the inverse probability of remaining uncensored up to that time. These probabilities are estimated from the data using the Kaplan–Meier estimator of the censoring distribution (Hernán and Robins 2020). As a result, events that occur in time intervals with higher censoring are given more weight, balancing the bias introduced by the censored observations.

High censoring levels are common in clinical studies with limited follow-up durations. When censoring is unevenly distributed over time, standard survival models may underestimate the survival probabilities and misrepresent long-term outcomes.

IPCW provides a method to address this issue by accurately reconstructing the survival experience of censored individuals.

Cure models, and flexible survival models more broadly, are sensitive to such censoring-induced bias, especially in situations of extrapolation where the model predictions extend beyond the range of observed events. By applying IPCW, the estimation procedure accounts for the incomplete information and allows for more accurate curve fitting and RMST estimation.

Implementation in the Experiment:

In this study, IPCW is implemented at the level of each bootstrapped sample. For every iteration, we derive inverse probability weights from the Kaplan–Meier estimate of the censoring distribution. These weights are used to compensate for information lost as a result of right-censoring by reweighting the observed events. Both parametric and spline-based survival models are fitted with and without IPCW weights, and their extrapolation accuracy is assessed. Performance is evaluated using the absolute error in restricted mean survival time (RMST), calculated against a reference standard RMST, obtained from the whole, original dataset with low censoring. This approach enables a direct comparison of IPCW versus traditional methods in settings characterized by high levels of censoring.

4.2.2 Experiments

The analysis follows a structured simulation framework that includes data preparation, controlled censoring, bootstrap resampling, dual model fitting (with and without IPCW), and evaluation using RMST error.

Step 1: Dataset Preparation and Selection

The following three individual patient-level datasets were selected for their low censoring rates and long follow-up periods to allow reliable estimation of long-term survival:

- **SEER Pancreatic Cancer Cohort:** This dataset was derived from the Surveillance, Epidemiology, and End Results (SEER) Program, a cancer registry maintained by the U.S. National Cancer Institute. It includes N = 35,225 patients diagnosed with pancreatic cancer. The observed censoring rate is relatively low (approximately 6.5%), with follow-up times ranging from 1.0 to 510.0 months. This long-term registry dataset provides a highly mature survival curve, making it a strong candidate for use as a "ground truth" reference in evaluating extrapolation methods.(National Cancer Institute 2023)
- **SEER Small-Cell Lung Cancer (SCLC) Cohort:** This dataset was also sourced from the SEER database and it comprises N = 25,855 patients diagnosed with small-cell lung cancer. Like the pancreatic cohort, it exhibits a low censoring rate (around 7%), providing a relatively complete survival profile suitable for simulating artificially censored scenarios. The use of large, population-based SEER cohorts allows the modeling of real-world survival patterns in oncology.(National Cancer Institute 2023)
- **NSCLC Immunotherapy Trial Dataset:** This dataset was reconstructed from the Kaplan–Meier overall survival curve published in (Peters et al. 2025), based on pooled patient-level data from the CheckMate 227 and CheckMate 9LA trials. It includes $N = 322$ patients with metastatic non-small cell lung cancer (NSCLC) and low PD-L1 expression, treated with immune checkpoint inhibitors. The dataset exhibits a low censoring rate (21.7%) and follow-up ranging from 0.6 to 85.7 months.

Each dataset goes through standard cleaning procedures, including the removal of missing or zero follow-up times. Variables were standardized to *time* (survival duration) and *status* (1 = event, 0 = censored). The immunotherapy dataset was handled with the possibility of a cured patient subgroup in mind. The low censoring in these datasets supports the generation of credible reference benchmarks before artificial censoring.

Step 2: Reference RMST Calculation

For each original dataset, we estimate the RMST using the Kaplan-Meier method up to the 90th percentile of observed survival times. These RMST values served as reference standards for evaluating extrapolation accuracy. This approach follows recommendations for setting clinically meaningful follow-up horizons (Royston and Altman 2013).

Step 3: Increasing Censoring in the Original Datasets

To simulate higher censoring scenarios, each dataset was modified to produce three new versions with target censoring rates of approximately 40%, 60%, and 80%. The censoring cutoff was defined using the Kaplan–Meier survival curve of the original dataset. Observations beyond the cutoff time were administratively censored, and their status was updated accordingly. This method provides consistent censoring conditions for simulation across all datasets (Grant et al. 2019).

Step 4: Bootstrap Sampling and IPCW Weighting

For each version of the dataset with increased censoring, 100 bootstrap samples (n = 300, with replacement) were drawn. IPCW weights were computed for each sample using the Kaplan–Meier estimator of the censoring distribution. The weights for the uncensored events were set as the inverse of the probability of remaining uncensored just before the event time (Robins, Rotnitzky, and Zhao 1994). Censored observations received zero weight. To prevent instability from large weights, a 95th percentile cap was applied (Cole and Hernán 2004).

Step 5: Survival Model Fitting

Each bootstrap sample was analyzed using both unweighted and IPCW-weighted methods. The survival models used included:

- Weibull and log-normal accelerated failure time (AFT) models (`survreg`),
- Weibull proportional hazards models (`flexsurvreg`),
- Royston–Parmar spline models (`flexsurvspline`),
- Weibull mixture cure models (`flexsurvcure`) for the immunotherapy dataset.

The spline models were initially configured with one or two internal knots, and when convergence failed, simpler settings or alternate scale functions (e.g., hazard, odds, or normal) were used. The model configuration and convergence outcome were recorded. This model set supports the evaluation of IPCW's effect across standard and flexible modeling approaches.

Spline Model Convergence Strategy: In this experiment, Spline models were initially fit with one or two internal knots. When convergence failed, a fallback strategy was triggered, sequentially trying simpler configurations and alternate scale types (hazard, odds, normal) until a model converged. This callback approach ensured flexible models were included in the IPCW evaluation while maintaining computational stability. This design supports a robust assessment of IPCW's performance across both traditional parametric and flexible spline-based modeling frameworks, in line with good practices in survival analysis (Gray, Hernandez, and N. Latimer 2020)(Gray et al., 2020).

Step 6: RMST and Error Calculation

The RMST for each fitted model was estimated using the same time horizon defined in Step 2. For supported models, RMST was extracted using summary functions; for others, numerical integration was used. The absolute RMST error was calculated as:

$$\text{Error} = |\text{RMST}_{\text{model}} - \text{RMST}_{\text{reference}}| \tag{2}$$

The improvement from IPCW was defined as the reduction in absolute error compared to the traditional model:

$$\text{Improvement} = \text{Error}_{\text{Traditional}} - \text{Error}_{\text{IPCW}} \tag{3}$$

Positive values indicated that IPCW led to more accurate extrapolation.

Step 7: Aggregating Simulation Results

For each iteration, the model results were stored, including RMST estimates, errors, AIC values, convergence status, and IPCW weight statistics. These results were aggregated by dataset, censoring level, and model type.

4.2.3 Statistical Implementation

All analyses were conducted in R (version 4.5.0). Among the primary packages utilized were:

- `survival` (for basic survival functions like `survreg` and `Surv` objects),
- `flexsurv` (for flexible parametric survival models including `flexsurvspline` and `flexsurvreg`),
- `flexsurvcure` (for fitting mixture and non-mixture cure models).

5 Results

5.1 Model Selection in Cure Fraction Models using k-fold Cross Validation

This section presents the findings from the primary research question, which is an investigation into whether k-fold cross-validation (CV) enhances model selection for survival extrapolation, with a particular focus on its performance with mixture cure models in datasets potentially exhibiting long-term survival plateaus.

5.1.1 RMST Interpretation

In order to understand the comparative analyses better, we first clarify our primary outcome measure, the Restricted Mean Survival Time (RMST). The RMST up to a specific time point τ (tau) quantifies the average time a patient remains alive (or event-free) within that defined observation window from time 0 to τ. For example, if time is measured in months and we calculate RMST up to $\tau = 60$ months, an RMST value of 45 months signifies that, on average, patients in the group survived for 45 months out of that initial 60 month period. This metric provides an easily interpretable summary of the survival experience over a chosen, clinically relevant time frame and is calculated as the area under the survival curve up to τ.

We compared model selection guided by traditional information criteria (AIC and BIC) with a CV-based approach, evaluating performance by absolute error in RMST and relative RMST relative to a reference standard RMST from minimally censored data.

5.1.2 Absolute RMST Error

Table 1: Comparison of Model Selection Methods on RMST Absolute Error

Dataset	**Reference RMST**	**AIC-based Selection**			**BIC-based Selection**		
		Traditional	CV	Improvement	Traditional	CV	Improvement
OS Ipilimumab	25.88	4.08	4.05	0.8	3.54	3.73	–5.4
OS AntiPD1	35.46	1.81	1.82	–0.2	1.77	1.79	–0.7
OS IpiNivo	35.90	1.47	1.49	–1.4	1.45	1.47	–1.4
NSCLC Immunotherapy	29.92	5.22	4.97	4.9	1.16	2.65	–128.1
NSCLC Chemotherapy	19.30	2.88	2.91	–0.8	4.23	3.70	12.6
SEER Breast	100.39	38.28	37.61	1.8	48.38	38.64	20.1
Ovarian	1374.49	664.47	652.91	1.7	745.22	687.28	7.8
GBSG	1659.74	47.58	47.95	–0.8	48.21	48.05	0.3

Table 1 presents the mean absolute RMST errors (in months) for models selected via traditional (AIC/BIC) and cross-validated (CV) information criteria. The 'Improvement (%)' column indicates the percentage reduction in absolute RMST error when using CV. Positive values favor CV.

AIC-based Selection: When using AIC as the underlying criterion, CV-guided selection led to a reduction in absolute RMST error in four of the eight datasets: OS Ipilimumab (0.8% improvement), NSCLC Immunotherapy (4.9% improvement), SEER Breast (1.8% improvement), and Ovarian

(1.7% improvement). For the remaining datasets, traditional AIC performed marginally better or equally.

BIC-based Selection: When comparing CV BIC to traditional BIC, CV resulted in lower absolute RMST error in three datasets: NSCLC Chemotherapy (12.6% improvement), SEER Breast (20.1% improvement), and Ovarian (7.8% improvement). However, for OS Ipilimumab and notably NSCLC Immunotherapy (−128.1% improvement, meaning traditional BIC was substantially better), traditional BIC selection led to more accurate RMST estimates. Performance was similar for the other datasets.

Overall, the impact of CV on reducing absolute RMST error was inconsistent; while some improvements were noted, CV did not universally outperform traditional information criteria in this regard. Figure 2 presents the comparative performance.

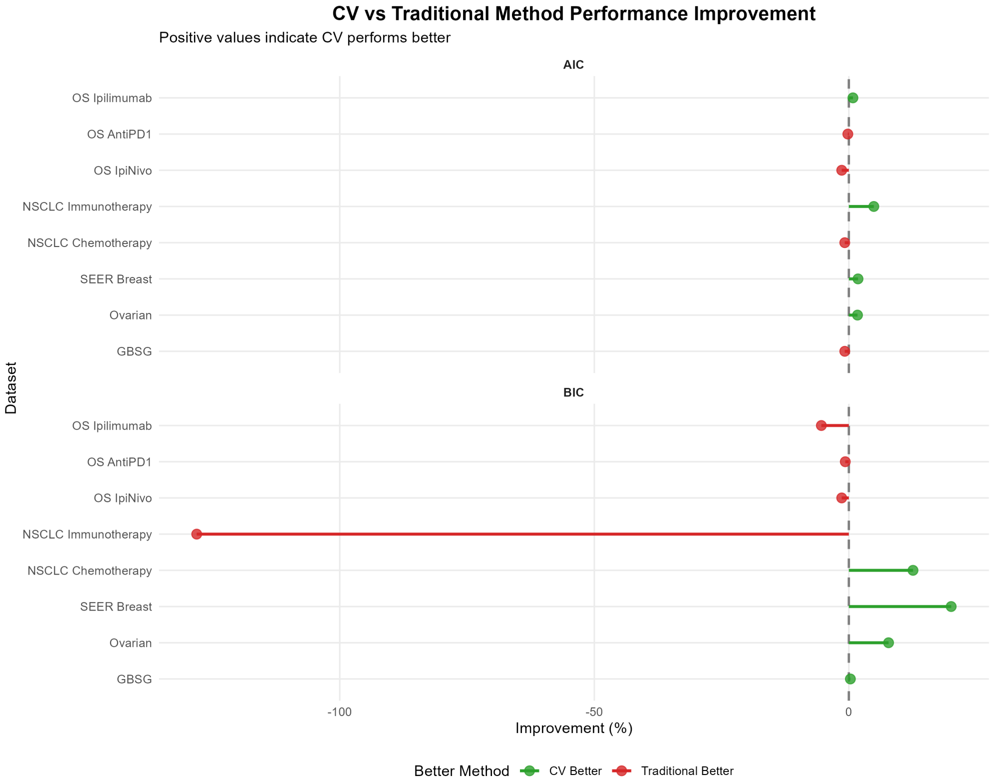

Figure 2: Cross-validation versus traditional model selection improvement analysis. The lollipop chart displays the percentage improvement in RMST prediction accuracy when using cross-validation compared to traditional AIC and BIC selection methods. Positive values indicate superior CV performance, while negative values favor traditional approaches. The mixed pattern of improvements across datasets demonstrates the context-dependent nature of cross-validation benefits in survival model selection.

Table 2: Relative RMST Prediction Error (%) by Model Selection Method

Dataset	**AIC (Trad)**	**AIC (CV)**	**BIC (Trad)**	**BIC (CV)**
OS Ipilimumab	15.8	15.6	13.7	14.4
OS AntiPD1	5.1	5.1	5.0	5.0
OS IpiNivo	4.1	4.2	4.0	4.1
NSCLC Immunotherapy	17.5	16.6	3.9	8.9
NSCLC Chemotherapy	14.9	15.1	21.9	19.2
SEER Breast	38.1	37.5	48.2	38.5
Ovarian	48.3	47.5	54.2	50.0
GBSG	2.9	2.9	2.9	2.9

5.1.3 Relative RMST Error:

The relative RMST error is a performance measure that adjusts for differences in time scale, making it possible to compare results across diseases with varying follow-up periods. Table 2 shows the relative RMST prediction errors for each dataset.

The lowest relative errors (under 5%) were observed in the OS AntiPD1 and OS IpiNivo datasets, suggesting that the predicted RMSTs closely matched the reference values. By contrast, the SEER Breast and Ovarian datasets showed high relative errors (35%), reflecting the greater difficulty of extrapolating survival over long follow-up periods. In most datasets, cross-validation (CV) improved or maintained prediction accuracy; however, it occasionally led to overfitting, as seen in the NSCLC Immunotherapy data under BIC selection, where relative error increased notably from 3.9% to 8.9%. The relative error analysis presented in Figure 3 demonstrates significant variation in model prediction accuracy across the eight cancer datasets.

AIC-based Selection: CV-guided AIC selection resulted in lower relative RMST error compared to traditional AIC in OS Ipilimumab (15.6% vs. 15.8%), NSCLC Immunotherapy (16.6% vs. 17.5%), SEER Breast (37.5% vs. 38.1%), and Ovarian (47.5% vs. 48.3%). In other datasets, performance was similar or marginally favored traditional AIC.

BIC-based Selection: For BIC-based selection, CV led to lower relative error in NSCLC Chemotherapy (19.2% vs. 21.9%) and SEER Breast (38.5% vs. 48.2%), and Ovarian (50.0% vs. 54.2%). However, traditional BIC produced substantially lower relative error for NSCLC Immunotherapy (3.9% vs. 8.9%) and slightly better or similar performance in the remaining datasets.

Overall Observation: The relative error perspective confirms the mixed impact of CV. For example, while CV-BIC showed a large percentage improvement in absolute error for SEER Breast, its relative error (38.5%) was still substantial, though better than traditional BIC (48.2%). For NSCLC Immunotherapy, traditional BIC achieved a very low relative error of 3.9%, which was considerably better than CV-BIC (8.9%).

For the NSCLC Immunotherapy dataset, cross-validation helped AIC slightly but hurt BIC. The best prediction came from the traditional BIC-selected model, which was off by less than 4% compared to the actual average survival estimate. This tells us that sometimes more straightforward selection rules (like traditional BIC) can outperform more complex strategies like CV, depending on the dataset.

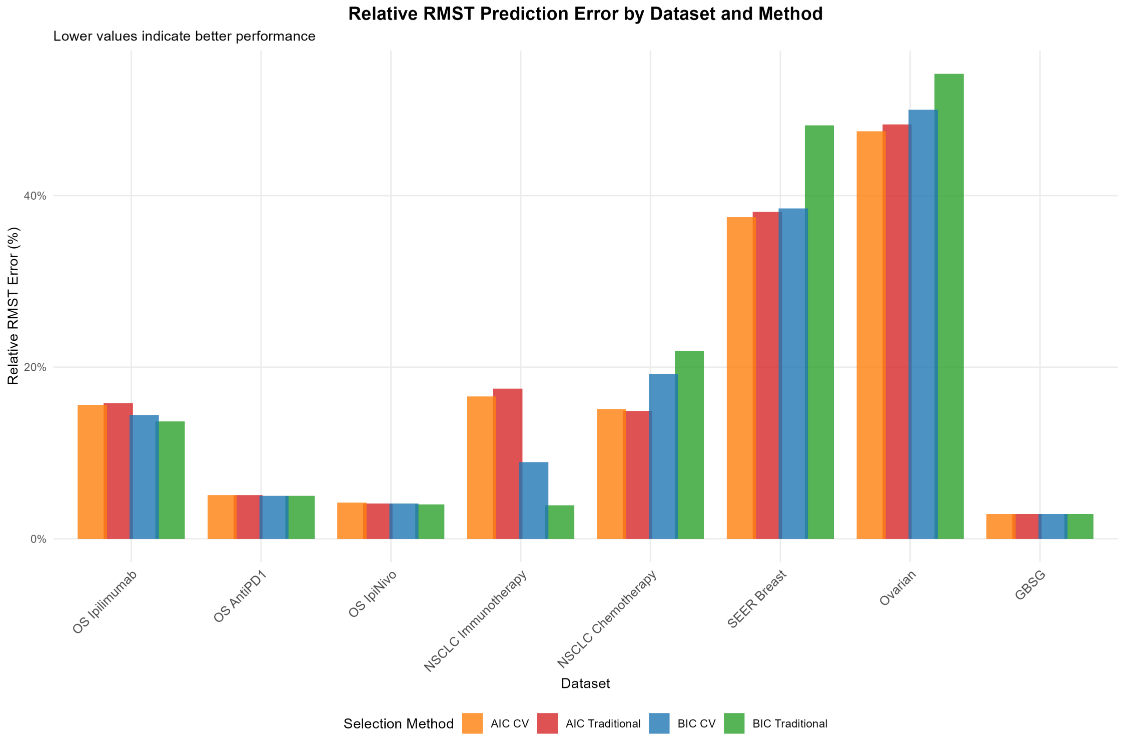

Figure 3: Relative RMST prediction error comparison across datasets and model selection methods. The grouped bar chart displays the relative error percentages for traditional AIC, cross-validation AIC, traditional BIC, and cross-validation BIC selection approaches across all eight cancer datasets. Lower values indicate higher prediction accuracy.

Model Selection Patterns

To understand which types of models were being selected, we analyzed the frequency of model selection across simulations for each dataset.

Table 3: Model Selection Patterns: Most Frequently Selected Models (n = 100)

Dataset	**Traditional Model**	**Frequency (%)**	**CV Model**	**Frequency (%)**
OS Ipilimumab	Spline k=1 normal	69	Spline k=1 normal	71
OS AntiPD1	Spline k=4 hazard	35	Spline k=4 hazard	35
OS IpiNivo	Spline k=1 normal	57	Spline k=1 normal	61
NSCLC Immunotherapy	Spline k=4 odds	30	Spline k=4 odds	37
NSCLC Chemotherapy	Spline k=3 normal	53	Spline k=3 normal	56
SEER Breast	Spline k=4 normal	31	Spline k=2 normal	29
Ovarian	Spline k=2 hazard	35	Spline k=2 hazard	31
GBSG	Spline k=1 hazard	37	Spline k=1 hazard	37

- A notable observation was the consistent preference for flexible spline models by traditional (AIC/BIC) and CV-guided selection methods across all eight datasets. As shown in Table 3, spline models were the most frequently selected model type in 100% of datasets for both approaches.

- Notably, despite the inclusion of mixture cure models in the candidate set and the presence of datasets where long-term survival plateaus might be expected (e.g., OS Ipilimumab, OS AntiPD1, OS IpiNivo, NSCLC Immunotherapy)—mixture cure models were not selected as the most frequent best-fitting model by either traditional criteria or CV in any of the eight datasets.

- There was also a high level of agreement between the model families chosen by traditional and CV-guided methods. In all datasets, both approaches selected spline-based models. In many cases, the same spline configuration was selected (e.g., Spline k=1 normal in OS Ipilimumab and OS IpiNivo, and Spline k=1 hazard in GBSG), indicating strong consistency in model preference.

5.2 RMST estimation with IPCW

This section presents the findings from the secondary research question, which investigates how IPCW improves long-term survival extrapolation when working with highly censored data. The model accuracy was evaluated by comparing estimated RMST values to a reference RMST derived from the complete, minimally censored datasets.

Overall Effectiveness of IPCW

On average, IPCW reduced RMST error by 1.062 units. When interpreted in time units, it corresponds to an average improvement of approximately 1.06 months in RMST prediction compared to traditional (unweighted) models. IPCW resulted in improved extrapolation accuracy in 91.0% of all comparisons, demonstrating a consistent benefit across a wide range of settings.

Artificial Censoring and Achieved Censoring Rates

To simulate higher censoring scenarios, we modified the dataset to produce three new versions with target censoring rates of approximately 40%, 60%, and 80%. Observations beyond the cutoff time—determined from the Kaplan–Meier survival curve of the original dataset—were administratively censored, and their status was updated accordingly, as described in the Methods section (Step 3 of Section 4.2.2) (Grant et al. 2019).

Across all datasets, the achieved censoring rates were generally close to the target values, with mean deviations ranging from 1% to 3.4% and a maximum deviation of 8% (Table 4). The highest deviations were observed in the 60% SEER_Pancreas and 80% SEER_SCLC datasets.

Although exact target censoring levels were not always achieved, deviations were minor and the overall pattern of censoring was consistent across datasets, thus preserving the validity of subsequent analyses.

Table 4: Achieved censoring rates across datasets.

Master Dataset	**Target %**	**Achieved %**	**Error %**
SEER_Pancreas_Censored_40pct	40	38.5	1.5
SEER_Pancreas_Censored_60pct	60	52.0	8.0
SEER_Pancreas_Censored_80pct	80	77.5	2.5
NSCLC_Immunotherapy_Censored_40pct	40	38.8	1.2
NSCLC_Immunotherapy_Censored_60pct	60	59.0	1.0
NSCLC_Immunotherapy_Censored_80pct	80	79.5	0.5
SEER_SCLC_Censored_40pct	40	37.6	2.4
SEER_SCLC_Censored_60pct	60	58.7	1.3
SEER_SCLC_Censored_80pct	80	74.6	5.4

Effect at different Censoring levels

We further analyzed IPCW performance across varying levels of artificial censoring to assess how the benefit changes with censoring intensity:

- **40% Censoring:** IPCW reduced RMST error by 0.342 units, equivalent to approximately .34 months. It was beneficial in 6% of comparisons.
- **60% Censoring:** The largest benefit was observed at this level, with an average RMST error reduction of 2.02 units or 2.02 months, improving accuracy in 91.6% of comparisons.
- **80% Censoring:** Even with very high censoring, IPCW improved RMST accuracy by 0.933 units (about .93 months in 84.2% of cases.

These findings indicate that IPCW is especially helpful in moderate to high censoring scenarios, where traditional models often lack sufficient event information to extrapolate survival accurately.

Performance by Censoring Level

Table 5 summarizes IPCW's effectiveness across different artificial censoring scenarios. The greatest improvement in RMST accuracy occurred at 60% censoring.

Table 5: Performance by Censoring Level

Target Censoring (%)	Number of Comparisons	Mean RMST Improvement	Beneficial Comparisons (%)
40	1,000	0.342 units	96.0
60	856	2.020 units	91.6
80	800	0.933 units	84.2

These results indicate that IPCW is helpful across a range of high censoring levels, with powerful benefits observed when approximately 60% of patient data is censored. Notably, the performance drops slightly at 80% censoring, possibly due to fewer available events and increased uncertainty. This can be viewed in the Figure 4 below.

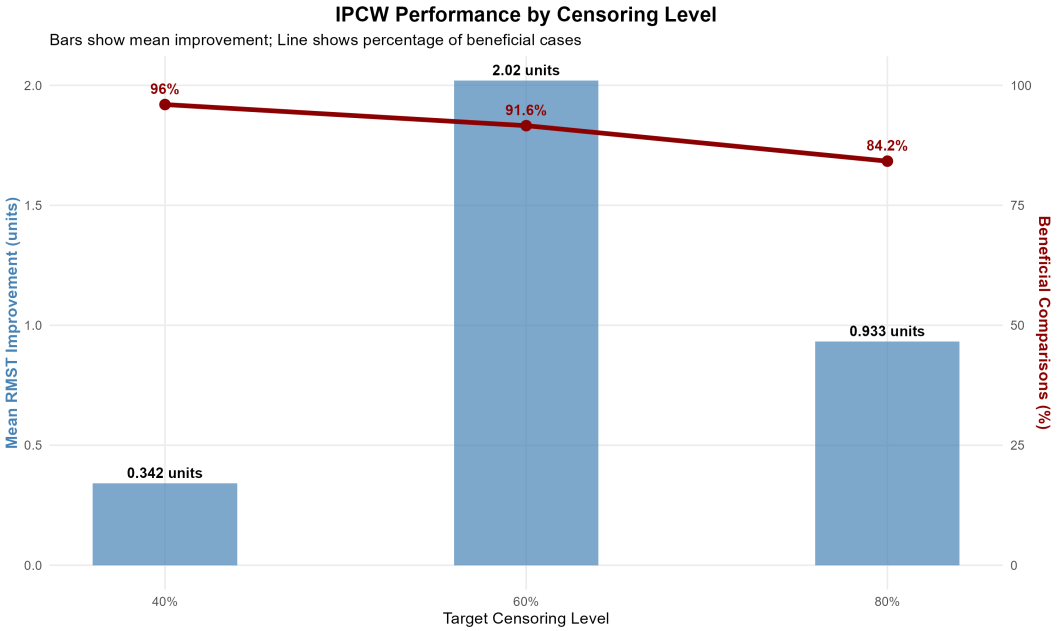

Figure 4: IPCW performance across different censoring levels. The dual-axis plot displays mean RMST improvement (blue bars) and percentage of beneficial comparisons (red line with points) for target censoring levels of 40%, 60%, and 80%. Peak IPCW effectiveness occurs at 60% censoring with a mean improvement of 2.02 units and 91.6% beneficial cases. Performance declines at higher censoring levels (80%) where mean improvement drops to 0.93 units, though 84.2% of comparisons still show benefits. The consistent high success rates across all censoring levels (84-96%) demonstrate the effectiveness of IPCW adjustment.

The number of comparisons differs slightly across censoring levels due to occasional model convergence failures. These failures were more frequent at higher censoring levels, where the limited number of observed events made it difficult for some complex models (e.g., splines or cure models) to estimate parameters reliably. Only successfully converged models were included in each evaluation.

These results indicate that IPCW is helpful across a range of high censoring levels, with powerful benefits observed when approximately 60% of patient data is censored. Notably, the performance drops slightly at 80% censoring, possibly due to fewer available events and increased uncertainty.

Performance by Survival Model Type

Table 6 shows IPCW's impact across different survival model types. RMST improvements are reported in both raw units and interpreted as approximate time in months.

IPCW demonstrated the most significant benefit when applied to cure models and parametric Weibull models. Log-normal models exhibited more variability and lower overall benefit, likely due to their heavier-tailed survival shape, which may interact less predictably with the weighting

Table 6: Performance by Model Type

Model Type	Number of Comparisons	Mean RMST Improvement (units)	Success Rate (%)	Best	Worst
Cure Weibull	300	1.590	99.0	7.39	-0.52
Spline	556	1.300	97.3	7.76	-0.52
Weibull	900	1.290	99.4	4.36	-2.35
Log-normal	900	0.512	76.1	5.43	-4.28

scheme. The model-specific IPCW performance illustrated in Figure 5 reveals the various responsiveness to censoring adjustment across survival model types.

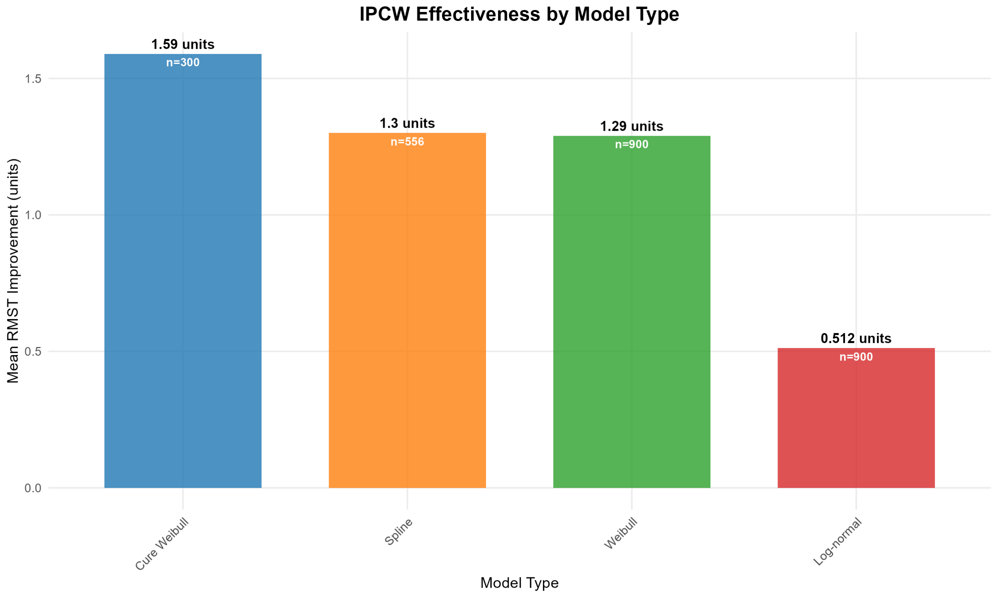

Figure 5: IPCW effectiveness by model type showing mean RMST improvement across all censoring levels. The bar chart displays the average improvement in RMST prediction accuracy achieved through inverse probability of censoring weighting for different survival model types. Cure Weibull models demonstrated the highest mean improvement (1.59 units), followed by Spline models (1.30 units) and standard Weibull models (1.29 units). Log-normal models showed the lowest improvement (0.51 units) but benefited from IPCW adjustment. Sample sizes (n) represent the total number of model comparisons across all censoring scenarios. Results demonstrate that IPCW provides consistent benefits across diverse model types, with parametric cure models and flexible spline approaches showing the greatest responsiveness to censoring adjustment.

The number of comparisons varies by model type because not all models were applicable to every dataset. For instance, cure models were only fitted to the immunotherapy dataset. Standard parametric and spline models were applied across all datasets.

Dataset Level Insights

To better understand IPCW's impact, we examined results by dataset. Full detailed results are included in Appendix 7, stratified by dataset, model type, and censoring level. Below are highlights for each dataset:

- **NSCLC Immunotherapy (Clinical Trial):** IPCW consistently improved RMST accuracy across all model types and censoring levels. For instance, the Weibull cure model at 60% censoring showed a mean RMST improvement of 3.73 units (3.73 months), with a 95% CI of [0.46, 6.62]. Similarly, spline models at 60% censoring improved by 3.86 units (3.86 months), 95% CI: [0.49, 6.50].

- **SEER Pancreatic Cancer (Registry):** This real-world dataset showed significant IPCW benefit for the Weibull model, especially at 80% censoring (mean improvement = 3.23 units; 95% CI: [0.01, 4.20]). Spline models were moderately helpful (e.g., 0.70 units at 40% censoring), although some confidence intervals included zero, suggesting potential uncertainty. Log-normal models showed mixed results, with some scenarios like 80% censoring resulting in worse performance (mean improvement = -2.14 units; 95% CI: [-3.44, 0.29]).

- **SEER Small-Cell Lung Cancer (SCLC):** IPCW was particularly beneficial at higher censoring levels. For the Weibull model at 80% censoring, the RMST improvement averaged 2.86 units (2.86 months), 95% CI: [1.91, 3.58]. Spline models also showed benefit, such as at 60% censoring (mean improvement = 2.34 units; 95% CI: [-0.02, 3.80]). However, spline model convergence failures were common at 80% censoring, limiting their evaluation under extreme censoring conditions.

6 Discussion and Interpretation of the results

6.1 Summary of findings

This study explored advanced survival modeling techniques for improving models' selection and extrapolation accuracy in the context of health economic evaluations, particularly when faced with complex data characteristics such as long-term survival plateaus or high levels of censoring. Two key research questions guided the investigation.

The first research question aimed to determine whether k-fold cross-validation(CV) could improve model selection and long-term survival extrapolation in mixture cure models, especially for datasets with survival plateaus. The analysis revealed that CV led to marginal improvements in RMST prediction accuracy compared to traditional information criteria (AIC and BIC), though the benefit was not uniform across all datasets. Specifically, cross-validation based on AIC selected a more accurate model (in terms of RMST error) in 4 out of 8 datasets, while cross-validation based on BIC showed improved accuracy in 3 out of 8 datasets. This indicates a mixed but meaningful impact, whereby in some settings, CV helped identify models with better generalizability, while in others, traditional methods were equally effective or slightly better.

An important pattern observed across all eight datasets was the dominant selection of spline-based models by both traditional criteria and CV approaches. In every case, flexible splines emerged as the most frequently chosen model type, reflecting their adaptability in capturing complex hazard shapes and long-term survival behavior. However, despite the inclusion of mixture cure models in the candidate model set and the presence of datasets where long-term survival plateaus were plausible, none of the selection methods identified cure models as the best fit in any dataset.

The second research question focused on the effect of inverse probability of censoring weighting (IPCW) on extrapolation accuracy when datasets are highly censored. Simulation experiments conducted on three datasets (two population registries and one clinical trial) demonstrated that IPCW consistently improved RMST accuracy, particularly under moderate to high censoring conditions. On average, IPCW improved extrapolation by approximately 1.06 RMST units and was beneficial in over 91% of model comparisons. The greatest improvements were observed at 60% censoring, with the Weibull and spline models benefiting the most. Cure models, where applicable, also benefited from IPCW adjustment.

In general, the findings indicate that while traditional model selection techniques remain robust, cross-validation and IPCW offer tangible improvements in specific contexts — particularly when survival data is incomplete or contains a mixture of cured and uncured patients.

6.2 Implications of Findings

The results of this study has several implications for researchers and practitioners involved in survival modeling, particularly in informing health economic evaluations.

First, the mixed performance of cross-validation-based selection methods highlights the importance of context when applying model selection techniques. This suggests that cross-validation may provide advantages over traditional criteria in certain settings. The assumption that more sophisticated selection methods always outperform traditional approaches is challenged in this study, highlighting the context-dependent nature of model selection performance.

Secondly, the consistent selection of flexible spline-based models by CV and traditional methods across the datasets suggests a preference for models that offer sufficient flexibility to capture varying hazard shapes. This suggests that, in the absence of strong parametric assumptions, splines serve as a reliable default modeling strategy.

Furthermore, the fact that cure models were not selected, even in datasets suggesting long-term survival plateaus based on biological or clinical evidence, demsontrates a limitation in current model selection techniques. It suggests that standard information criteria and cross-validation approaches may not be suited for cure model detection. This points to a need for alternative selection criteria specifically designed for cure models.

IPCW and Extrapolation: The IPCW analysis demonstrated that weighting to adjust for censoring can greatly improve extrapolation accuracy across a range of models and datasets. On average, IPCW improved RMST extrapolation by approximately 1.06 units (interpreted as months), and in over 91% of model comparisons, it outperformed unweighted approaches. This reinforces the idea that high censoring rates, common in oncology and real-world datasets, can meaningfully distort survival estimates if not appropriately addressed.

IPCW demonstrated the most benefit at moderate-to-high censoring levels (e.g., 60%), with a slight attenuation at 80%. This might indicate a threshold beyond which the data becomes so sparse that even IPCW, despite weight stabilization, struggles to compensate fully. It was especially effective for cure and Weibull models, which are widely used in health technology assessments, which suggests that IPCW can strengthen their utility in these evaluations.

This has important implications for analysts working with immature or early-phase data, as it provides a statistically grounded method to mitigate information loss due to censoring.

6.3 Possible Drawbacks of the Used Methods

While the study outlines a useful framework for survival model selection and extrapolation under censoring, a number of limitations should be considered:

1. Lack of covariate adjustment: The models in this study were fitted without covariates, in order to focus on extrapolation performance and ensure comparability across datasets. However, in real-world settings, covariates such as age, stage, and treatment arm are often critical for accurate survival prediction.

2. Use of low-censoring data as reference: The reference RMST values were calculated using the original datasets, which had relatively low levels of censoring (typically under 25%). Though this provided a practical benchmark, it is not equivalent to having fully observed survival times. The presence of even moderate censoring in the reference data introduces some degree of approximation in evaluating extrapolation accuracy.

3. Nature of Artificial Censoring: The artificial censoring applied was not based on covariates. As a result, the performance of IPCW might differ in scenarios with significant covariate-dependent informative censoring. The covariate-free IPCW application primarily used in this study tested adjustment for the amount of censoring.

4. Digitization of clinical trial KM curves: Due to the limited availability of individual patient data (IPD) from immunotherapy trials, some of the datasets were reconstructed from published Kaplan–Meier survival curves. Although digitization methods are widely used, they may introduce a few inaccuracies in event times or censoring status. Such errors could affect both model fitting and the evaluation of extrapolation accuracy.

6.4 Future research

1. Explore cure model identification methods: Future research could focus on improving the detection and validation of cure models within survival datasets. This includes developing advanced strategies for identifying when a cure model is appropriate, especially in complex or noisy datasets.

2. Use of covariates in survival and censoring models: All models in this study were univariate to isolate extrapolation behavior, but future studies can include patient-level covariates. This may enable personalized survival predictions, especially important in real world clinical settings.

3. Improved ground truth estimation in the presence of censoring: Although the ground truth in this study used low-censoring datasets as a proxy for the reference survival distribution, future research could leverage fully observed synthetic datasets or apply multiple imputation strategies to more rigorously assess extrapolation accuracy when no gold standard is available.

5. Extension to diverse clinical datasets: The study can be extended to include a wider range of diseases datasets, treatment approaches, and trial structures. This will increase the robustness of the findings and broaden their application in health economic evaluations.

7 Ethical Thinking, Societal Relevance, and Stakeholder Awareness

The findings and methodological advancements presented in this thesis are intended to support survival model selection in health economic evaluations, particularly in the context of oncology and immunotherapy. As such, their use may inform decisions that have direct implications for patient care, treatment funding, and public health policy.

Transparency and Reproducibility: Given the potential real-world impact of this work, the models, code, and analytical decisions remain fully transparent and reproducible. All survival analyses in this study were implemented using open-source statistical software (`R`) and are supported by well documented code. The data used were obtained from publicly available datasets, including registry-based cohorts and digitized clinical trial data, ensuring transparency.

Fairness and Interpretation: The results from the experiments were interpreted fairly and responsibly, particularly in cases where model selection showed instability. Performance metrics such as RMST were chosen for their clinical relevance and interpretability, and results were reported in a way that acknowledges both their potential benefits and limitations.

With regards to Data Privacy and Compliance, all datasets used in this study were either publicly available or reconstructed from published Kaplan-Meier curves. No personally identifiable information was accessed in the analysis.

Societal Relevance: This research aligns with the goals of improving health outcomes and cost-effectiveness in healthcare decision making. These have an overall effect on public health funding, reimbursement decisions, and access to care.

In terms of stakeholder awareness, the research holds relevance for a range of groups including HTA (Health Technology Assessment) bodies, regulatory agencies, clinical researchers, insurance companies and pharmaceutical companies. For HTA agencies, the insights discovered may support the refinement of survival modeling guidelines. For clinicians and trial designers, the findings stress the importance of considering data maturity when planning follow-up durations or interpreting interim results.

8 Conclusion

This study investigated whether advanced statistical techniques, specifically k-fold cross-validation and inverse probability of censoring weighting (IPCW), could enhance survival model selection and improve long-term survival extrapolation accuracy in health economic evaluations.

We used a comprehensive simulation framework across eight diverse cancer datasets to assess the performance of cross-validation-enhanced model selection compared to traditional AIC and BIC criteria. Our candidate models included parametric distributions, flexible splines, and mixture cure models. The results indicated that cross-validation yielded mixed benefits, with improvements seen in approximately half of the tested datasets. When improvements did occur, they were meaningful but modest, suggesting that cross-validation provides selective rather than universal advantages over established selection methods.

Although mixture cure models were incorporated into the analysis and several datasets exhibited distinct survival plateaus, flexible spline models were consistently chosen across all datasets and selection approaches. This was a surprising discovery, which may indicate that conventional cure model formulations might not fully capture the complexity of real-world survival patterns, or that existing model selection criteria may be insufficient to identify situations where cure models are most appropriate.

The evaluation of inverse probability of censoring weighting yielded more encouraging results. IPCW improved RMST-based extrapolation accuracy in over 90% of comparisons, particularly when applied to datasets with moderate to high censoring levels. The technique showed peak effectiveness at approximately 60% censoring and demonstrated particular strength with Weibull and spline-based models. These consistent improvements across diverse scenarios suggest that IPCW represents a valuable and practical enhancement to standard survival modeling approaches.

Several study limitations were observed. Our reference RMST values were derived from datasets with existing censoring rather than completely uncensored data, potentially affecting the accuracy of our benchmarks. In addition, the use of digitized Kaplan-Meier curves for some immunotherapy datasets may have introduced measurement imprecision.

Despite these limitations, this work provides practical guidance for researchers and decision-makers in health technology assessment. Cross-validation appears most valuable in specific contexts rather than as a universal replacement for traditional methods, while IPCW shows promise for routine implementation in analyses involving heavy censoring.

Future research should focus on developing cure-specific model selection criteria that can better identify when mixture cure models are appropriate, and investigating alternative cure model formulations that may better capture the survival patterns observed in cancer datasets.

The findings suggest that while traditional survival modeling approaches remain robust, advanced techniques can provide meaningful improvements when appropriately applied.

A Appendix

A.1 Detailed RMST Analysis by Model and Censoring

Table 7 below presents a detailed, stratified comparison of restricted mean survival time (RMST) estimates—both traditional and IPCW-adjusted—across datasets (NSCLC Immunotherapy, SEER Pancreas, SCLC), varying censoring percentages (40%, 60%, 80%), and model types (Cure Weibull, Log-normal, Spline, Weibull). The "Improvement" column shows the absolute RMST gain contributed by IPCW, with associated 95% confidence intervals.

Table 7: Aggregated Results by Dataset, Censoring Level, and Model Type

Dataset	Model	Censoring (%)	RMST Traditional	RMST IPCW	Improvement (95% CI)
4*NSCLC Immunotherapy	Cure Weibull	40	31.6	31.1	0.44 (0.08, 0.86)
	Log-normal	40	31.0	30.5	0.39 (−0.31, 0.84)
	Spline	40	31.9	31.5	0.42 (0.07, 0.83)
	Weibull	40	39.1	38.8	0.31 (0.11, 0.59)
	Cure Weibull	60	41.3	37.6	3.73 (0.46, 6.62)
	Log-normal	60	36.6	33.9	2.70 (0.52, 4.50)
	Spline	60	39.7	35.8	3.86 (0.49, 6.50)
	Weibull	60	43.3	41.4	1.90 (0.35, 3.17)
	Cure Weibull	80	53.8	53.2	0.60 (0.08, 3.89)
	Log-normal	80	47.4	46.7	0.65 (0.11, 3.54)
	Spline	80	51.2	50.6	0.68 (0.10, 4.25)
	Weibull	80	51.3	50.9	0.45 (0.07, 2.41)
3*SEER Pancreas	Log-normal	40	5.81	5.50	0.26 (−0.14, 0.60)
	Spline	40	6.20	5.32	0.70 (−0.35, 1.34)
	Weibull	40	8.55	8.25	0.30 (0.14, 0.50)
	Log-normal	60	5.68	4.29	−0.36 (−1.27, 1.42)
	Weibull	60	8.81	7.76	1.05 (0.78, 1.36)
	Log-normal	80	5.80	1.98	−2.14 (−3.44, 0.29)
	Weibull	80	9.98	6.02	3.23 (0.01, 4.20)
3*SEER SCLC	Log-normal	40	9.93	9.77	0.16 (−0.05, 0.37)
	Spline	40	9.76	9.48	0.28 (0.04, 0.58)
	Weibull	40	12.40	12.30	0.16 (0.05, 0.29)
	Log-normal	60	11.10	9.10	1.79 (0.31, 2.52)
	Spline	60	11.60	8.29	2.34 (−0.02, 3.80)
	Weibull	60	13.40	12.10	1.33 (0.96, 1.81)
	Log-normal	80	11.60	6.81	1.16 (−3.40, 4.45)
	Weibull	80	14.20	11.40	2.86 (1.91, 3.58)

A.2 Project Code Repository

The code used for the analyses in this dissertation is publicly available on GitHub at:

`https://github.com/Fidelsia/Masters-Theses-analysis-code`

The repository contains two main folders corresponding to the two research questions:

- **Cure_model_Analysis**: Contains the scripts and data, related to the primary research question, which focuses on model selection with cross validation.
- **IPCW_Analysis**: Contains the scripts and data related to the secondary research question, which focuses on inverse probability of censoring weighting (IPCW).

References

Aho, Ken, DeWayne Derryberry, and Teri Peterson (2014). "Model selection for ecologists: the worldviews of AIC and BIC". In: *Ecology* 95.3, pp. 631–636. DOI: 10.1890/13-1452.1.

Amdahl, Jordan (2022). *flexsurvcure: Flexible Parametric Cure Models*. R package version 1.3.1. URL: https://cran.r-project.org/package=flexsurvcure.

Arlot, Sylvain and Alain Celisse (2010). "A survey of cross-validation procedures for model selection". In: *Statistics Surveys* 4, pp. 40–79. DOI: 10.1214/09-SS054. URL: https://projecteuclid.org/euclid.ssu/1268143839.

Berkson, Joseph and Robert P. Gage (1952). "Survival curve for cancer patients following treatment". In: *Journal of the American Statistical Association* 47.259, pp. 501–515. DOI: 10.1080/01621459.1952.10501187.

Bermejo, Inigo and Sabine Grimm (2024). "MSR17 Can Machine Learning Support Survival Model Selection to Inform Economic Evaluations? Exploring K-Fold Cross Validation Based Model Selection in Seven Datasets". In: *Value in Health* 27.12, S441. DOI: 10.1016/j.jval.2024.10.2251. URL: https://doi.org/10.1016/j.jval.2024.10.2251.

Boag, J. W. (1949). "Maximum likelihood estimates of the proportion of patients cured by cancer therapy". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 11.1, pp. 15–53. DOI: 10.1111/j.2517-6161.1949.tb00020.x.

Browne, Michael W. (2000). "Cross-validation methods". In: *Journal of Mathematical Psychology* 44.1, pp. 108–132.

Burnham, Kenneth P. and David R. Anderson (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd. New York: Springer. ISBN: 978-0-387-95364-9. DOI: 10.1007/b97636.

Cole, Stephen R. and Miguel A. Hernán (2004). "Adjusted survival curves with inverse probability weights". In: *Computer Methods and Programs in Biomedicine* 75.1, pp. 45–49.

Edmunson, J. H. et al. (1979). "Different Chemotherapeutic Sensitivities and Host Factors Affecting Prognosis in Advanced Ovarian Carcinoma vs. Minimal Residual Disease". In: *Cancer Treatment Reports* 63, pp. 241–247.

Gallacher, David, Peter Kimani, and Nigel Stallard (2021). "Extrapolating parametric survival models in health technology assessment: a simulation study". In: *Medical Decision Making* 41.1, pp. 37–50.

Grant, Robert L. et al. (2019). "Regression modelling for survival data: Methods beyond the Cox model". In: *European Journal of Cardio-Thoracic Surgery* 56.2, pp. 210–216.

Gray, L., M. Hernandez, and N. Latimer (2020). "Extrapolation of survival curves using parametric models: Current practice in health technology assessment". In: *Medical Decision Making* 40.6, pp. 745–756.

Guyot, Patricia et al. (2012). "Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves". In: *BMC Medical Research Methodology* 12.1, pp. 1–13. DOI: 10.1186/1471-2288-12-9.

Harrell, Frank E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd. New York: Springer. ISBN: 978-3-319-19425-7. DOI: 10.1007/978-3-319-19425-7.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. New York: Springer. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7.

Hernán, Miguel A. and James M. Robins (2020). *Causal Inference: What If*. Boca Raton: CRC Press. URL: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

Jackson, Christopher (2025). *flexsurvspline: Flexible parametric survival regression using the Royston–Parmar spline model*. R documentation, accessed via RDocumentation and rdrr.io. URL: https://www.rdocumentation.org/packages/flexsurv/versions/2.3.2/topics/flexsurvspline.

Jackson, Christopher et al. (2017). “Extrapolating survival from randomized trials using external data: A review of methods”. In: *Medical Decision Making* 37.4, pp. 377–390.

Lambert, Paul C. et al. (2007). “Estimating and modeling the cure fraction in population-based cancer survival analysis”. In: *Biostatistics* 8.3, pp. 576–594.

Latimer, Nicholas R. (2013). “Survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data: Inconsistencies, limitations, and a practical guide”. In: *Medical Decision Making* 33.6, pp. 743–754.

Molinaro, Annette M., Richard Simon, and Ruth M. Pfeiffer (2005). “Prediction error estimation: a comparison of resampling methods”. In: *Bioinformatics* 21.15, pp. 3301–3307.

National Cancer Institute (2023). *Surveillance, Epidemiology, and End Results (SEER) Program*. https://seer.cancer.gov/. https://seer.cancer.gov/.

NICE Decision Support Unit (2013). *Technical Support Document: Survival analysis for economic evaluations alongside clinical trials*. Technical Support Document. London, UK: National Institute for Health and Care Excellence.

Othus, Megan et al. (2012). “Cure models as a useful statistical tool for analyzing survival”. In: *Clinical Cancer Research* 18.14, pp. 3731–3736.

Patel, S. et al. (2016). “Sequential administration of high-dose interleukin-2 and ipilimumab in patients with metastatic melanoma”. In: *Journal of Clinical Oncology* 34.15_suppl, e21041. DOI: 10.1200/JCO.2016.34.15_suppl.e21041. URL: https://doi.org/10.1200/JCO.2016.34.15_suppl.e21041.

Patilea, Valentin and Ingrid Van Keilegom (2017). “A general approach for cure models in survival analysis”. In: *The Annals of Statistics* 45.4, pp. 1612–1646. DOI: 10.1214/17-AOS1511.

Peng, Yingwei and Keith B. Dear (2000). “A nonparametric mixture model for cure rate estimation”. In: *Biometrics* 56.1, pp. 237–243.

Peters, Solange et al. (2025). “Long-term survival outcomes with first-line nivolumab plus ipilimumab–based treatment in patients with metastatic NSCLC and tumor programmed death-ligand 1 lower than 1%: a pooled analysis”. In: *Journal of Thoracic Oncology* 20.1. Open access article under CC BY license, pp. 94–108. ISSN: 1556-0864. DOI: 10.1016/j.jtho.2024.09.1439. URL: https://doi.org/10.1016/j.jtho.2024.09.1439.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao (1994). "Estimation of regression coefficients when some regressors are not always observed". In: *Journal of the American Statistical Association* 89.427, pp. 846–866.

Royston, Patrick and Douglas G. Altman (2013). "External validation of a Cox prognostic model: principles and methods". In: *BMC Medical Research Methodology* 13, p. 33. DOI: 10.1186/1471-2288-13-33.

Royston, Patrick and Mahesh K. Parmar (2002). "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects". In: *Statistics in Medicine* 21.15, pp. 2175–2197.

van Not, Olivier J. et al. (2024). "Long-Term Survival in Patients With Advanced Melanoma". In: *JAMA Network Open* 7.8. Open access article, e2426641. DOI: 10.1001/jamanetworkopen.2024.26641. URL: https://doi.org/10.1001/jamanetworkopen.2024.26641.