Journal of Data Science 0 (0), 1–20
June 2025

DOI: 0000
Research Article

# Predictive Modeling of Short-Term Recidivism: A Mixture Cure Rate Approach with Diverse Link Functions

Chaeyeon Yoo[1,*] and Dipak K. Dey[1]

[1]*215 Glenbrook Rd, Storrs, University of Connecticut, USA*

**Abstract**

Addressing short-term recidivism presents a significant challenge within the criminal justice system, necessitating robust predictive models that can accurately identify risk factors and predict outcomes. This paper introduces application of Bayesian cure rate models to predict short-term recidivism using data from individuals released from Iowa prisons in 2018. The cure rate models effectively capture the recidivism risk by distinguishing between cured individuals—those unlikely to re-offend-and those susceptible to relapse. By implementing Accelerated Failure Time (AFT) models combined with logistic regression and various link functions, including logit, skewed logit, reversed power logit, and flexible generalized logit, we enhance the model's flexibility and interpretability. Through a comparative analysis of these models, key predictors of recidivism are identified and discussed in detail.

**Keywords** *Accelerated failure time; Bayesian inference; generalized logistic distribution; time-to-event*

## 1 Introduction

Addressing short-term recidivism presents a significant challenge within the criminal justice system, necessitating robust predictive modelst that can accurately identify risk factors and predict outcomes. Recidivism, defined as the return to criminal behavior following intervention by the criminal justice system, has serious implications for both rehabilitation and public safety. Statistical modeling has become a crucial tool in analyzing recidivism data, supporting evidence-based decisions.

Survival analysis has gained substantial attention since its early adoption in criminology by Partanen (1969), offering a natural framework for modeling time-to-event outcomes such as time to reincarceration. Within the framework of survival analysis, individuals undergo a period of treatment through correctional programs aimed at rehabilitation, and then released, with interest focused on the probability and timing of recidivism. Applications have spanned traditional survival models, such as Cox proportional hazard models and accelerated failure time (AFT) models for offering critical insights into dynamic risk factors and subgroup differences in recidivism trajectories (Schmidt and Witte, 1989; Visher et al., 2005). However, despite this extensive literature, many existing approaches assume homogeneity in the risk population and fail to account for individuals who may be permanently immune to the event of interest.

One approach that has earned attention is the use of cure rate models (Maltz, 1984; Maller and Zhou, 1996), which acknowledge the presence of distinct sub-populations within the justice-involved community. Although these models have been widely applied in medical statistics, par-

*Corresponding author. Email: chaeyeon.yoo@uconn.edu.

ticularly in oncology, they remain relatively underexplored in criminological research. These models recognize two distinct groups: individuals who are never subject to the event of interest, and those who remain susceptible to relapse over time despite interventions. This partitioning of the population (Schmidt and Witte, 2012) acknowledges that a nontrivial fraction of individuals may never re-offend, leading to decreasing hazard rates over time. Recent work by De la Cruz et al. (2022) introduced a Bayesian mixture cure rate model tailored to recidivism analysis, allowing for the simultaneous estimation of short-term and long-term re-offending risks. This approach enhances the modeling of heterogeneous re-offending behavior by distinguishing between individuals who are temporarily versus permanently desisted from crime.

Mixture cure rate models consist of two key components: the incidence model for the probability of an individual belonging to the susceptible group and the latency model, which captures the survival function among the susceptible individuals. Parametric and semiparametric versions of these models have evolved since their early conceptualization of the cure fraction in cancer studies by Boag (1949) and Berkson and Gage (1952). In particular, Farewell (1977, 1982) and Yamaguchi (1992) introduced frameworks that combined logistic regression for the incidence component with Weibull or accelerated failure time (AFT) models for the latency distribution. More recently, Kuk and Chen (1992) and Sy and Taylor (2000) adopted Cox proportional hazards model for the conditional survival function, while Li and Taylor (2002) and Zhang and Peng (2007) focused on a semiparametric AFT model.

While most studies focus on latency, the mixture model also incorporates the estimation for the proportion of cure individuals with logistic regression. Most studies default to the symmetric logit or probit link functions, which may inadequately model real world data, especially when the fraction of susceptible and insusceptible individuals are unbalanced. Previous research has highlighted the potential issues that arise from using symmetric link functions like logit and probit inappropriately. For instance, Nagler (1994) introduced the scobit model as an alternative to address asymmetric cases by allowing for different rates of change in the response variable. Similarly, Chen et al. (1999) and Kim et al. (2008) utilized generalized t-link models as skewed links for dichotomous response data, offering alternatives for analyzing asymmetric behavior. Building on well-known link functions, Bazán et al. (2017) proposed power and reversal power links, while Prasetyo et al. (2020) developed the flexible generalized logit link function, expanding the repertoire of available link functions. These links enable more accurate estimation of susceptibility probabilities, particularly when asymmetries exist in the data.

In this paper, we present an application of the mixture cure rate model in the context of short-term recidivism prediction with a Bayesian approach. We extend the AFT cure rate model by integrating a variety of link functions for logistic regression: logit, skewed logit, reversed power logit, and flexible generalized logit, which have not been explored in recidivism modeling. These link functions allow for improved flexibility and interpretability in estimating the probability of re-offending, in the presence of skewed group membership. us to better capture the complex relationship between predictors and recidivism likelihood. While this study focuses on extensions of the logit family, the framework can readily accommodate alternatives such as the probit or cauchit links, offering further avenues for model generalization.

The data used in this paper consist of information on a cohort of releases from the Iowa prison system in 2018, with information on demographics, original offense, and risk level. This dataset, available from the Iowa government correctional system database (`https://data.iowa.gov/Correctional-System/Iowa-Prison-Recidivism-and-Change-by-Cohort/dnzw-paxg/about_data`). Bayesian methods are employed for inference and model fitting with `stan`, and model performance is evaluated through Deviance Information Criterion (DIC), Leave-One-Out Infor-

Table 1: Summary statistics for numerical covariates.

| Variable | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| Age | 17.00 | 28.00 | 34.00 | 35.83 | 42.00 | 84.00 |
| Months Supervised | 0.00 | 4.00 | 7.00 | 12.34 | 13.00 | 536.00 |

mation Criterion (LOOIC) and Watanabe-Akaike Information Criterion (WAIC).

The remainder of this paper is organized as follows. Section 2 provides a detailed description of the Iowa 2018 short-term recidivism dataset used in this study. Section 3 introduces the mixture cure rate model framework, outlining both incidence and latency components. Section 4 presents the link functions used in the incidence model, including flexible and asymmetric alternatives to the traditional logit link. Section 5 discusses the model assessment criteria used to evaluate predictive performance. Section 6 presents simulation studies to assess the behavior of the proposed models under controlled conditions. Section 7 applies the methodology to the Iowa recidivism data, offering insights into the practical implications of our approach. Finally, Section 8 concludes with a summary of findings and directions for future research.

## 2 Data Description

The Iowa 3 year recidivism dataset identifies case-level records of 5010 individuals leaving prison and reentering the community in the 2018 cohort (Iowa Department of Corrections, 2024). It provides individual's status over three years to assess whether their reentry into the community was successful or whether the individual was subject to recidivism. For individuals who returned to a correctional facility, the dataset records the time until that event occurred. The dataset encompasses information on age, sex, race, original offense committed, and the individual's level of risk. Tables 1 and 2 provides a detailed overview of variables in our dataset, summarizing both the distribution of numerical variables and the proportion of individuals within each category who experienced recidivism. The first level of each category is set as the baseline. Specifically, the proportion column presents the percentage of the total 5010 subjects represented by each category. For example, females constitute 14.77% of the total subjects.

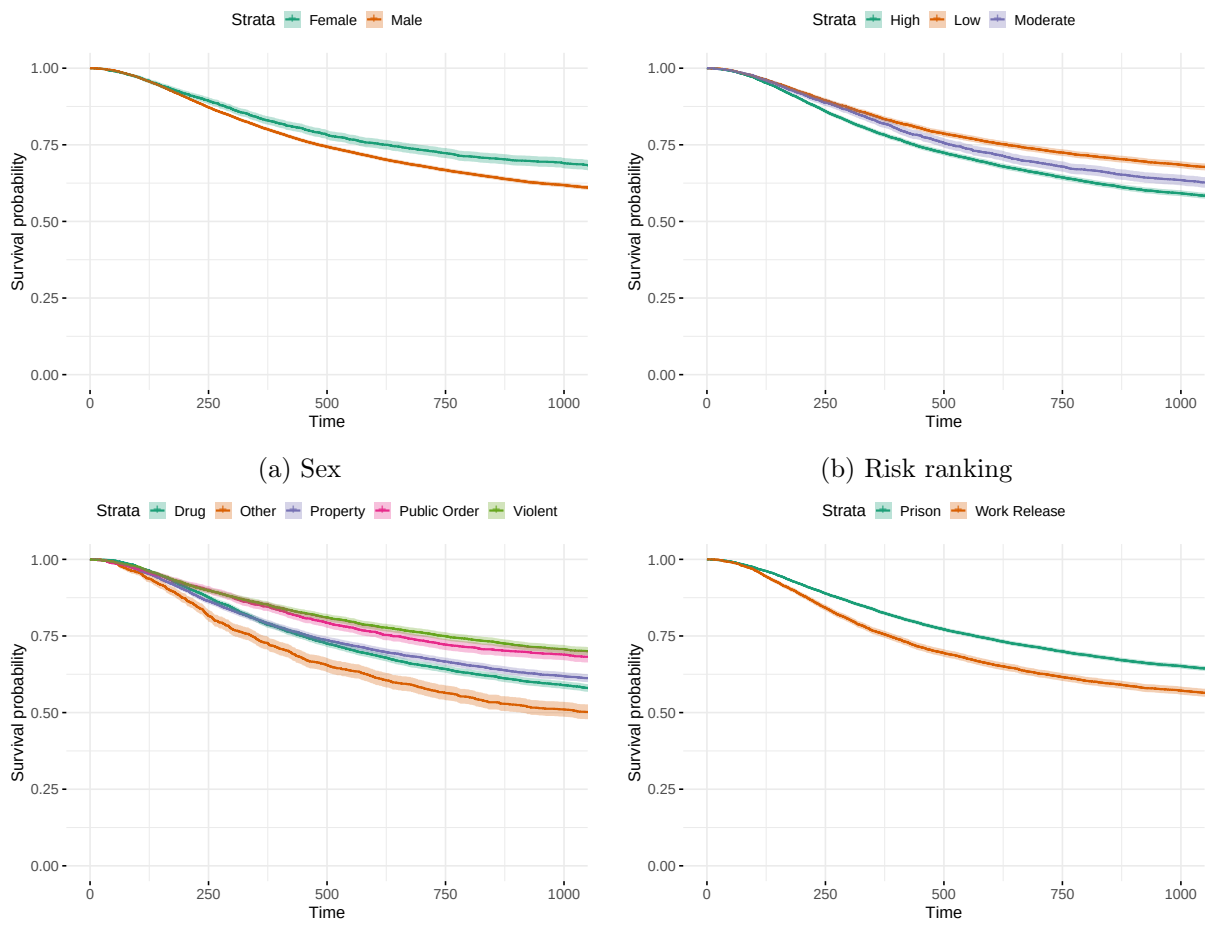
Kaplan–Meier estimates were used to evaluate the proportion of individuals free from recurrence across different covariate groups. For individuals with no observed re-incarceration event, the time to recurrence was set to the maximum follow-up duration of 1,095 days. As illustrated in Figure 1, the survival curves indicate that supervision type, offense type, risk ranking, and sex are all associated with time to re-incarceration among the susceptible population. Comparing the supervision types, those under work release show a higher recurrence of events compared to those in prison. Regarding offense type, subjects with public order and violence related offenses tend to have a higher rate of recidivism. The Kaplan-Meier curves are considerably flat, giving an implication of the survival curve not converging to 0. The flattening of the curves implies the presence of a cured or non-susceptible subpopulation, making standard survival models inappropriate. In addition, the observed imbalance in event occurrence in Table 3 suggests that the assumption of symmetry inherent in standard link functions, such as the logit, may not hold. These issues underscore the need for a mixture cure rate with flexible link functions that can accommodate skewed group memberships. We address these challenges by developing and evaluating a Bayesian mixture cure rate framework that incorporate asymmetric link functions for the incidence model.

Table 2: Overview of categorical covariates

| **Category** | **Level** | **Proportion (%)** |
|---|---|---|
| Race | American Indian or Alaska Native | 2.06 |
| | Asian or Pacific Islander | 0.68 |
| | Black | 21.40 |
| | Hispanic | 5.75 |
| | White | 70.12 |
| Sex | Female | 14.77 |
| | Male | 85.23 |
| Supervising Unit | Anamosa State Penitentiary | 3.59 |
| | Clarinda Correctional Facility | 10.00 |
| | Districts Compact Regions | 0.66 |
| | Fort Dodge Correctional Facility | 38.63 |
| | Iowa Correctional Institution for Women | 14.71 |
| | Iowa Medical Classification Center | 13.91 |
| | Iowa State Penitentiary | 1.46 |
| | Mount Pleasant Correctional Facility | 15.71 |
| | Newton Correctional Facility | 17.13 |
| | North Central Correctional Facility | 9.40 |
| Supervision Type | Work Prison | 70.92 |
| | Work Release | 29.08 |
| Supervision End Reason | Discharged – Expiration of Sentence | 23.79 |
| | Parole Granted | 64.05 |
| | Paroled to Detainer (INS) | 0.90 |
| | Paroled to Detainer (Iowa) | 0.68 |
| | Paroled to Detainer (Out of State) | 1.18 |
| | Paroled to Detainer (U.S. Marshall) | 0.98 |
| | Paroled with Immediate Discharge | 2.59 |
| | Released to Special Sentence | 5.82 |
| Offense Type | Drug | 30.32 |
| | Other | 8.12 |
| | Property | 26.05 |
| | Public Order | 12.12 |
| | Violent | 23.40 |
| Risk Ranking | Low | 51.54 |
| | Moderate | 15.91 |
| | High | 51.54 |

Table 3: Proportion of individuals experiencing recidivism (1) versus non-recidivism (0) in the cohort.

| 0 | 1 |
|---|---|
| 0.613 | 0.387 |




(a) Sex

(b) Risk ranking

(c) Supervision offense type

(d) Supervision release type

Figure 1: Kaplan-Meier curve plot by categorical variable

Let $T_i$ denote the time to recidivism or censoring for individual $i$, and let $\delta_i \in \{0, 1\}$ indicate the event status, where $\delta_i = 1$ if recidivism occurred, and $\delta_i = 0$ if the individual was censored. The observed time is recorded as $y_i = \min(T_i, C_i)$, with $C_i$ is the censoring time. Covariates used in the model include demographics and case-level variables such as age, sex, offense type, and supervision type. These will later be grouped into covariate vectors $\boldsymbol{Z}_i$ and $\boldsymbol{X}_i$ for the survival and cure components of the model, respectively. In the following section, we introduce a Bayesian mixture cure rate model that incorporates flexible link functions to improve the modeling of the cure probability. This framework is designed to better capture the underlying structure of the short-term recidivism in the Iowa dataset.

## 3 Mixture Cure Rate Model

In survival analysis, the survival function $S(t)$ quantifies the probability of observing a survival time $T$ greater than a given value t, expressed as $S(t) = P(T > t)$. Conventionally, it is assumed that $S(t)$ approaches 0 as t approaches $\infty$. However in real-world scenarios, not all subjects

experience the event of interest, leading to situations where $S(t)$ does not converge to 0. This highlights the significance of cure rate models, which are particularly effective in accounting for a subset of subjects who are cured and thereby permanently immune to the event.

The observed survival time for subject $i$ is defined as $y_i = \min(T_i, C_i)$, where $T_i$ represents the true survival time and $C_i$ denotes the censoring time. The censoring status is indicated by the variable $\delta_i$, where

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leqslant C_i \text{ (Uncensored)}, \\ 0 & \text{if } T_i > C_i \text{ (Censored)}. \end{cases}$$

The model includes two sets of covariates: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n\times p}$, which represents the matrix of covariates used in the cure rate component, and $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^T \in \mathbb{R}^{n\times q}$, which corresponds to the covariates in the survival component. The covariate matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ consist of subsets of available covariates, and they may share common elements.

The cure rate model is comprised of two distinct components: a binary component that models the probability of being cured or permanently non-vulnerable to the event of interest, and a survival component that models the time until event occurs for the uncured group. The cured group represents individuals who will never encounter the event.

Defining the cure indicator for individual $i$ as

$$w_i = \begin{cases} 0 & \text{if subject } i \text{ is uncured}, \\ 1 & \text{if subject } i \text{ is cured}. \end{cases}$$

With this indicator, the decomposition for event time $T$ consisting of the susceptible and insusceptible individuals is

$$T = (1 - w)T_s + w\infty,$$

where $T = \infty$ is considered that the event never happens for the cured group. Let $F(t)$ and $S(t)$ denote the cumulative distribution function (cdf) and survival function, respectively, for the overall population. In contrast, $F_s(t)$ and $S_s(t)$ represent the corresponding functions for the susceptible group. The functions $F_s(t)$ and $S_s(t)$ are proper distribution functions of the event time $T$, meaning they are well-defined only for the susceptible group and satisfy the properties of a complete probability distribution (e.g., $S_s(0) = 1$, $\lim_{t\to\infty} S_s(t) = 0$).

We define the survival function conditional on cure status $w$ as

$$\begin{aligned} &\Pr(T \leqslant t \mid w = 0) = F_s(t), \\ &\Pr(T \leqslant t \mid w = 1) = 0. \end{aligned}$$

The cumulative distribution function (CDF) for the overall population, using $\Pr(w = 1) = p_w$, becomes

$$\begin{aligned} F(t) = \Pr(T \leqslant t) &= \Pr(T \leqslant t \mid w = 1)\Pr(w = 1) + \Pr(T \leqslant t \mid w = 0)\Pr(w = 0) \\ &= (1 - p_w)F_s(t). \end{aligned} \tag{1}$$

Equivalently, the survival function for the overall population can be written as

$$S(t) = p_w + (1 - p_w)S_s(t). \tag{2}$$

# 4 Link Functions

In the context of cure rate models, the binary component often relies on logistic regression for modeling the probability of being cured or permanently non-vulnerable to the event of interest. Consider a vector $\boldsymbol{w} = (w_1, w_2, ..., w_n)'$ consisting of n independent dichotomous random variables with probability $p_{w_i} = \Pr(w_i = 1|\boldsymbol{x}_i)$. $\boldsymbol{x}_i = (x_{i1}, ..., x_{ik})'$ denote a $k \times 1$ vector of covariates for the $i$th subject, and $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_k)'$ denotes a $k \times 1$ vector of regression coefficients. Then the probability $p_{w_i}$, which is the probability of the $i$th individual to be insusceptible to the event, is predicted by applying a link function to a linear combination of predictor variables.

The logit link function is a standard choice in logistic regression, serving to connect the linear predictor to to probability of the binary response. However, it is not entirely accurate to claim that the logit link function is suitable for all types of data in the context of binary response models. While the logit link is indeed one of the most commonly used link in logistic regression, it is important to recognize that it is a symmetric link function. The symmetry implies that it assumes an equal rate of change in the response variable with respect to the predictor variable, which may not always be the case in real world scenarios. The relationships between predictors and binary outcomes can be asymmetric and even exhibit varying degrees of nonlinearity. In these situations, link functions based on the inverse of non-standard cdfs can better accommodate the asymmetry and nonlinearity present in the data.

## 4.1 Logit Link

The inverse of the logistic function, or the logit function transforms probabilities into their corresponding log-odds. Then the probability $p_{w_i}$ of the binary response variable $w_i = 1$ is expressed as

$$p_{w_i} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}, \quad i = 1, \ldots, n, \tag{3}$$

and the logit link is represented as

$$\text{logit}(p_{w_i}) = \log\left(\frac{p_{w_i}}{1 - p_{w_i}}\right) = \boldsymbol{x}_i'\boldsymbol{\gamma}. \tag{4}$$

## 4.2 Skewed and Reversed Power Logit Link

While the standard logit link function is symmetric there are instances where asymmetry in the relationship between the predictors and the response variable exists in real life. This is where the use of an asymmetric link distribution is required. The asymmetric link distributions offer greater flexibility in modeling complex relationships between predictors and binary outcomes, while effectively capturing skewness. Couple of the well-known distributions to derive the asymmetric link functions are derived from the Type I and II generalized logistic distribution (Bazán et al., 2017; Nagler, 1994).

The Type I generalized logistic distribution is also called exponentiated logistic distribution or the skewed logistic distribution. It adds an additional parameter $\alpha_s$ to the definition of the distribution. Simplifying the formula with $u_i = \boldsymbol{x}_i'\boldsymbol{\gamma}$, we have

$$F_{sl}(u_i) = \frac{1}{1 + \exp(-u_i)^{\alpha_s}}, \qquad -\infty < u_i < \infty,\ \alpha_s > 0, \tag{5}$$

and the corresponding probability density function (pdf) is given by

$$f_{sl}(u_i) = \frac{\alpha_s \exp(-u_i)}{1+\exp(-u_i)^{\alpha_s+1}}, \qquad -\infty < u_i < \infty,\ \alpha_s > 0. \tag{6}$$

This distribution is used to generate an alternative estimator to the logit function, and is referred to as the skewed-logit, because it allows for a skewed response curve, with $\alpha$ serving as a parameter to measure skewness. Compared to the standard logit function, the shape of the curve changes as the value of $\alpha_s$ varies. For instance, for $\alpha_s < 1$, the curve becomes left skewed. The link function, denoted by slogit, is represented as

$$\text{slogit}(p_{w_i}) = \log\left\{(1-p_{w_i}^{1/\alpha_s})^{-1} - 1\right\} = \boldsymbol{x}_i'\boldsymbol{\gamma}. \tag{7}$$

Another alternative to the logit function is derived in a similar direction. Under Type II generalized logistic distribution, with an additional parameter $\lambda_{rp}$, giving the reversed exponentiated cdf form as

$$F_{rp}(u_i) = 1 - \left\{1 - \frac{1}{1+\exp(-u_i)}\right\}^{\lambda_{rp}}, \qquad -\infty < u_i < \infty,\ \lambda_{rp} > 0, \tag{8}$$

with the corresponding pdf

$$f_{rp}(u_i) = \frac{\lambda_{rp}\exp(u_i)}{\{1+\exp(u_i)\}^2}\left\{\frac{1}{1+\exp(u_i)}\right\}^{\lambda_{rp}-1}, \qquad -\infty < u_i < \infty,\ \lambda_{rp} > 0. \tag{9}$$

Finally the link function denoted by rplogit is represented by

$$\text{rplogit}(p_{w_i}) = \log\left\{(1-p_{w_i})^{-1/\lambda_{rp}} - 1\right\} = \boldsymbol{x}_i'\boldsymbol{\gamma}. \tag{10}$$

As the $\lambda_{rp}$ value varies, the shape of the curve changes with some skewness. A value of $\lambda_{rp} = 1$ recovers the standard logit, while values $\lambda_{rp} < 1$ stretch the curve, creating heavier tails.

### 4.3 Flexible Generalized Logit Link

The flexible generalized logit (fglogit) link function is derived from the exponentiated-exponential logistic (EEL) distribution (Ghosh and Alzaatreh, 2018), which itself is based on the exponentiated-exponential distribution introduced by Gupta and Kundu (2001). This distribution combines features of both the exponentiated-exponential and logistic distributions, resulting in a flexible family capable of modeling skewness and heavy tails. With $\mu \in \mathrm{R}$ and $s > 0$ as the location and scale parameters of logistic distribution and $\alpha_{fg} > 0$ and $\lambda_{fg} > 0$ as the shape and scale parameters, the pdf and cdf are given as

$$f_{fg}(u_i) = \frac{\alpha_{fg}\lambda_{fg}\exp\left(u_i-\mu\right)/s}{s\{1+\exp\left(u_i-\mu\right)/s\}^{\lambda_{fg}+1}}\left[1-\{1+\exp\left(u_i-\mu\right)/s\}^{-\lambda_{fg}}\right]^{\alpha_{fg}-1}, \tag{11}$$

$$F_{fg}(u_i) = [1-\{1+\exp(u_i-\mu)/s\}^{-\lambda_{fg}}]^{\alpha_{fg}}. \tag{12}$$

When $\mu = 0$ and $s = 1$, specific values of $\lambda_{fg}$ and $\alpha_{fg}$ lead to well-known distributions: when $\lambda_{fg} = \alpha_{fg} = 1$, the distribution reduces to that of the logistic distribution, while $\lambda_{fg} = 1$ results in the Type I Generalized logistic distribution, and $\alpha_{fg} = 1$ yields the Type II Generalized logistic distribution. Additionally, depending on the inequality of the parameters, the direction

of skewness differs. The distribution exhibits negative skewness for $\alpha_{fg} < \lambda_{fg}$, positive skewness for $\lambda_{fg} < \alpha_{fg}$, and symmetry for $\alpha_{fg} = \lambda_{fg}$.

Utilizing the inverse of the standard EEL cdf, where $\mu = 0$, and $s = 1$, a flexible generalized logit function, denoted by fglogit, is formed (Prasetyo et al., 2020) by

$$\text{fglogit}(p_{w_i}) = \log\left\{(1 - p_{w_i}^{1/\alpha_{fg}})^{-1/\lambda_{fg}} - 1\right\} = \boldsymbol{x}_i'\boldsymbol{\gamma}. \tag{13}$$

For the following sections on simulation and real data application, we employ a mixture cure rate model, which combines three accelerated failure time models (AFT): Weibull (WB), lognormal (LN), and loglogistic (LLG), for the latency component with logistic regression models using different link functions for the incidence component: Logit (L), skewed logit (S), reversed power logit (RP), and flexible generalized logit (FG). This result in 12 different models, where each model is denoted by the corresponding combination of acronyms, for example, Weibull with reversed power logit link is represented as WBRP.

# 5 Bayesian inference

The proposed model combines the elements of the accelerated failure time model and logistic regression to account for both the latency and incidence components of survival data. Let $\boldsymbol{D}_{\text{obs}} = (\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n)^T = (\boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{X}, \boldsymbol{Z})$ be the observed data. The model consists of two components: incidence and latency. The incidence component is governed by the parameter vector $\boldsymbol{\nu} = (\boldsymbol{\gamma}^T, \boldsymbol{\eta}^T)^T$, where $\boldsymbol{\gamma}$ represents the regression coefficients for the incidence model, and $\boldsymbol{\eta} = (\alpha_s, \lambda_{rp}, \alpha_{fg}, \lambda_{fg})^T$ includes additional parameters associated with the chosen link function. The latency component is characterized by the parameter vector $\boldsymbol{\upsilon} = (\boldsymbol{\beta}^T, \boldsymbol{\sigma}^T)^T$, where $\boldsymbol{\beta}$ denotes the regression coefficients for the latency model, and $\boldsymbol{\sigma} = (\sigma_{wb}, \sigma_{ln}, \sigma_{llg})^T$ contains parameters for the latency distribution. The survival function for the uncured individuals can be express as $S_s(y_i|\boldsymbol{z}_i;\boldsymbol{\upsilon}) = S_s(y_i; \boldsymbol{z}_i^T\boldsymbol{\beta}, \boldsymbol{\sigma})$, where $S_s(\cdot)$ denotes the survival function corresponding to the chosen latency distribution and $f_s(\cdot)$ denotes the corresponding pdf. Let $p_{w_i}(\boldsymbol{x}_i;\boldsymbol{\nu}) = \Pr(w_i = 1 \mid \boldsymbol{x}_i;\boldsymbol{\nu})$. The corresponding likelihood function with the observed data is

$$L_{\text{obs}} = \Pi_{i=1}^{n}\{p_{w_i}(\boldsymbol{x}_i;\boldsymbol{\nu})f_s(y_i|\boldsymbol{z}_i;\boldsymbol{\upsilon})\}^{\delta_i}\{1 - p_{w_i}(\boldsymbol{x}_i;\boldsymbol{\nu}) + p_{w_i}(\boldsymbol{x}_i;\boldsymbol{\nu})S_s(y_i|\boldsymbol{z}_i;\boldsymbol{\upsilon})\}^{(1-\delta_i)}. \tag{14}$$

## 5.1 Model assessment

Model performance is evaluated using three Bayesian model comparison criteria: Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002), Watanabe-Akaike Information Criterion (WAIC) (Watanabe, 2010), and Leave-One-Out Information Criterion (LOOIC) (Vehtari et al., 2017). These criteria balance model fit and complexity, providing insight into the predictive accuracy and generalizability of the models.

The DIC assesses model fit by penalizing model complexity, making it suitable for comparing models with different numbers of parameters. It is computed as

$$\text{DIC} = \text{Dev}(\overline{\boldsymbol{\Theta}}) + 2p_D,$$

where $\boldsymbol{\Theta}$ represents the collection of all model parameters, $\text{Dev}(\boldsymbol{\Theta}) = -2\log L(\boldsymbol{\Theta}|\boldsymbol{D}_{\text{obs}})$ is the deviance function, and $p_D = \overline{\text{Dev}}(\boldsymbol{\Theta}) - \text{Dev}(\overline{\boldsymbol{\Theta}})$ represents the effective number of parameters.

The WAIC is a fully Bayesian alternative to DIC, which asymptotically approximates Bayesian cross-validation. It is computed as

$$\text{WAIC} = -2\sum_{i=1}^{n} \log \mathbb{E}_{\boldsymbol{\Theta}}[p(y_i|\boldsymbol{\Theta})] + 2p_{\text{WAIC}},$$

where the first term represents the log pointwise predictive density (lpd) averaged over the posterior, and $p_{\text{WAIC}} = \sum_{i=1}^{n} \text{Var}_{\boldsymbol{\Theta}}(\log p(y_i|\boldsymbol{\Theta}))$ accounts for model complexity by estimating the effective number of parameters.

The LOOIC evaluates out-of-sample predictive performance using leave-one-out cross-validation (LOO-CV). It estimates the expected log predictive density for new observations by approximating

$$\text{LOOIC} = -2\sum_{i=1}^{n} \log p(y_i|y_{-i}),$$

where $p(y_i|y_{-i})$ represents the leave-one-out predictive density for observation $i$. A practical approximation is given by

$$\text{LOOIC} = -2\sum_{i=1}^{n} \log \mathbb{E}_{\boldsymbol{\Theta}}[p(y_i|\boldsymbol{\Theta})] + 2p_{\text{LOOIC}},$$

where $p_{\text{LOOIC}}$ is an estimate of the effective number of parameters. Since LOOIC is computed directly from posterior samples, it provides a more robust assessment of model generalizability than DIC.

These model selection criteria are used to compare different models and determine the best fit for the given data, where smaller values indicate better model performance.

# 6 Simulation Study

## 6.1 Simulation setting

In this simulation study, we assessed the performance of various Bayesian mixture cure rate models under different link function settings for the cure fraction. Specifically, we evaluated models using the four links, logit (L), slogit (S), rplogit (RP), and fglogit (FG), combined with three different parametric latency distributions: Weibull (WB), lognormal (LN), and loglogistic (LLG), resulting in a total of 12 model configurations.

We adopted a Bayesian estimation framework, specifying independent priors for all model parameters to allow flexibility while ensuring regularization. For logistic regression coefficient vector $\boldsymbol{\gamma}$, we used weakly informative normal priors to provide reasonable regularization without imposing strong assumptions. Specifically, each coefficient follows $\gamma_j \sim N(0,1),\ j = 1,\ldots,m$. For the link function parameters of slogit, rplogit, and fglogit, we assigned lognormal priors, such that, $\alpha_s, \lambda_{rp}, \alpha_{fg}, \lambda_{fg} \sim \text{Lognormal}(0,1)$. These priors reflect the assumption that the shape parameters are likely near 1 while allowing flexibility in their estimation. For the latency component of the cure rate model, the scale parameter $\sigma$ follows a Cauchy prior to accommodate heavy-tailed behavior, with $\sigma_{wb}, \sigma_{ln}, \sigma_{llg} \sim \text{Cauchy}(0, 2.5)$. The regression coefficient $\boldsymbol{\beta}$, which model the survival component, follow the same weakly informative normal prior as $\gamma$, which is $\beta_k \sim N(0,1),\ k = 1,\ldots,p$.

To evaluate model performance, we conducted 1,000 replications for each of the 12 models under two sample sizes: $n = 500$ and $n = 1000$. For each simulated dataset, covariates for the latency include an intercept and two numeric data generated from the standard normal distribution, while the incidence component include two scaled covariates generated from a uniform distribution $U(-1, 1)$, without an intercept. For true parameter values, we have $\boldsymbol{\beta} = (1, -1, -2)^T, \boldsymbol{\gamma} = (0.5, 0.6)^T, \boldsymbol{\sigma} = (1.5, 1.5, 1.5)^T, \boldsymbol{\eta} = (1.5, 1.5, 0.5, 1.5)^T$. Event times for uncured individuals were simulated based on the specified latency distribution, while cure status was drawn from a Bernoulli distribution according to the calculated cure probabilities. Random uniform censoring times with Uniform(0, 100) were applied to simulate right-censoring, and individuals identified as cured were automatically assigned censored status. The simulation study focused on three key evaluation aspects: coverage of 95% credible intervals for parameter recovery, effectiveness of model selection criteria (WAIC, DIC, LOOIC) in identifying the true model configuration, and goodness-of-fit diagnostics. Convergence diagnostics indicated reliable estimation across all simulations, with all R-hat values close to 1 and no divergent transitions. In this paper, only the Weibull model plots are presented, while analogous results for lognormal and loglogistic latency models are included in the supplementary materials.

## 6.2 Simulation result

**True value recovery**: Figure 2 presents the posterior mean estimates and 95% credible intervals for key model parameters under the Weibull latency model across the four link functions: logit, slogit, rplogit, and fglogit. Across all link functions and both sample sizes, the estimated posterior means closely align with the true values, demonstrating strong point estimate accuracy.

The credible intervals are generally narrow, especially for logit, slogit, and rplogit links, indicating high estimation precision. As expected, intervals widen under the smaller sample size ($n = 500$), but the true values remain well-captured within the intervals. The fglogit link, while producing slightly wider intervals for its incidence section parameters, still yields accurate posterior means and acceptable uncertainty bounds. This increased interval width is expected due to the higher flexibility and parameterization of the fglogit link function, which introduces greater variability, but retains good coverage performance. Notably, while not included in this paper, even under a more informative prior setting with reduced variance for $\lambda_{fg}$ and $\alpha_{fg}$, the fglogit model continues to exhibit reliable performance, producing stable posterior estimates and maintaining acceptable interval coverage.

The coverage rates in table 4 for all parameters are close to the nominal 95% level for both sample sizes, with slightly conservative coverage for fglogit parameters. These results confirm the accuracy and reliability of the proposed Bayesian mixture cure models under finite sample sizes.
**Model assessment validation**: To evaluate the ability of model selection criteria to identify the true model, we conducted a simulation study with $n = 1000$, using LOOIC, WAIC, and DIC as comparison metrics. For each of the 12 true model configurations and 100 replications, we fit a set of competing models that included the correct and incorrect combinations of distribution and link function, specifically comparing the proposed asymmetric link functions (slogit, rplogit, and fglogit) to the standard logit link across different latency forms.

Table 4 summarizes how often each criterion selected the true model out of 100 comparing the suggested link function models to the standard logit link with different distributions. The selection rates for Weibull-based models were very high across all criteria, often reaching 100 out of 100. However, for lognormal and loglogistic structures, selection rates were slightly lower. This decline can be attributed to the similar shape and fit of the competing distributions under

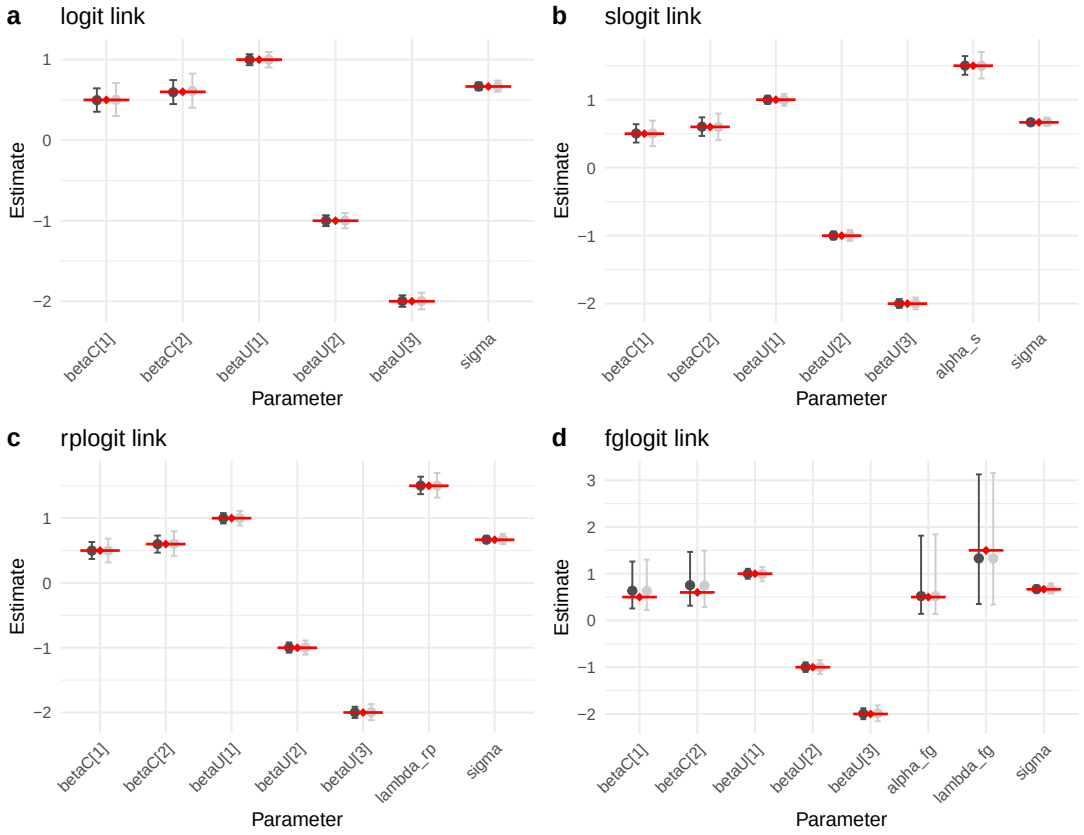



Figure 2: Posterior mean estimates and 95% credible intervals under the Weibull latency model. Black lines represent $n = 1000$, gray lines represent $n = 500$, and the red horizontal line indicates the true parameter value.

the specified parameter values. However, even in the presence of mild misspecification of the latency distribution, the model comparison criteria consistently favored the true link function over the standard logit link. These results highlight the robustness of criteria and their strong overall ability to recover the correct link function for the incidence model, even when the survival component is not perfectly specified.

**Goodness-of-fit evaluation**: To assess the goodness-of-fit of the proposed models, we conducted a Cox-Snell residual analysis (Scolas et al., 2016). The Cox-Snell residuals,

$$r_{CS}(y_i) = -\log\{\widehat{S}(y_i)\} = -\log\{1 - \hat{p}_{w_i} + \hat{p}_{w_i}\widehat{S}_s(y_i|\boldsymbol{z}_i)\}, \quad i = 1, \ldots, n\,, \tag{15}$$

where $\hat{p}_{w_i} = p_{w_i}(\boldsymbol{x}_i; \hat{\nu})$ is the estimated cure probability, and $\hat{S}_s(y_i|\boldsymbol{z}_i)$ is the estimated survival function for the susceptible group. Under a good fit model, these residuals should follow an exponential distribution with rate 1, and thus, plotting the Kaplan-Meier estimate of $r_{CS}(y_i)$ against the exponential(1) reference provides a visual diagnostic for model fit.

We generated event times from the true Weibull model with sample size of $n = 1000$, paired with each of the four link functions and then evaluated model fit by examining the Cox-Snell residuals. For each data generating mechanism, we fitted the model fit using all four link functions, allowing direct comparison between the true model and alternatives. As shown in Figures 3 and 4, when the data were generated using the logit link, the simpler log model as well

Table 4: Coverage probabilities for each parameter for Weibull latency across difference link functions, and sample sizes.

| N | Distribution | Parameter | Logit | SLogit | RPLogit | FGLogit |
|---|---|---|---|---|---|---|
| 500 | Weibull | $\beta_C[1]$ | 0.939 | 0.951 | 0.937 | 1.000 |
| | | $\beta_C[2]$ | 0.941 | 0.951 | 0.944 | 1.000 |
| | | $\beta_U[1]$ | 0.945 | 0.951 | 0.945 | 0.934 |
| | | $\beta_U[2]$ | 0.948 | 0.940 | 0.948 | 0.951 |
| | | $\beta_U[3]$ | 0.944 | 0.953 | 0.947 | 0.961 |
| | | $\alpha$ | – | 0.932 | – | 1.000 |
| | | $\lambda$ | – | – | 0.934 | 1.000 |
| | | $\sigma$ | 0.950 | 0.945 | 0.960 | 0.941 |
| 1000 | Weibull | $\beta_C[1]$ | 0.943 | 0.938 | 0.945 | 1.000 |
| | | $\beta_C[2]$ | 0.950 | 0.946 | 0.934 | 1.000 |
| | | $\beta_U[1]$ | 0.936 | 0.946 | 0.947 | 0.950 |
| | | $\beta_U[2]$ | 0.936 | 0.935 | 0.951 | 0.950 |
| | | $\beta_U[3]$ | 0.934 | 0.939 | 0.940 | 0.934 |
| | | $\alpha$ | – | 0.944 | – | 1.000 |
| | | $\lambda$ | – | – | 0.946 | 1.000 |
| | | $\sigma$ | 0.947 | 0.951 | 0.950 | 0.950 |

Table 5: Number of times each model was selected under different true models with Weibull latency

| | Criterion | True slogit | True rplogit | True fglogit |
|---|---|---|---|---|
| Weibull | LOOIC | 100 | 100 | 99 |
| | WAIC | 100 | 100 | 99 |
| | DIC | 100 | 100 | 98 |

as the proposed asymmetric links provided comparable fits. However, when data were generated under asymmetric links, the standard logit link model showed systematic deviation from the reference exponential(1) curve, indicating a lack of fit. In contrast, models using the correct asymmetric link captured the Cox-Snell residual behavior more accurately, with the estimated survival curves aligning well within the posterior predictive envelope. These results put emphasis the robustness of the proposed flexible link models, offering clear advantages when the incidence structure is skewed. The plots presented in this paper correspond to a single simulation replicate from the Weibull case with $n = 1000$. However, similar patterns were consistently observed across other replications. Additional Cox-Snell residual plots with lognormal and loglogistic latency distributions and for models fitted to data generated under their own true specifications, are included in the supplementary materials.

# 7 Application to Iowa recidivism data

We revisit the Iowa 3-year recidivism dataset introduced in Section 2, focusing on evaluating recidivism risk through the Bayesian mixture cure rate framework. The objective is to assess how flexible link functions perform in capturing the cure proportion and time-to-recidivism among formerly incarcerated individuals. Covariates used in the analysis include demographic and institutional variables likely to influence recidivism outcomes. The incidence model included age,

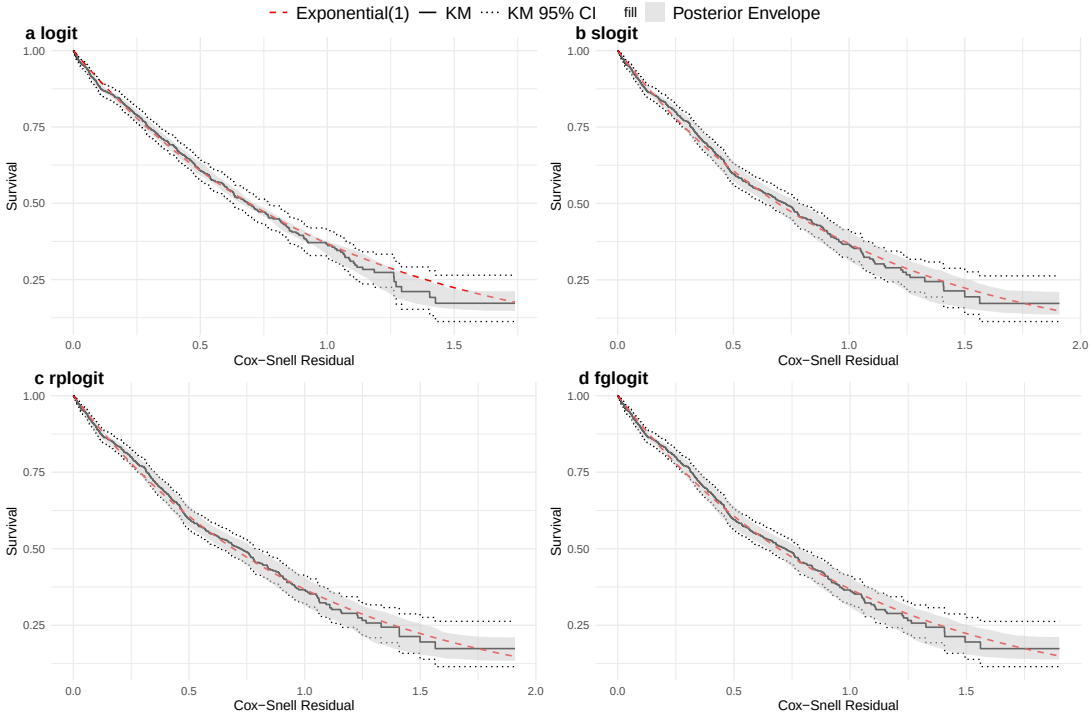



Figure 3: Cox-Snell residual plot with Weibull logit link being true model

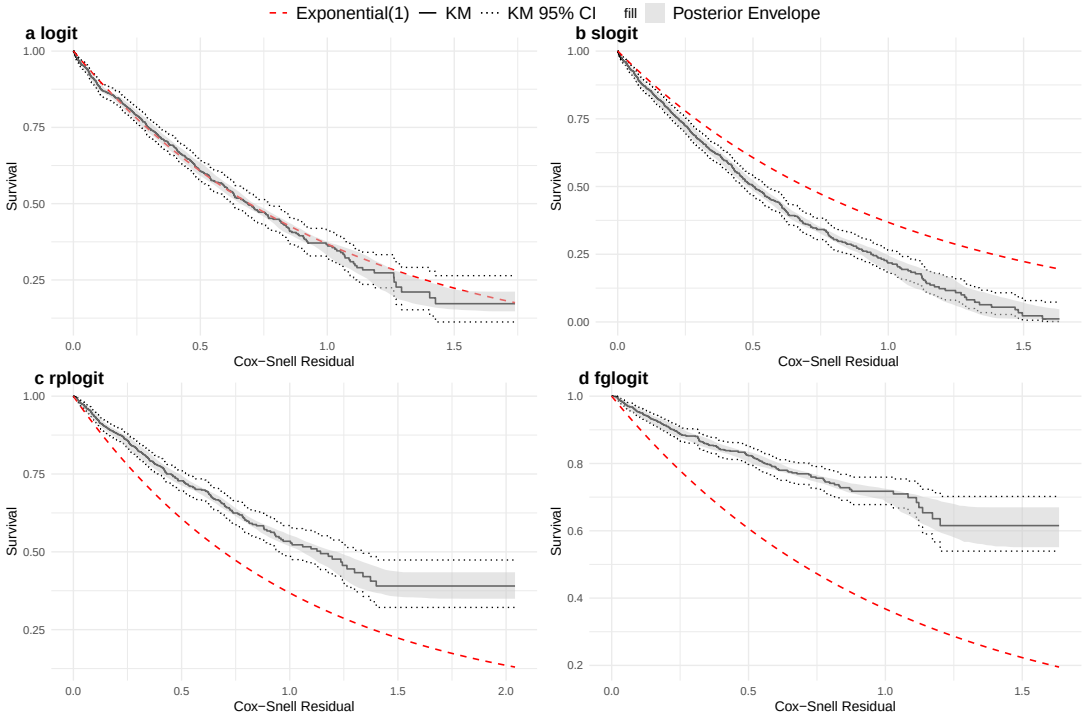



Figure 4: Cox-Snell residual plot with Weibull logit link fitted to other true link models

Table 6: Bayesian assessment values with smaller values indicating better fit.

| LOOIC | Weibull | Lognormal | Loglogistic |
|---|---|---|---|
| logit | 32464.5 | 32498.6 | 32439.1 |
| slogit | 32466.8 | 32499.1 | 32438.8 |
| rplogit | **32464.3** | 32495.8 | **32438.3** |
| fglogit | 32467.8 | **32494.8** | 32440.2 |
| WAIC | Weibull | Lognormal | Loglogistic |
| logit | **32464.0** | 32498.0 | **32439.0** |
| slogit | 32466.5 | 32498.2 | 32439.3 |
| rplogit | 32464.3 | 32494.6 | 32439.1 |
| fglogit | 32468.7 | **32493.7** | 32441.3 |
| DIC | Weibull | Lognormal | Loglogistic |
| logit | **32403.8** | 32434.7 | 32378.8 |
| slogit | 32405.8 | 32434.4 | 32378.7 |
| rplogit | 32404.1 | **32432.7** | **32378.0** |
| fglogit | 32407.1 | 32432.8 | 32379.7 |

months under supervision, race, sex, supervising unit, supervision type, supervision end reason, and supervision offense type, as well as the individual's risk ranking. The latency model used the same set of covariates except for risk ranking, which was excluded under the rationale that it more directly reflects the overall risk of recidivism rather than influencing the timing of the event. Additionally, only high-level categorical groupings were used, such as the offense type, omitting finer subcategories to support the interpretability.

We apply the twelve mixture cure rate models, the combinations of three AFT models (Weibull (WB), Lognormal (LN), and Loglogistic(LLG)) with four link functions (logit (L), slogit (S), rplogit (RP), and fglogit(FG)), to predict the short-term recidivism. Using stan, each model configuration was sampled across four independent chains with 2,000 iterations per chain, half of which were allocated to warm-up phases. The convergence of these models was rigorously evaluated through standard diagnostic checks including trace plots and assessments of the potential scale reduction factor (R-hat). All models achieved satisfactory convergence as indicated by R-hat values close to 1.0, and there were no divergent transitions nor any instances where the maximum tree depth was saturated, indicating reliable posterior estimation.

## 7.1 Comparison

Table 6 compares model fit across the 12 models using LOOIC, WAIC, and DIC, where smaller values indicate better fit. The LGRP model emerges as the top-performing model. Within each AFT family, WBL and LNFG models achieved the best assessment values, though they were consistently outperformed by the LLGRP. The assessment differences between LLGRP and LLGL are modest, suggesting that the data structure is reasonably well captured even by the simpler logit link within the loglogistic family. Nonetheless, we proceed with inference based on the LLGRP model due to its greater flexibility and slighter superior overall performance. Importantly, the covariates identified as significant in the LLGRP model were also found to be significant in most other model specifications, reinforcing the robustness and consistency of the findings.

To assess goodness-of-fit of the model, we evaluated the loglogistic mixture cure rate models

using posterior predictive p-values (PPP) derived based on Cox-Snell residuals from equation 15 (Gelman et al., 1996). We construct the PPP as

$$p_{\text{ppp}} = \Pr[T(r^{\text{rep}} \geqslant T(r^{\text{obs}})|\boldsymbol{y}],$$

where $T(\cdot)$ is a discrepancy measure, the integrated squared error between the empirical survival function of the residuals and the Exponential(1) survival curve, $r^{\text{rep}}$ are the residuals computed from the posterior predictive replicates, and $r^{\text{obs}}$ are the residuals from the observed data. A PPP close to 0.5 indicates adequate fit, while values close to 0 or 1 (under 0.05, or above 0.95) suggest model misfit. Among the loglogistic models, LLGRP yielded a PPP of 0.257, which, while slightly lower than the ideal value of 0.5, still suggests an acceptable fit, indicating acceptable agreement between model and data. With the same loglogistic latency structure, LLGL, LLGS, and LLGFG produced PPPs ranging from 0.278 to 0.384, all within a reasonable range. The Cox-Snell residual plot for the LLGRP model in figure 5 further supports this, with the empirical Kaplan-Meier curve closely tracking the reference curve and remains well within the 95% posterior predictive envelope, demonstrating a overall fit and supporting the suitability of the loglogistic distribution in capturing recidivism dynamics.

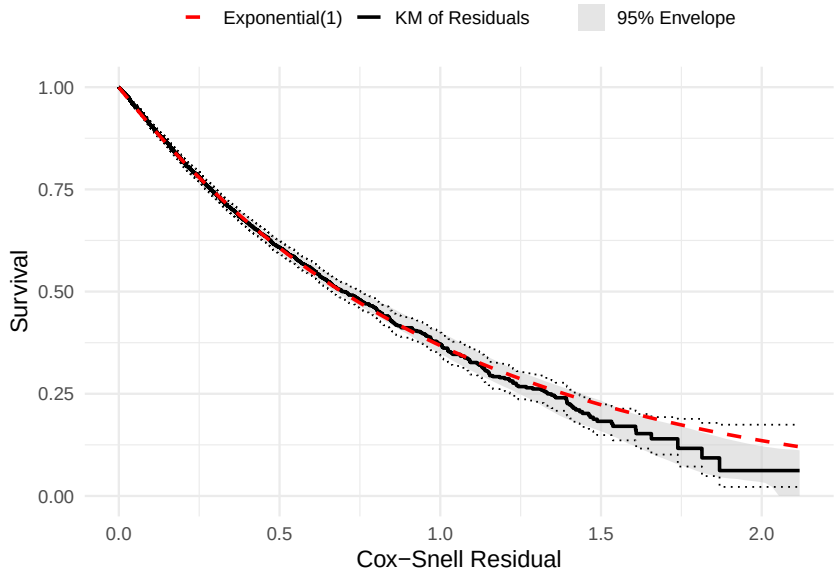



Figure 5: Kaplan Meier Cox-Snell residual plot with LLGRP

## 7.2 Inference

Table 7 highlights the parameter estimates of LLGRP, but similar patterns of statistical significance and effect direction were observed across all model specifications. The persistence of key covariates as significant predictors across varying model structures underscores the robustness of the findings and supports the reliability and interpretability of the proposed framework across a range of plausible modeling scenarios.

The reference categories for the categorical variables were set at the first level in table 2. In the latency component of the model, age was positively associates with time to recidivism across

all specifications, suggesting that older individuals tended to remain out of prison for longer durations after release. Race also exhibited significant effects, suggesting that compared to the baseline American Indian or Alaska Native race group, Hispanic and White individuals showed extended survival times, indicating racial disparities in post release outcomes. Facility specific effects were notable as well where it indicates those released from the Clarinda Correctional Facility, Iowa Medical Classification Center, and the Iowa State Penitentiary had significantly longer times to re-offend compared to the baseline Anamosa State Penitentiary, potentially reflecting institutional differences in release practices or rehabilitation programs.

In terms of the incidence component, covariates were interpreted in terms of their effect on the probability of being cured. A positive coefficient indicates a higher log-odds of being cured. Several supervision related variables stood out. Months under supervision were positively associates with the probability of being cured, suggesting that extended supervision periods may lead to lower underlying risk. Among supervision units, individuals released from the Iowa State Penitentiary, were also significantly more likely to be cured, suggesting institution-specific factors that may support long-term resistance from crime. Supervision end reason of being paroled to an INS detainer was associated with reduced cure probability, where the baseline is being discharged due to expiration of sentence. Regarding offense type, compared to drug offense, individuals with other offenses were significantly more likely to be uncured, while public order and violent offenses were associated with higher cure probabilities, which may reflect a more structured release processes in these groups.

While the shape parameter $\lambda_{rp}$ in the incidence model has a posterior mean of 0.623, its 95% interval includes 1, suggesting no strong evidence of deviation from the standard logit link. Additionally, model comparison criteria such as DIC, WAIC, and LOOIC show only marginal differences between the flexible link models and the standard logit within the same latency distribution. This indicates that the data structure in the Iowa recidivism cohort may be relatively well-aligned with the symmetry assumption of the logit link. However, it is important to note that this result does not diminish the value of the flexible link functions. In datasets where the cure fraction is more skewed or where group membership is more imbalanced, the asymmetric link models may provide clear advantages. Overall, the flexible link models performed comparably well, and given that the logit link is nested within these more general forms, they offer a robust and adaptable framework for incidence modeling in a wide range of applications.

# 8 Conclusion

This study explored Bayesian mixture cure rate models with flexible link functions to analyze recidivism outcomes in Iowa. We assessed 12 models, which combined three AFT models: Weibull, lognormal, and loglogistic, and four link functions: logit, skewed logit, reversed power logit, and flexible generalized logit link. Simulations confirmed the robustness and flexibility of the proposed models, with all methods effectively recovering true parameters and demonstrating strong coverage properties. While the logit link performed well under symmetric assumptions, the flexible asymmetric links offered improved fit when the data exhibited skewness in the cure fraction.

In the real data application of Iowa three year recidivism dataset, the best performing model based on LOOIC, WAIC, and DIC was the loglogistic with reversed power logit link (LLGRP). The model fit was also supported by the Cox-Snell residual plot with posterior predictive p-values. Although the improvemenet over the standard logit model were modest, the LLGRP

Table 7: LLGRP model: Posterior means and 95% HPD intervals for latency and cure rate variables.

| **Parameter** | **Mean** | **HPD 95% Interval** |
|---|---|---|
| **Latency Variables** | | |
| Age | 4.34 | [3.319, 5.336] |
| *Race* | | |
| Hispanic | 0.545 | [0.228, 0.867] |
| White | 0.549 | [0.166, 0.920] |
| *Supervising Unit* | | |
| Clarinda Correctional Facility | 1.511 | [0.553, 2.527] |
| Iowa Corr. Institution for Women | 0.465 | [0.191, 0.733] |
| Iowa Medical Classification Center | 1.954 | [0.988, 2.963] |
| Iowa State Penitentiary | 0.512 | [0.260, 0.765] |
| Newton Correctional Facility | 0.389 | [0.122, 0.654] |
| *Supervision Type* | | |
| Work Release | 0.325 | [0.035, 0.604] |
| *Supervision End Reason* | | |
| Paroled to Detainer (INS) | −0.718 | [−0.885, −0.554] |
| Paroled to Detainer (Iowa) | 1.926 | [0.896, 3.148] |
| Paroled to Detainer (Out-of-State) | −1.106 | [−1.573, −0.631] |
| Immediate Discharge | 0.967 | [0.052, 1.700] |
| *Offense Type* | | |
| Other | −1.120 | [−1.357, −0.872] |
| Public Order | −0.146 | [−0.274, −0.018] |
| *Risk Ranking* | | |
| Moderate | 0.218 | [0.102, 0.334] |
| $\sigma_{\text{llg}}$ | 1.853 | [1.756, 1.954] |
| **Incidence Variables** | | |
| Months Supervised | 0.380 | [0.237, 0.564] |
| *Supervising Unit* | | |
| Iowa State Penitentiary | −2.557 | [−3.355, −1.912] |
| *Supervision End Reason* | | |
| Paroled to Detainer (INS) | −0.578 | [−0.960, −0.227] |
| *Offense Type* | | |
| Other | −1.021 | [−1.605, −0.484] |
| Public Order | 0.507 | [0.213, 0.863] |
| Violent | 0.952 | [0.507, 1.505] |
| $\lambda_{\text{rp}}$ | 0.623 | [0.362, 1.105] |

model offers greater adaptability and flexibility, particularly in settings with asymmetric cure structures. Across models, the key covariates such as race, months supervised, offense type, and supervision conditions were revealed to be significant, emphsizing the interplay between demographic factors and institutional effects on recidivism outcomes.

This work demonstrates the value of incorporating flexible link functions into the mixture cure rate model framework, offering insights into the factors driving both the likelihood and timing of recidivism. Future research could incorporate alternative latency structure. While parametric AFT models offer interpretability and efficiency, it may impose strict assumptions on the survival patterns. To address this, parametric mixture distributions or semiparametric AFT models, such as those using spline-based hazard functions, can be used to allow for greater modeling flexibility. Additionally, replacing the linear predictor in the incidence component with Gaussian process priors, as suggested by Li et al. (2016), could enhance modeling of nonlinear covariate effects while maintaining flexibility. This approach could lead to the development of more sophisticated models that better capture the complex dynamics of survival data with cure fraction.

## Supplementary Material

The code supplement is available as a compressed folder. It is written in R and readily allows to replicate of the simulation results and Iowa data application. Some additional simulation results are provided within the folder.

## Data availability

The data used in this paper are openly available in the Iowa government data source for the correctional system (Iowa Department of Corrections, 2024).

## References

Bazán JL, Torres-Avilés F, Suzuki AK, Louzada F (2017). Power and reversal power links for binary regressions: An application for motor insurance policyholders. *Applied Stochastic Models in Business and Industry*, 33(1): 22–34.

Berkson J, Gage RP (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259): 501–515.

Boag JW (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society: Series B (Methodological)*, 11(1): 15–53.

Chen MH, Dey DK, Shao QM (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, 94(447): 909–919.

De la Cruz R, Fuentes C, Padilla O (2022). A bayesian mixture cure rate model for estimating short-term and long-term recidivism. *Entropy*, 25(1): 56.

Farewell VT (1977). A model for a binary variable with time-censored observations. *Biometrika*, 64(1): 43–46.

Farewell VT (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38(3): 1041–1046.

Gelman A, Meng XL, Stern H (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4): 733–760.

Ghosh I, Alzaatreh A (2018). A new class of generalized logistic distribution. *Communications in Statistics - Theory and Methods*, 47(14): 3459–3473.

Gupta RD, Kundu D (2001). Exponentiated exponential family: An alternative to gamma and weibull distributions. *Biometrical Journal*, 43: 117–130.

Iowa Department of Corrections (2024). Iowa prison recidivism and change by cohort. Accessed: January 2024. Available at https://data.iowa.gov/Correctional-System/Iowa-Prison-Recidivism-and-Change-by-Cohort/dnzw-paxg/about_data.

Kim S, Chen MH, Dey DK (2008). Flexible generalized t-link models for binary response data. *Biometrika*, 95(3): 535–548.

Kuk AY, Chen C (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79: 531–541.

Li CS, Taylor JMG (2002). A semi-parametric accelerated failure time cure model. *Statistics in Medicine*, 21: 3235–3247.

Li D, Wang X, Dey DK (2016). A flexible cure rate model for spatially correlated survival data based on generalized extreme value distribution and gaussian process priors. *Biometrical Journal*, 58(5): 1178–1197.

Maller RA, Zhou X (1996). *Survival Analysis with Long-Term Survivors*. Wiley.

Maltz MD (1984). *Recidivism*. Academic Press, Florida.

Nagler J (1994). Scobit: An alternative estimator to logit and probit. *American Journal of Political Science*, 38(1): 230–255.

Partanen J (1969). On waiting time distributions. *Acta Sociologica*, 12(3): 132–143.

Prasetyo RB, Kuswanto H, Iriawan N, Ulama BSS (2020). Binomial regression models with a flexible generalized logit link function. *Symmetry*, 12(2): 221.

Schmidt P, Witte AD (1989). Predicting criminal recidivism using 'split population' survival time models. *Journal of Econometrics*, 40(1): 141–159.

Schmidt P, Witte AD (2012). *Predicting Recidivism Using Survival Models*. Springer Science and Business Media.

Scolas S, Legrand C, Oulhaj A, Ghouch A (2016). Diagnostic checks in mixture cure models with interval-censoring. *Statistical Methods in Medical Research*, 27: 2114–2131.

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64: 583–639.

Sy JP, Taylor JMG (2000). Estimation in a cox proportional hazards cure model. *Biometrics*, 56: 227–236.

Vehtari A, Gelman A, Gabry J (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5): 1413–1432.

Visher CA, Winterfield L, Coggeshall MB (2005). Ex-offender employment programs and recidivism: A meta-analysis. *Journal of Experimental Criminology*, 1(3): 295–316.

Watanabe S (2010). Asymptotic equivalence of bayes cross-validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11: 3571–3594.

Yamaguchi K (1992). Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of 'permanent employment' in japan. *Journal of the American Statistical Association*, 87: 284–292.

Zhang J, Peng Y (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine*, 26(16): 3157–3171.