MASTER'S DEGREE THESIS

Master in Statistics and Operations Research/Master in Engineering Mathematics and Computational Science UPC/CTH

Title: Estimating the probability of discharge among Covid-19 hospitalizations using cure models

Author: Ida Höglund Persson

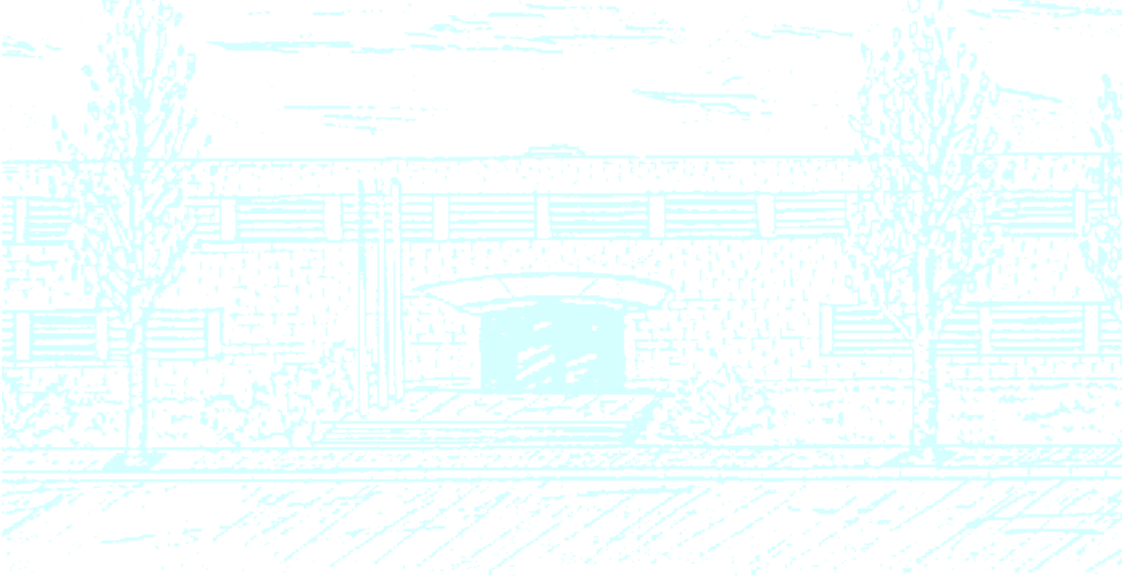
Advisor: Guadalupe Gómez Melis

Co-Advisor: Klaus Langohr

Department: Statistics and Operations Research

University: Universitat Politècnica de Catalunya/
Chalmers University of Technology

Academic year: 2022-2023



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

CHALMERS
UNIVERSITY OF TECHNOLOGY

Universitat Politécninca de Catalunya

Master's Thesis

Estimating the probability of discharge among Covid-19 hospitalizations using cure models

Author:
Ida Höglund Persson

Advisor:
Guadalupe Gómez Melis
Co-Advisor:
Klaus Langohr

Department of Statistics and Operations Research
2023

UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

CHALMERS
UNIVERSITY OF TECHNOLOGY

Abstract

Cure models have gained increased attention in recent years due to the advances within several disease treatments. This thesis focuses on cure models and their applicability in handling survival models with long-term survival due to immunity. A dataset including a large proportion of right-censored individuals at the follow-up time can be suspected to have individuals who will never experience the event of interest due to being cured (also expressed as being immune to the event).

The population is assumed to be divided into two subpopulations - one which is susceptible to the disease, and one which is immune to the event. In this study, the main objective is to estimate the proportion of cure, referred to as the incidence. Various cure models will be investigated and applied to a dataset comprising 2074 hospitalized Covid-19 patients from the metropolitan area of Barcelona during the first wave of the pandemic. The purpose of the study is to explore if there are some cure models suitable to model the behaviour of Covid-19 patients.

Mixture cure models allow the subpopulations to have different survival distributions while non-mixture cure models are easily interpreted due to similarities with proportional hazards models. The estimation of the survival of the uncured population, called the latency function, has previously mainly been done parametrically. However, recent research provides new complex and more accurate nonparametric methods that will be applied in this study. A comparison of the various estimation methods are provided with a discussion of the advantages and disadvantages of the approaches.

The estimation of the incidence function further provides an estimation of the immune proportion in our data set indicating the predicted percentage of hospitalized patients discharged from the hospitals. Due to the impossibility of distinguishing if a right-censored observation is uncured and not yet experienced the event or cured, a sufficient follow-up time and big data set is vital for the analysis.

Keywords: *cure models, latency, incidence, Covid-19*

Acknowledgements

I would like to express my sincere appreciation to my professors,Guadalupe Gómez Melis and Klaus Langohr, for their support and guidance during my Master's thesis. Their expertise, encouragement, and dedication have played a significant role in shaping the outcome of this research.

I am also grateful to the Department of Statistics and Operations Research at Universitat Politécninca de Catalunya (UPC) for giving me the opportunity to write my thesis within their institution. The resources, facilities, and collaborative environment provided by the department have greatly helped my research for this thesis.

Lastly, I want to express my appreciation to Chalmers University of Technology for providing me with an excellent education during my five years of study at Engineering Mathematics. The knowledge and skills I have gained at Chalmers have been crucial in completing this Master's thesis. I am also grateful to Chalmers for allowing me to write my thesis at UPC.

Ida Höglund Persson, Barcelona, May 2023

Contents

List of Figures

List of Tables

1

Introduction

With the world at the end of overcoming a major pandemic, a sense of relief is spread across the world. However, it is recognized that the potential occurrence of future pandemics is a matter of concern. A recent study in Italy (2021) assessed the annual probability of a new pandemic with a similar impact to Covid-19 to be approximately 2% [1]. Diseases transmitted from animals to humans have more than once been responsible for outbreaks, not least the Covid-19 pandemic. Factors such as an increased global travel, intrusion in natural habitats and the climate change can contribute to the emergence and spread of new diseases. Climate change is one of the most significant factors since it can force more interaction between humans and animals. This emphasizes the importance of being prepared and making proactive measurements to prevent as much damage as possible for our world.

One of the major challenges faced during the recent pandemic was the limited availability of hospital beds, which made it difficult to accommodate the high number of patients requiring hospitalization. In an eventual future similar situation it would therefore be facilitated to be well educated with statistical models which can prepare for a vital situation like this. *How can statistical models derived from Covid-19 data provide valuable knowledge to prevent shortages of critical resources, such as hospital beds in potential future pandemics or similar situations?* This Master's thesis will explore statistical models with the objective to answer this question.

The area of cure models might, for some, gone unnoticed. However, due to the success in several medical research areas the patient outcomes have improved, which brings a new attention to cure models within survival studies. Their theory violates previous assumptions about the shape and behavior of the survival function and is instead visualized as an improper survival function. This thesis will examine whether cure models could be appropriate to model Covid-19 data and, if so, what are the most efficient and most accurate estimation methods. Finally, a conclusion concerning its usability in the future will be made.

2

Theory

Modeling diseases and survival is a significant area within the field of statistics, given its wide applicability in the development of medicines and other aspects of hospital logistics. It has been an objective to create models for diseases and conditions that take into account the potential for a cure within the population. This chapter will first introduce the survival analysis followed by a brief introduction of cure models and their different applications.

2.1 Survival analysis with right censoring

Survival analysis is a field within statistics with the objective to analyze the time it takes for an event of interest to occur, such as death, failure or recovery [2]. It is commonly used in medical research, engineering, social sciences, and other fields where the goal is to model the data with a survival function. This function describes the probability of an individual surviving up to a certain time point. The survival function is defined as

$$S(t) = P(T > t) = 1 - F(t), \quad t \geq 0,$$

where $F(t)$ is the cumulative distribution function of time t.

When working with survival data, it is crucial to consider the presence of censored observations, which is indicated by incompleteness in the data. Right censoring means that some individuals in the study have not experienced the event of interest by the end of the study period. This type of censoring is common in many fields, for instance in medical research where patients may be lost to follow-up or still alive at the end of the study. Right censoring complicates the analysis because the survival time of the censored individuals is unknown, and their contribution to the analysis is limited. However, with the right methods it is still possible to estimate the probability of an event occurring over time, taking into account both the observed and censored data.

To analyze survival data with right censoring, two commonly applied methods are the Kaplan-Meier estimator and the Cox proportional hazards model. The Kaplan-Meier estimator is a nonparametric method that estimates the survival function from the observed data and the proportion of censored individuals. The Cox proportional hazards model is a semiparametric method that models the relationship between the hazard function and the covariates, while accounting for censoring. Both methods are widely used and provide valuable information about the time-to-event data, such as the median survival time or the effect of covariates on the survival outcome.

Define Y as the survival time, a non-negative continuous or discrete variable. Further, let C denote the time to censoring, indicating the time where the individual will stop being followed given that the event has not yet occurred. For an individual i the survival time is defined as Y_i and the censoring time as C_i. We assume C_i to be independent from Y_i, which indicates a non-informative censoring. Furthermore, we define $T_i = \min\{Y_i, C_i\}$ and

$$\delta_i = \begin{cases} 1, & \text{if } Y_i \leq C_i \\ 0, & \text{if } Y_i > C_i \end{cases}, \tag{2.1}$$

where δ_i is the event indicator taking the value 1 if the individual experiences the event before the end of the study and 0 if the individual does not.

Having a study with a lot of right-censored observations might indicate a plateau in the survival curve after a certain amount of time. This makes us believe that some individuals will not experience the event at all. A proportion of the population are considered to be statistically cured implying that they instead have the same mortality rate as the non-sick population. To have a population with a proportion of cured/immune individuals might introduce bias in the result from traditional survival methods. Cure models have been developed to account for this and the different kinds of models will be explained in this report.

2.2 Mixture cure models

The first cure models were introduced by Boag in 1949 to model a proportion of cured individuals after cancer treatment [2]. The general idea of the mixture models is that the observed population is divided into two groups, one that will be susceptible for the disease and one that will be immune to the event of interest. In this context, being immune to the event and having been cured from the disease are synonymous. The survival function will have two parts, one representing the cured individuals and one for the uncured. The proportion of cured patients in the population is represented by $1 - \pi$ and referred to as the incidence component of the expression. The survival function for a mixture cure model is given by

$$S(t) = (1 - \pi) + \pi S_u(t), \tag{2.2}$$

where $S_u(t)$ is the survival function of the uncured population called the latency function. The estimation of the incidence and the latency will be discussed in the next chapter. An important benefit with the mixture model is that it allows the covariates to have different influence of each group.

An extension of the mixture cure models

An extension to the mixture cure model also takes into account the background survival of the population, which refers to the underlying distribution of the non-sick population [3]. Previous research on mixture cure models has mainly considered children to minimize the risk of dying from other causes than the disease in matter. Nevertheless, when wanting to apply a cure model to an adult population, with a higher risk of dying from non-disease related causes, an extension of the previous method has been presented. This investigates the relative survival which is the ratio between the observed survival (observed events) and the expected survival function for the given group. This provides a measure of the excess survival experienced by the diagnosed patients. From this ratio follows the expression for the overall survival:

$$S_o(t) = S^*(t)R(t),$$

where $S^*(t)$ is the background survival function, $R(t)$ is the relative survival and $S(t)$ is the survival function for the entire population. The distribution for the underlying population, $S^*(t)$, is found externally (life tables for the general population is found in the Human Mortality Database (HMD)) and chosen for the country and age that will be analyzed. The relative survival is expressed as $R(t) = (1-\pi)+\pi S_u(t)$, which is the survival in Equation (2.2), while $S^*(t)$ is the background function obtained from the mortality data base. Hence, the overall survival is derived by

$$S_o(t) = S^*(t)(\pi + (1-\pi)S_u(t)). \tag{2.3}$$

The estimation of the various components in Equation (2.3) can be performed using parametric, semiparametric or nonparametric methods, as well as for the mixture models that do not consider background survival.

2.3 Non-mixture cure models

To model the growth of cancer cells more accurately another approach was presented in [4]. Let N be the number of cancer cells each individual has after cancer treatment. A cured individual has $N = 0$ cancer cells while an uncured has $N > 0$. The number of cancer cells of the uncured can grow rapidly and if at least one of the N cancer cells produces a detectable cancer mass the individual will develop cancer. Define

$$T = \min\{\tilde{T}_1, ..., \tilde{T}_N\} = \tilde{T}_{(1)},$$

where $\tilde{T}_i$ are the i.i.d. latent event times, i.e. the activation times for cancer cells to develop a detectable cancer mass, and $\tilde{T}_{(1)}$ is the first order statistic of $\tilde{T}_1, ..., \tilde{T}_N$, also called the first-activation scheme suitable for tumour kinetics. There are several different schemes developed for this purpose involving different distributions of $\tilde{T}_i$ and T leading to different cure models. Considering the most common model, N follows a Poisson distribution with mean λ. If $\tilde{T}_i \sim F^H(t)$ is a proper cumulative distribution function, then the unconditional survival function of T is

$$\begin{aligned}
P(T > t) = S(t \mid \boldsymbol{z}) &= P(N = 0) + P\left(\tilde{T}_1 > t, \ldots, \ldots, \tilde{T}_N > t, N \geq 1\right) \\
&= e^{-\lambda} + \sum_{k=1}^{\infty} \frac{\left[S^H(t)\lambda\right]^k}{k!} e^{-\lambda} \\
&= e^{-\lambda F^H(t)} = p^{F^H(t)},
\end{aligned}$$

and $S^H(t) = 1 - F^H(t)$. In this expression $p = e^{-\lambda}$ is the probability of being cured since $\lim_{t \to \infty} S^H(t) = p = e^{-\lambda}$, which implies that $S^H(t)$ is an improper survival function. Furthermore, the hazard function of this non-mixture model is given by $h(t|z) = -\ln(p) f^H(t)$ where $f^H(t) = dF^H(t)/dt$.

If the parameters in $f^H(t)$ do not vary by covariates then the expression for the hazard function is a proportional hazards model, which is another advantage of a non-mixture cure model. This model also has an extension incorporating the background survival and works similarly as the mixture cure model extension using the same calculations. The overall survival function is given by

$$S_o(t) = S^*(t) p^{F^H(t)}.$$

Furthermore, in case of no varying parameters this is a proportional hazards model. To be able to use these cure models the latency function has to be estimated which can be done parametrically or nonparametrically and will be introduced below.

2.4 Estimation methods for incidence and latency

There are various ways to estimate the two parts of the cure models which are based on different assumptions. The parametric estimation assumes that the baseline distribution of the time to event in the uncured population is known. This distribution is usually assumed to be exponential, Weibull or log-normal. In addition, the incidence is modeled as a binary regression model for the cured population which is assumed to follow a parametric distribution such as logistic regression or probit regression. Parametric estimation requires the assumption of a specific distribution and subsequently the estimation of the parameters of the distribution. Parametric methods can provide very precise estimates of the survival function. It needs however the fit to be good, i.e. to closely follow a parametric distribution, to provide an accurate estimation.

The difference between parametric and semiparametric estimation in cure models is that semiparametric does not assume a specific distribution for baseline survival in the uncured population. However, a similarity is that it also assumes a parametric form for the regression model for the cured population. Semiparametric methods use nonparametric estimation methods such as Kaplan-Meier or Cox proportional hazards regression to estimate the survival function of the susceptible population. Semiparametric methods can be more flexible than parametric methods, but they may require larger sample sizes and can be computationally heavy.

A nonparametric estimation for the cure models does not assume a specific distribution for either the latency or the incidence for the cured population. Instead, the nonparametric estimation methods use a special form of Kaplan-Meier to estimate the survival function in the susceptible population. Nonparametric methods are the most flexible but can be less precise than parametric or semiparametric methods, particularly for small sample sizes.

In summary, the choice of estimation method in cure models depends on the assumptions made about the underlying distributions and the desired level of flexibility and precision in the estimates. Parametric methods are more precise but require more assumptions, while nonparametric methods are more flexible but may be less exact. Nevertheless, semiparametric methods offer a balance between flexibility and precision. In the following chapter, various methodologies for the different estimation approaches will be explained.

2.5 Previous applications

As previous mentioned, cure models were first introduced to model cancer survival. A study in Norway described in [5] applied the mixture cure models to twenty three kinds of cancer, where the majority of them (fifteen types) resulted in an accurate cure fraction. The model gave valid results for cancers of the mouth and pharynx, oesophagus, stomach, colon, rectum, liver, gallbladder, pancreas, lung and trachea, ovary, kidney, bladder, CNS, non-Hodgkin lymphoma (only for males) and leukemia. Thus, there were some cancer types where cure models did not provide a good fit.

Maller and Zhou discuss in their book *Survival Analysis with Long-Term Survivors* [2] the application of cure models in the context of recidivism among released prisoners. The survival time of an individual will be denoted as the time that elapses before a rearrest. The prisoner's release from the prison was observed as well as the time to return to prison. The success was measured by how large the proportion of the prisoners that never came back was. A cut-off date was chosen and the persons who had not returned to the prison were labelled as right-censored. Due to the big data set and other assumptions, they could further be labelled as immune to the event of going back to prison. Since the prisoners were released at different times left truncation had to be considered. Extending this analysis to include covariates, such as country of origin and gender, makes these kind of models important within criminology. The first model of this study was done by Partanen (1969) who successfully fitted through maximum likelihood a mixture cure model using the exponential distribution. The data set consisted of times to first return to Finnish prisons out of 606 Finnish convicts. The conclusion was that this fit was better than a model with same distribution but without considering immunes. Several different cure models on convicts were constructed during the following year by applying different parametric distributions. Maller and Zhou applied their model on the data from Western Australian prisoners and eventually extended the method into being able to handle covariates.

Another area of application is engineering reliability where the individuals are components instead of persons. Nelson (1982) explains an experiment with motorettes components [2]. Ten motorettes are put on test at $t = 0$ on four different temperatures, where the survival time was the time to breakdown of the insulation. Since all motors eventually will fail, the immunes in this case were the motors that lived much longer relative to the components. In this study competing risks had to be considered since only breakdown due to insulation was the event in the study and the motor could break down due to other reasons.

3

Methods for the estimation of mixture cure models

This section describes the methodology to estimate mixture cure models. Consider a population divided into two subpopulations, one susceptible for the disease and one immune of the disease. The first group involves the individuals that will experience the event of interest given a sufficiently large follow-up time and the second the individuals who will not experience the event regardless of the follow-up time.

Let Y be the random variable for time until the event of interest and C the random variable for the censored times. The random variables are assumed to be independent from each other and the minimum of Y and C for each individual will be observed and defined as $T = \min\{Y, C\}$. Moreover, the indicator of death $\delta = \mathbf{1}\{Y < C\}$ is introduced. Denote $\tau > 0$ as the upper bound of the survival time for the non-susceptible individuals and let v be the indicator of cure that takes the value of 1 for cured individuals and 0 otherwise. We assume $P(T < \tau|\nu = 0) = 1$, which implies that the probability of experiencing the event before time τ given that the individual is not cured is 1. For simplicity, τ is usually set to ∞ which indicates that $v = 1$ can not be observed. Thus, a cured subject will never be observed but always right-censored. Hence it is not possible to distinguish if ν takes the value of 1 or 0 for a right-censored individual. However, in practice it is possible to give τ a finite value which implies a possibility for ν to take the value 1 .

Let $S_u(t) = P(Y > t|v = 0)$ and $S_c(t) = P(Y > t|v = 1)$ be the survival functions for the susceptible and the cured populations for any $t < \tau$, respectively. According to the definition of cured objects, $S_c(t)$ is a degenerate function due to the fact that it is constantly equal to 1 for $t < \tau$. The mixture cure model can be defined as the unconditional survival function of T for any $t < \tau$, that is

$$S(t) = \pi S_u(t) + (1 - \pi)S_c(t) = \pi S_u(t) + (1 - \pi). \tag{3.1}$$

The model expression comprises two components: the incidence, representing the probability of being cured $(1-\pi)$, and the latency, denoted as $S_u(t)$, which represents the survival distribution for the uncured population. This formulation allows for diverse effects of covariates on the different populations.

3.1 Parametric and semiparametric estimation

Several methods to estimate these two submodels have been presented since the mixture cure models were introduced in 1949. The following section will introduce several methods to estimate the incidence and latency parametrically or semiparametrically. The information for these approaches is mainly retrieved from the book *Cure models: Methods, Applications and Implementation* (Yingwei Peng and Binbing Yu) [6]. Consider $\boldsymbol{X}$ and $\boldsymbol{Z}$ to be the covariate vectors for the latency and incidence, respectively. The sample is represented by $(T_i, \delta_i, X_i, Z_i), i = 1, ..., n$, where $T_{(i)}$ denotes the i-th order statistic observation from the sample $(T_1, T_2, ..., T_n)$, and $\delta_{(i)}, X_{(i)}$, and $Z_{(i)}$ refer to the concomitants corresponding to the δ, X, and Z samples, respectively.

Consider the covariate vectors $\boldsymbol{X}$ and $\boldsymbol{Z}$. Then, the expression in (3.1) can be further written as

$$S(t) = \pi(z)S_u(t|x) + (1 - \pi(z)), \tag{3.2}$$

where $1 - \pi(z)$ is the incidence and $S_u(t|x)$ the conditional survival function for the susceptible subpopulation (the conditional latency function).

3.1.1 Parametric incidence submodel

The primary approach for estimating the incidence function will be through parametric methods. Given that the random variable Y exhibits a binary property, it will be modeled with the Bernoulli distribution, where the parameter π is used to represent the probability of a success. Let $\boldsymbol{z}$ be the vector of covariates for the incidence, starting with value 1, and $\boldsymbol{\gamma}$ the corresponding coefficient vector. Following, link functions will be presented to express the effect of the covariates on π, i.e. the link between π and the linear predictor $\boldsymbol{z}'\boldsymbol{\gamma}$. The most common parametrization is the logit link:

$$\log(\pi(\boldsymbol{z})/(1 - \pi(\boldsymbol{z}))) = \boldsymbol{z}'\gamma.$$

An interpretation of the value e^{γ} for a certain covariate $\boldsymbol{z}$ is that it represents the odds ratio of being uncured when the covariate is increased by one unit, while holding all other covariates constant. Another link function is the complementary loglog link function:

$$\log(-\log(\pi(\boldsymbol{z}))) = -\boldsymbol{z}'\boldsymbol{\gamma}.$$

In this case, the interpretation of e^{γ} is the relative log risk of remaining uncured when the corresponding covariate $\boldsymbol{z}$ is increased by one unit, assuming all other covariates remain constant. A third link that can be applied is the probit link, where the linear predictors effect is linked to π through the cumulative distribution function of the standard normal distribution:

$$\Phi^{(-1)}[\pi(\boldsymbol{z})] = \boldsymbol{z}'\boldsymbol{\gamma}.$$

However, there is no clear interpretation of $\boldsymbol{\gamma}$ here due to its restricted form. This link function, as well as the logit function will symmetrically model $\pi(\boldsymbol{z})$ and $1 - \pi(\boldsymbol{z})$ since the corresponding coefficient will only differ by a sign. However, the complementary loglog link approaches 0 slowly and 1 fast, and does not model the complementary probabilities equivalently. With a large sample size it might be useful to choose a link function based on the Box-Cox transformation (Box and Cox, 1964):

$$\pi(\boldsymbol{z}) = \left(1 + \lambda e^{-\boldsymbol{z}'\gamma}\right), \quad 0 \leq \lambda \leq 1,$$

which has both the logit and complimentary log-log links as special cases.

3.1.2 Parametric latency submodel

To parametrically model the second part of the mixture cure model we specify the distribution of the survival times of the uncured population together with the effect of the covariates on the distribution. This is done by expressing the survival function $S_u(t)$ in terms of the baseline survival function $S_{u0}(t)$ and a function $e^{\boldsymbol{x}'\boldsymbol{\beta}}$ where $\boldsymbol{x}$ is the covariate vector. For the case with no covariates, i.e. when $x = 0$, the survival function will be equal to the baseline survival function making only estimation of the parameters of the chosen parametric distribution necessary. Specification of the latency can be done in different ways. An important note is that the covariate vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ do not necessarily have to be equal. The first model that will be explained is the parametric latency submodel under the proportional hazards (PH) assumption.

3.1.2.1 Parametric PH latency submodel

The survival function for the susceptible population will be modeled as

$$S_u(t|\boldsymbol{x}) = S_{u0}(t)^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}. \tag{3.3}$$

Here, $S_{u0}(t)$ is the baseline survival function which will be modeled with a parametric distribution. For instance, $S_{u0}(t)$ can be modeled by the exponential distribution which implies the baseline function $S_{u0}(t) = e^{-\lambda t}$ with rate λ. This further implies the survival function $S_u(t|\boldsymbol{x}) = e^{-\lambda \exp(\boldsymbol{x}'\boldsymbol{\beta})t}$ from the exponential distribution but with the rate $\lambda \exp(\boldsymbol{x}'\boldsymbol{\beta})$. Another choice for the baseline survival function is the Weibull distribution resulting in $S_{u0}(t) = e^{-\lambda t^p}$ where λ is the scale parameter and p is the shape. This further implies the survival function $S_u(t|\boldsymbol{x}) = e^{-\lambda \exp((\boldsymbol{x}'\boldsymbol{\beta}))t^p}$ which as well is a Weibull function, but with the scale parameter $\lambda \exp(\boldsymbol{x}'\boldsymbol{\beta})$. The interpretation of $\boldsymbol{\beta}$ is equivalent to the one for the Cox model where $\boldsymbol{\beta}_i$ for covariate $\boldsymbol{x}_i$ denotes the log-hazard ratio when the covariate is increased with 1 unit holding all other covariates are fixed.

3.1.2.2 Parametric AFT latency submodel

An alternative parametric submodel is the accelerated failure time model, where the survival for the uncured population will be modeled as

$$S_u(t|\boldsymbol{x}) = S_{u0}(te^{-\boldsymbol{x}'\boldsymbol{\beta}}), \tag{3.4}$$

where $S_{u0}(t)$ is the baseline survival. If $\log(T|Y=1) = \boldsymbol{x}'\boldsymbol{\beta} + \sigma\varepsilon$ satisfies $P(e^{\sigma\varepsilon} > t) = S_{u0}(t)$ where σ is the scale parameter and ε the error term, then $T|Y=1$ will satisfy the accelerated failure time assumption. Usually, the distribution for $e^{\sigma\varepsilon}$ is modeled with a parametric distribution. For instance, if e^{ε} follows the exponential distribution with $\sigma = 1$ or if ε follows the extreme value distribution with survival function $P(\varepsilon > s) = \exp(-e^{s})$ then the latency is given by $S_u(t|\boldsymbol{x}) = \exp(-(te^{-\boldsymbol{x}'\boldsymbol{\beta}})^{1/\sigma})$. Thus, the corresponding mixture cure model is a Weibull AFT mixture cure model. Additionally, this model can also be considered a PH mixture cure model, as it satisfies the proportional hazards assumption. However, the term $e^{\sigma\varepsilon}$ is not limited to the Weibull model alone; it can also be modeled using a normal distribution, leading to a lognormal mixture cure model. While there are other parametric models available, this project will focus specifically on these two mentioned models.

3.1.2.3 Direct maximization of observed likelihood function

Suppose the observed data is given by: $(t_i, \delta_i, z_i, x_i), i = 1, 2, ..., n$ where n is the sample size. For each individual i, t_i represents the observed survival time, δ_i is the event indicator where $\delta_i = 1$ if the event is uncensored and $\delta_i = 0$ otherwise. The covariate vectors for the incidence and latency for each individual are denoted $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$, respectively. Let $\boldsymbol{\alpha}$ be a vector with the unknown parameters in the parametric baseline survival and $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha})'$. The likelihood function for the mixture cure model is expressed as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [\{\pi(z_i) f_u(t_i \mid \boldsymbol{x}_i)\}]^{\delta_i} [\{1 - \pi(z_i)\}\{\pi(z_i) S_u(t_i \mid \boldsymbol{x}_i)\}]^{1-\delta_i},$$

and the log-likelihood $\ell(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta}))$ can be maximized using the Newton-Raphson method. Let

$$\boldsymbol{U}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}} \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} \end{pmatrix}, \quad I(\boldsymbol{\theta}) = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = -\begin{pmatrix} \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\beta}'} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\alpha}'} \\ \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}'} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}'} \\ \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\gamma}'} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\beta}'} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} \end{pmatrix}.$$

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is obtained through iterating over the formula $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \boldsymbol{I}^{-1}(\boldsymbol{\theta}^{(k)})\boldsymbol{U}(\boldsymbol{\theta}^{(k)})$ until the difference between $\boldsymbol{\theta}^{(k+1)}$ and $\boldsymbol{\theta}^{(k)}$ is small enough. The variance of $\hat{\boldsymbol{\theta}}$ can be approximated by $I^{-1}(\hat{\boldsymbol{\theta}})$.

At the maximum likelihood estimator, $\boldsymbol{U}(\boldsymbol{\theta}) = 0$ means that the partial derivatives of the log-likelihood function with respect to the parameters, evaluated at $\hat{\boldsymbol{\theta}}$, are zero. This indicates that we have found the parameter values that maximize the likelihood of the observed data. In other words, the score equation, which represents the first-order condition for maximum likelihood estimation, is satisfied at $\hat{\boldsymbol{\theta}}$. This property confirms that the Newton-Raphson method has converged to the optimal solution, ensuring that our estimates are the most likely values given the data.

3.1.2.4 Estimation via EM algorithm

Another way to maximize the likelihood is with the EM algorithm. Let y_i be the value of Y for subject i, where $y_i = 1$ when $\delta_i = 1$. Since there is no way to know if a censored individual has been cured or has not experienced the event yet when $\delta_i = 0$, y_i will be unknown for the censored individuals. Consider the log-likelihood for the observed data $(t_i, \delta_i, \boldsymbol{z}_i, \boldsymbol{x}_i, y_i), i = 1, 2, ..., n$, given that all values of y_i are observed, as

$$\ell^c(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \log \prod_{i=1}^{n} [\pi\left(\boldsymbol{z}_i\right) f_u\left(t_i \mid \boldsymbol{x}_i\right)^{\delta_i} S_u\left(t_i \mid \boldsymbol{x}_i\right)^{1-\delta_i}]^{y_i} \left[1 - \pi\left(\boldsymbol{z}_i\right)\right]^{1-y_i},$$

which can be written as $\ell^c(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \ell_1(\boldsymbol{\gamma}) + \ell_2(\boldsymbol{\beta}, \boldsymbol{\alpha})$ where

$$\ell_1(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \Big(y_i \log\left[\pi\left(\boldsymbol{z}_i\right)\right] + (1 - y_i) \log\left[1 - \pi\left(\boldsymbol{z}_i\right)\right] \Big)$$

and

$$\ell_2(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1} y_i \Big(\delta_i \log[f_u(t_i | \boldsymbol{x}_i)] + (1 - \delta_i) \log[S_u(t_i \mid \boldsymbol{x}_i)] \Big).$$

Denote the probability of an individual with covariates $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ to be uncured given that it survived until time t by

$$w_{0i}(t) = \frac{\pi\left(\boldsymbol{z}_i\right) S_u\left(t \mid \boldsymbol{x}_i\right)}{1 - \pi\left(\boldsymbol{z}_i\right) + \pi\left(\boldsymbol{z}_i\right) S_u\left(t \mid \boldsymbol{x}_i\right)}.$$

The algorithm consists of two steps, the expectation step (E-step) and the maximization step (M-step). The initial state of the EM algorithm is $(\boldsymbol{\gamma}^{(0)}, \boldsymbol{\beta}^{(0)}, \boldsymbol{\alpha}^{(0)})$. Given the estimates of the parameters in the $(k-1)$th iteration the E-step calculates the posterior expectation of y_i as

$$w_i^{(k)} = \delta_i + (1 - \delta_i) w_{0i}(t_i),$$

where $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = (\boldsymbol{\gamma}^{(k-1)}, \boldsymbol{\beta}^{(k-1)}, \boldsymbol{\alpha}^{(k-1)})$. This follows from Bayes' Theorem and the properties of y_i. Hereafter, the purpose of the M-step is to maximize the two terms in the log-likelihood. In the kth iteration y_i is replaced with $w_i^{(k)}$ and to update the current estimates of $\boldsymbol{\theta}$ the sums

$$\ell_1(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \Big(w_i^{(k)} \log[\pi\left(\boldsymbol{z}_i\right)] + (1 - w_i^{(k)}) \log[1 - \pi(\boldsymbol{z}_i)] \Big) \tag{3.5}$$

and

$$\ell_2(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} w_i^{(k)} \Big(\delta_i \log\left[f_u\left(t_i \mid \boldsymbol{x}_i\right)\right] + (1 - \delta_i) \log\left[S_u\left(t_i \mid \boldsymbol{x}_i\right)\right] \Big)$$

are maximized.

The algorithm is iterated until the difference between the new and old estimates is small enough, hence the algorithm converged. ℓ_1 can be seen as the log-likelihood function for logistic regression with the vector of $w_i^{(k)}, k = 1, ..., n$ as response while ℓ_2 can be treated as a weighted log-likelihood for censored data due to their similarities. Hence, both expressions can be maximized with the Newton-Raphson method or through computational methods for logistic regression or standard survival methods allowing weights, respectively.

3.1.3 Software for parametric estimation

For the simplest parametric cure models, standard statistical packages can be enough to fit and obtain the maximum likelihood for the models. This could be if for instance the latency function is a Weibull or exponential function and the model does not consider any covariates. In cases where the latency distribution is complex or the effects of covariates exhibit nonlinearity in either the incidence or latency, standard survival analysis packages may not be sufficient. The following R packages implement the methods described above to provide estimates of the incidence and latency function [6].

3.1.3.1 `gfcure`

The package `gfcure` in R can fit different parametric accelerated failure time mixture cure models. The function `gfcure` works similar as `survreg` from the `survival` package, but includes two formula arguments for the two parts in the mixture cure model. Most parametric distributions can be specified for the baseline distribution in the AFT latency submodel part; lognormal, exponential, Rayleigh, Weibull, gamma, loglogistic, general loglogistic, generalized F, and extended generalized gamma.

3.1.3.2 `mixcure`

`mixcure` was developed to fit a parametric mixture cure model using already existing packages in R. Using `survreg` from the `survival` package or `flexsurvreg` from `flexsurv` the latency function can be fitted with any of the parametric distributions available in the corresponding package. The incidence can be fitted with the function `glm` by adding an argument which specifies what link function that will be used.

3.1.3.3 `flexsurvcure`

The R package `flexsurvcure` utilizes the `flexsurv` package to estimate a range of parametric cure models, including both mixture and non-mixture models. To achieve this, the function incorporates custom distributions that correspond to specific cure models and adds them to the `flexsurv` package. This approach allows for more flexibility in the modeling process, as additional parameters in a distribution can be dependent on covariates. As a result, the cure models generated using flexsurvcure are more versatile than those previously mentioned.

3.1.4 Residuals to assess goodness of fit

The Schoenfeld residuals can be used to assess the goodness of the parametric fit for the mixture cure model. Assuming the same set of covariates in both the incidence and the latency part of the model, the Schoenfeld residuals are given by

$$\delta_i \left[\boldsymbol{x}_i - \frac{\sum_{j=1}^{n} Y_j(t_i)\boldsymbol{x}_j h(t_i \mid \boldsymbol{x}_j, \boldsymbol{z}_j)}{\sum_{j=1}^{n} Y_j(t_i) h(t_i \mid \boldsymbol{x}_j, \boldsymbol{z}_j)} \right], \quad i = 1, \dots, n,$$

where $Y_i(t) = I(t_i \geq t)$ is an indicator taking the value 1 if the condition is true and 0 otherwise. The function $h(t|\boldsymbol{x}, \boldsymbol{z})$ represents the unconditional hazard function corresponding to the survival function. These residuals represent the discrepancy between the covariate value of a certain uncensored individual and the weighted average of the covariate values for those still at risk, where the conditional hazard function at the uncensored time is the weight. The mean of the residuals for a perfectly fitted mixture model would be zero. Two good ways to detect patterns in the fit is to either plot the residuals against time or to fit a polynomial regression to the residuals.

3.1.5 Semiparametric latency submodels

In addition to parametric estimation, semiparametric estimation methods offer an alternative approach for estimating the mixture cure model. Similar to the parametric approach discussed earlier, the incidence is estimated parametrically, resulting in ℓ_1 as described in Equation (3.5). However, in the following sections, we will explore various semiparametric methods for estimating the latency component of the model.

3.1.5.1 Semiparametric PH latency submodel

Equivalently as for the parametric PH latency submodel, the survival function is given by $S_u(t|\boldsymbol{x}) = S_{u0}(t)^{\exp(\boldsymbol{x}'\boldsymbol{\beta})}$. The corresponding hazard function is $h_u(t \mid \boldsymbol{x}) = h_{u0}(t)e^{\boldsymbol{\beta}'\boldsymbol{x}}$ which implies the likelihood function for the parameters $\boldsymbol{\beta}$ and S_{u0}, denoted as

$$\ell_2(\boldsymbol{\beta}, S_{u0}) = \sum_{i=1}^{n} w_i \Big(\delta_i \log h_{u0}(t_i) + \delta_i \boldsymbol{\beta}'\boldsymbol{x}_i + e^{\boldsymbol{\beta}'\boldsymbol{x}_i} \log[S_{u0}(t_i)] \Big). \tag{3.6}$$

In the case where $\delta_i = 1$ we have that $w_i = 1$ and $\log w_i = 0$ and therefore we express it as

$$\ell_2(\boldsymbol{\beta}, S_{u0}) = \sum_{i=1}^{n} \Big(\delta_i \log h_{u0}(t_i) + \delta_i \boldsymbol{\beta}'\boldsymbol{x}_i + e^{\log w_i + \boldsymbol{\beta}'\boldsymbol{x}_i} \log[S_{u0}(t_i)] \Big). \tag{3.7}$$

Similarly as in the previous section, the expression in Equation (3.6) can be seen as a weighted PH likelihood function for all individuals with $w_i > 0$. Furthermore, Equation (3.7) can be seen as a likelihood function of a PH model with an offset term $\log w_i$. These two likelihood expressions can be maximized with either Newton-Raphson method or methods in R that allow weights or offset terms.

Due to the fact that a nonparametric or unspecified S_{u0} will result in a special case of the Cox PH model the likelihood functions in (3.6) and (3.7) can be maximized individually. Hence, the partial log-likelihood function for parameter $\boldsymbol{\beta}$ is

$$\log \prod_{j=1}^{k} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{s}_j)}{\left\{\sum_{i \in R_j} \exp(\log w_i + \boldsymbol{\beta}'\boldsymbol{x}_i)\right\}^{d_j}},$$

where $\boldsymbol{s}_j = \sum_{i:t_i=\tau_j} \delta_i \boldsymbol{x}_i$, k is the number of uncensored failure times $\tau_1 < \tau_2 < ... < \tau_k$, d_j is the number of uncensored times equal to τ_j and R_j the risk set at time τ_j.

With updated $\hat{\boldsymbol{\beta}}$, the estimated baseline survival function can be obtained by the Nelson-Aalen estimator, written as

$$\hat{S}_{u0}(t) = \exp\left(-\sum_{j:\tau_j<t} \frac{d_j}{\sum_{i\in R_j} \exp\left(\log w_i + \hat{\boldsymbol{\beta}}'\boldsymbol{x}_i\right)}\right).$$

3.1.5.2 Semiparametric AFT latency submodel

In this section we explore the effect on $\boldsymbol{x}$ in the latency under the accelerated failure time assumption. As for the semiparametric PH model a direct maximization of the likelihood is not plausible due to the unspecified baseline survival. An implementation of the EM algorithm in R is more complex than for the semiparametric PH model, making other methods reasonable to consider, for instance the linear rank method, M-estimation or Kernel smoothing estimation.

The linear rank method will now be briefly explained. Similar to the parametric AFT model, we model the effects on $\boldsymbol{x}$ using $S_u(t|\boldsymbol{x}) = S_{u0}(te^{-\boldsymbol{x}'\boldsymbol{\beta}})$, where $S_{u0}(t)$ represents the baseline survival. Considering the function $\log(T|Y=1) = \boldsymbol{x}'\boldsymbol{\beta} + \sigma\varepsilon$, we define $f_0(\cdot)$, $h_0(\cdot)$, and $S_0(\cdot)$ as the density, hazard, and survival function of ε, respectively. These three functions are employed to formulate the log-likelihood function for the latency component as

$$\begin{aligned}\ell_2(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \sum_{i=1, w_i>0}^{n} \Big(\delta_i \log f_0\left(\log t_i - \boldsymbol{\beta}'\boldsymbol{x}_i\right) + w_i\left(1-\delta_i\right)\log\left[S_0\left(\log t_i - \boldsymbol{\beta}'\boldsymbol{x}_i\right)\right]\Big) \\ &= \sum_{i=1, w_i>0}^{n} \Big(\delta_i \log h_0\left(\log t_i - \boldsymbol{\beta}'\boldsymbol{x}_i\right) + w_i \log\left[S_0\left(\log t_i - \boldsymbol{\beta}'\boldsymbol{x}_i\right)\right]\Big).\end{aligned}$$

To estimate $\boldsymbol{\beta}$ the linear programming method (Jin et al. 2003). Define the rank-like function (Zhang and Peng, 2007) as

$$\sum_{i=1}^{n} \delta_i g\left(\varepsilon_i\right)\left(\boldsymbol{x}_i - \frac{\sum_{j=1}^{n} \boldsymbol{x}_j w_j I\left(\varepsilon_j \geq \varepsilon_i\right)}{\sum_{j=1}^{n} w_j I\left(\varepsilon_j \geq \varepsilon_i\right)}\right),$$

where $\varepsilon_i = \log t_i - \boldsymbol{\beta}' x_i$ and $g(\cdot)$ is be defined as a Gehan-type weight function $g(u) = \sum_{j=1}^{n} I\left(\varepsilon_j \geq u\right) w_j / n$. Thus, the expression in Equation (3.1.5.2) can further be written as:

$$n^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} \delta_i\left(x_i - x_j\right) w_j I\left(\varepsilon_j \geq \varepsilon_i\right),$$

and interpreted as the gradient of a convex function. This function can then be minimized by the linear programming method.

Given an estimator $\hat{\boldsymbol{\beta}}$, the survival function of ε can be estimated nonparametrically using the Nelson-Aalen estimator:

$$\hat{S}_0(\varepsilon) = \exp\left(-\sum_{j:\tau_j^*<\varepsilon} \frac{d_j}{\sum_{i\in R_j} w_i}\right),$$

where $\tau_1^* < \tau_2^*, ..., < \tau_k^*$ are distinct uncensored values of $\log t_i - \boldsymbol{\beta}'\boldsymbol{x}_i$, d_j the corresponding amount of uncensored survival times and R_j is the risk set at τ_j^*. Moreover, the estimate of $S_{u0}(t)$ can be derived from $\hat{S}_0(\varepsilon)$. Further methods to estimate the AFT latency model semiparametrically are explained in [6].

3.1.5.3 Residuals to assess goodness of fit

To evaluate the fit of a semiparametric mixture cure model is challenging because of the limited information about the hazard function. However, it is still possible to investigate the fit by using the martingale residuals, which are also common in standard survival analysis fitting. The martingale residuals are given as

$$M_i = \delta_i + \log\left[1 + e^{\boldsymbol{\gamma}'\boldsymbol{z}_i - H_{u0}(t_i)\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}\right] - \log\left[1 + e^{\boldsymbol{\gamma}'\boldsymbol{z}_i}\right]. \tag{3.8}$$

Adding the estimations for $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and H_{u0} into (3.8) we get the martingale residual of the fit. This represents the difference between the observed number of events for an individual and the expected amount based on the fitted model, taking into account the covariates and follow-up time.

The package `smcure` provides estimations for both the semiparametric proportional hazards mixture cure model and the accelerated failure time mixture cure model by implementing the methods presented in this section. This is done using the functions `predictsmcure` and `plotpredictsmcure`.

3.2 Nonparametric estimation

When the data does not follow the distribution assumed by a given parametric model, using that model to estimate incidence and latency can introduce bias. An alternative is nonparametric estimation, where nonparametric methods can be employed to estimate these quantities without assuming any specific distribution for the data [7].

For simplicity, we consider $\boldsymbol{X}$ as the covariate vector for both the incidence and the latency. Let $F(t|\boldsymbol{x}) = P(Y \leq t|\boldsymbol{X} = \boldsymbol{x})$ and $G(t|\boldsymbol{x}) = P(C \leq t|\boldsymbol{X} = \boldsymbol{x})$ be the distribution functions of Y and C conditional on $\boldsymbol{X} = \boldsymbol{x}$, respectively. The sample is denoted by $(T_i, \delta_i, X_i), i = 1, ..., n$, where $T_{(i)}$ represents the observation of the i-th order statistic with respect to the sample $(T_1, T_2, ..., T_n)$, and $\delta_{(i)}$ and $X_{(i)}$ represent the concomitants to the $\boldsymbol{\delta}$ and $\boldsymbol{X}$ samples, respectively.

3.2.1 Nonparametric incidence submodel

First of all, to estimate the incidence for a model with no covariates define $\hat{S}_{KM}(t)$ as the Kaplan-Meier estimator of the survival function with observations $\{(T_i, \delta_i), i = 1, ..., n\}$ and $T^1_{\max} = \max\limits_{i:\delta_i=1}(T_i)$ as the largest uncensored time. Taking $\lim_{t\to\infty}$ in the expression in (3.1) we get the cure estimate $(1 - \pi)$ which corresponds to the last value of the $\hat{S}_{KM}(t)$. Hence, the estimate of π can be expressed as

$$\hat{\pi} = 1 - \hat{S}_{KM}(T^1_{\max}).$$

Furthermore, we consider a model with a univariate continous covariate X. The generalized Kaplan-Meier estimator of Beran (1981) [8] is used to estimate the conditional survival function denoted as

$$\hat{S}_h(t \mid x) = \prod_{T_{(i)} \leq t} \left(1 - \frac{\delta_{(i)} B_{h(i)}(x)}{\sum_{r=i}^{n} B_{h(r)}(x)} \right), \tag{3.9}$$

where $B_{h(i)}(\boldsymbol{x}) = K_h \left(x - x_{(i)} \right) / \sum_{j=1}^{n} K_h \left(x - x_{(j)} \right)$ are the Nadaraya-Watson weights and $K_h(\cdot) = \frac{1}{h} K(h)$ is the rescaled kernel with bandwidth $h > 0$. Using the estimator in Equation (3.9), Xu and Peng (2014) introduced the estimator for the incidence function and thereby the cure rate is given by

$$1 - \hat{\pi}_h(x) = \hat{S}_h(T^1_{\max} | x). \tag{3.10}$$

Xu and Peng (2014) further provided a condition which if fulfilled should strength the reliability of the method by ensuring a sufficient follow-up time. Define $\tau_{S_0}(x) = \sup\{t : S_0(t|x) > 0\}$ and $\tau_G(x) = \sup\{t : G(t|x) < 1\}$ and further $\tau_0 = \sup_{x \in D} \tau_{S_0}(x)$ where D is the support of X and consider the inequality:

$$\tau_0 < \tau_G(x), \forall x \in D. \tag{3.11}$$

This condition is vital for a couple of reasons. First of all, it ensures that when the probability that a susceptible individual survives beyond the largest censoring time $\tau_G(x)$ is zero, the estimations of the incidence and latency will still be consistent. Secondly, due to the fact that T^1_{max} converges to τ_0 the condition ensures that all times observed after T^1_{max} can be assumed to correspond to cured individuals. If this would not be true the estimator of the incidence function would be larger than the actual value. Another point is that if the last observation is uncensored, no matter how many late censored observations are in the study, the cure rate will be estimated to zero. Hence, it is very important to have a large sample size and be careful with the choice of the follow-up time.

To make sure that the condition in (3.11) is true Maller and Zhou (1996) proposed a nonparametric test that evaluates the constant right tail of the KM estimate, i.e. looks at the difference between the last failure time and the last censored time. A long plateau together with a heavy censoring in this interval should result in a sufficiently long follow-up time for the condition to hold [2].

3.2.2 Nonparametric latency submodel

Given the same covariate effects for the latency as for the incidence in the mixture cure model a nonparametric estimator for the conditional latency function follows from Equation (3.2) (López-Cheda et al. 2017b) and is given by

$$\hat{S}_{u0}(t \mid x) = \frac{\hat{S}_h(t \mid x) - (1 - \hat{\pi}_h(x))}{\hat{\pi}_h(x)}$$

with $\hat{S}_h(t|x)$ and $(1 - \hat{\pi}_h(x))$ as in (3.9) and (3.10), respectively. The optimal bandwidths for these two estimates are not necessarily equal, but it is recommended to be to ensure that $\hat{S}_h(t|x) \to 1 - \hat{\pi}_h(x)$ when $t \to \infty$. The bandwidth is chosen through bootstrap selection [7].

3.2.3 Software for nonparametric estimation

The package `npcure` in R is available for nonparametric estimation of the latency and incidence functions based on the theory in Section 3.2. For estimation of the Beran estimator in (3.9) `npcure` uses the Epanechnikov kernel: $K(u) = \frac{3}{4}(1-u^2)\mathbf{1}_{|u|\leq 1}$. The functions `probcure()` and `latency()` provide estimates of the incidence and latency functions, respectively [9]. The `probcure()` is called with:

```
probcure(x, t, d, dataset = NULL, x0, h, local = TRUE, conflevel = 0L
     bootpars = if (conflevel == 0 && !missing(h)) NULL else
          controlpars()),
```

where the `x` argument specifies the covariate included in the model, `t` the survival time, `d` indicates whether the event happened or not, `x0` specifies the covariate values the cure rate will be estimated for. The `h` argument specifies the bandwidth while `local` is either true or false depending on whether it is a local or a global bandwidth. The `conflevel` argument can be used in case a confidence interval is wanted. The `latency()` function has similar arguments to `probcure()` and is called with:

```
latency(x, t, d, dataset = NULL, x0, h, local = TRUE, testimate =
      NULL, conflevel = 0L, bootpars = if (conflevel == 0) NULL else
          controlpars(), save = TRUE).
```

Additionally, it has an argument `testimate` which determines the time t at which the function $S_0(t)$ is estimated. Furthermore, the functions `probcurehboot()` and `latencyhboot()` compute bootstrap bandwidths for the estimators. A full description of the functions of the package `npcure` is found in Table 3.1 below.

3.2.4 Nonparametric estimation when the cure status is partially known

The estimators in the previous method can be biased if some individuals in the data are known to be cured before the end of the follow-up time. An approach that considers a known cure status for some subject has been developed by Safari, López-de Ullibarri and Jácome [10]. The proposed estimator is based on the nonparametric estimator introduced in Section 3.2.1. We introduce a new indicator, denoted as ξ, which represents whether the cure status of a subject is known ($\xi = 1$) or unknown ($\xi = 0$). Recall that ν serves as the indicator of cure, taking the value of 0 for the individuals who are uncured and 1 for those who are cured. In contrast to the previously presented estimator, this new method incorporates the indicator ξ, which allows us to determine the cure indicator ($v = 1$) for patients when $\xi = 1$. However, for patients who have not experienced the event or whose cure status is unknown, ν remains unknown. To address this, we introduce the product $\nu\xi$, which equals 0 for the aforementioned group since ξ is 0.

The censoring distribution will be an improper distribution function $G(t|\boldsymbol{x}) = \{1 -$

Function	Description
`beran`	Computes Beran's estimator of the conditional survival function.
`berancv`	Computes the CV bandwidth for Beran's estimator of the conditional survival function.
`controlpars`	Sets the control parameters of the `latencyhboot()` and `probcurehboot()` functions.
`hpilot`	Computes pilot bandwidths for the nonparametric estimators of the cure rate and the latency.
`latency`	Computes the nonparametric estimator of the latency.
`print.npcure`	Method of the generic function `print` for 'npcure' objects.
`probcure`	Computes the nonparametric estimator of the cure rate.
`probcurehboot`	Computes the bootstrap bandwidth for the nonparametric estimator of the cure rate.
`summary.npcure`	Method of the generic function `summary` for 'npcure' objects.
`testcov`	Performs covariate significance tests for the cure rate.
`testmz`	Performs the nonparametric test of Maller and Zhou (1992).

Table 3.1: Descriptions of the functions in the `npcure` package.

$\pi(\boldsymbol{x})\}G_0(t|x)$. This can be interpreted as that the probability that C takes the value of ∞ is $\pi(\boldsymbol{x})$ while the probability that it takes the value of a random variable C_0 with proper continuous distribution function $G_0(t)$ is $1-\pi(\boldsymbol{x})$. Due to the assumption of conditional independence between Y and C given the covariates, a cured individual will be identified with probability

$$P(\xi = 1|\nu = 1, \boldsymbol{X} = \boldsymbol{x}) = P(C = \infty|\boldsymbol{X} = \boldsymbol{x}) = \pi(\boldsymbol{x}).$$

The method will assume three groups in the observed data $\{(T_i, \delta_i, \boldsymbol{X}_i, \xi_i\nu_i) : i = 1, ..., n\}$:

1. $(T_i, \delta_i = 1, \boldsymbol{X}_i, \xi_i\nu_i = 0)$ - The subjects who experienced the event of interest (uncensored).
2. $(T_i, \delta_i = 0, \boldsymbol{X}_i, \xi_i\nu_i = 0)$ - The subjects who have neither experienced the event nor been cured at the follow-up time (censored).
3. $(T_i, \delta_i = 0, \boldsymbol{X}_i, \xi_i\nu_i = 1)$ - The subjects where cure has been observed before the follow-up time.

The survival time will be given as $T_i = \min\{Y_i, C_i\}[1 - \mathbf{1}(Y_i = \infty, C_i = \infty)] + C_{0i}\mathbf{1}(Y_i = \infty, C_i = \infty)$ making the survival time for the last group be given as the constant C_{0i}. Similarly as previous introduced methods, the mixture cure model is constructed by defining the probability of cure as $1 - \pi(x) = P(Y = \infty, \boldsymbol{X} = x)$ and the latency as $S_0(t|x) = P(Y > t|Y < \infty, \boldsymbol{X} = \boldsymbol{x})$. We consider, for simplicity, a univariate continous covariate X with density function $m(x)$. An estimator of the conditional cumulative hazard function of Y, $\Lambda(t|x)$ when the cure status is partially known is

$$\widehat{\Lambda}_h^c(t|x) = \sum_{i=1}^{n} \frac{\delta_{[i]} B_{h[i]}(x)\mathbf{1}(T_{(i)} \leq t)}{\sum_{j=i}^{n} B_{h[j]}(x) + \sum_{j=1}^{i-1} B_{h[j]}(x)\mathbf{1}(\xi_{[j]}v_{[j]} = 1)},$$

where $X_{[i]}, \delta_{[i]}, \xi_{[i]}$ and $v_{[i]}$ are the concomitants of the ordered observed times $T_{(1)} \leq ... \leq T_{(n)}$. The corresponding proposed product-limit estimator of the conditional survival function $S(t|x)$ when the cure status is partially known is

$$\widehat{S}_h^c(t \mid x) = \prod_{i=1}^{n} \left\{ 1 - \frac{\delta_{[i]} B_{h[i]}(x) \mathbf{1}\left(T_{(i)} \leq t\right)}{\sum_{j=i}^{n} B_{h[j]}(x) + \sum_{j=1}^{i-1} B_{h[j]}(x) \mathbf{1}\left(\xi_{[j]} v_{[j]} = 1\right)} \right\}.$$

An important note is that the subjects who are known to be cured before the time $T_{(i)}$ remain in the risk set, and are thereby encountered in the denominator of the expression. The proof of this estimator is presented in [10].

In the case of no known cured individuals the estimator will be reduced to Beran's estimator in Equation (3.9). For the case where there are known cures with survival times only at one specific threshold the estimator will be reduced to Beran's as well. Lastly, when there is no censoring it will be reduced to the kernel-type estimator of the conditional survival function that was introduced by Nadaraya (1964), expressed as

$$\widetilde{S}_h(t \mid x) = \sum_{i=1}^{n} B_{h[i]}(x) \mathbf{1}\left(T_{(i)} > t\right).$$

The corresponding estimator of the cure probability $1 - \pi(x)$ (Safari et al. 2022) is

$$1 - \widehat{\pi}_h^c(x) = \widehat{S}_h^c\left(T_{\max}^1 \mid x\right).$$

Similarly as for the estimator in Section 3.2.2, the optimal bandwidth for $\widehat{S}_h^c(t \mid x)$ is not necessarily the same as the optimal for $1 - \widehat{\pi}_h^c(x)$ [11]. The latency was proposed (Safari et al. 2023) as

$$\widehat{S}_{0,h_1,h_2}^c(t \mid x) = \begin{cases} \frac{\widehat{S}_{h_2}^c(t|x) - \left\{1 - \widehat{\pi}_{h_1}^c(x)\right\}}{\widehat{\pi}_{h_1}^c(x)} & \text{if } 0 \leq t \leq T_{\max}^1 \text{ and } \widehat{S}_{h_2}^c(t \mid x) > 1 - \widehat{\pi}_{h_1}^c(x) \\ 0 & \text{otherwise.} \end{cases} \tag{3.12}$$

In the unconditional case the estimator in Equation 3.12 is expressed as

$$\widehat{S}_{0,n}^c(t) = \frac{\widehat{S}_n^c(t) - (1 - \widehat{p}_n^c)}{\widehat{p}_n^c} \tag{3.13}$$

where

$$\widehat{S}_n^c(t) = \prod_{i=1}^{n} \left\{ 1 - \frac{\delta_{[i]} \mathbf{1}\left(T_{(i)} \leq t\right)}{n - i + 1 + \sum_{j=1}^{i-1} \mathbf{1}\left(\xi_{[j]} v_{[j]} = 1\right)} \right\}$$

is the generalization of the Kaplan-Meier estimator of the survival function with a cured proportion in the data and where some of the subjects are identified as cured. Lastly, the unconditional estimator of the probability of cure is given by

$$1 - \widehat{p}_n^c = \widehat{S}_n^c\left(T_{\max}^1\right).$$

3.2.4.1 Bootstrap selection of the bandwidths

The optimal bandwidths h_1 and h_2 are selected through bootstrapping. The principle of the bootstrap-based selection methods is to choose the bandwidths that minimize a bootstrap estimate of the mean integrated squared error (MISE). An approximation of the bootstrap MISE can be written as

$$\mathrm{MISE}_x^*\left(h_1, h_2\right) \simeq \frac{1}{B} \sum_{b=1}^{B} \int \left\{\widehat{S}_{0,h_1,h_2}^{c,*b}(v \mid x) - \widehat{S}_{0,g_{1x},g_{2x}}^{c}(v \mid x)\right\}^2 \omega(v,x) dv \tag{3.14}$$

where $\widehat{S}_{0,h_1,h_2}^{c,*b}(t|x)$ is the estimator computed with the bth boostrap resample and the bandwidths, while $\widehat{S}_{0,g_{1x},g_{2x}}^{c}(v \mid x)$ is the estimator computed with the original sample using the pilot bandwidths (g_{1_x}, g_{2_x}). Furthermore, $\omega(v,x)$ is a non-negative weight with purpose of giving a lower weight to the right tail of the distribution. The bootstrap bandwidths will be computed through six steps presented in [11].

3.2.4.2 Asymptotic properties

In this section we will present the asymptotic properties for the estimators in the previous section. For further information we refer to the articles [10] and [11]. Consider the following (sub)distribution functions:

$$\begin{aligned}
H(t \mid x) &= P(T \leq t \mid X = x), \\
H^1(t \mid x) &= P(T \leq t, \delta = 1 \mid X = x), \\
H^{11}(t \mid x) &= P(T \leq t, \xi = 1, v = 1 \mid X = x), \\
J(t \mid x) &= 1 - H(t \mid x) + H^{11}(t \mid x).
\end{aligned}$$

The estimator $\widehat{S}_h^c(t|x)$ will be expressed as $1 - \widehat{F}_h^c(t|x)$ according to the definition of the survival function. Assumptions 1-8 in [10] state that the random variables Y and C are conditionally independent and that the derivatives and the second derivatives of the sub(distributions) exist and are continuous with respect to t and/or x. Moreover, the kernel function $K(v)$ is a symmetrical density with zero mean, vanishing outside of (-1,1) and the total variation is less than $\lambda < \infty$. Under these assumptions, Theorems 1-2 establish the asymptotic representations of $\widehat{\Lambda}_h^c$ and $1 - \widehat{F}_h^c(t|x)$. Furthermore, the strong consistency of the estimators is obtained in Corollary 1. Under same assumptions the bias and variance of $1 - \widehat{F}_h^c(t|x)$ is derived in Proposition 3 while Theorem 3 establishes that asymptotically normality holds. For further evidence of the asymptotic representation and normality, as well as the bias and variance for the estimator $\widehat{S}_{0,h_1,h_2}^{c}(t \mid x)$ the same assumptions are made together with Assumptions 9-10 which consider the speed of the convergence to 0 for the bandwidth h. Theorems 1-2 in [11] provide the asymptotic representation and the asymptotic normality, respectively. Furthermore, Proposition 1 expresses the bias and variance of $\widehat{S}_{0,h_1,h_2}^{c}(t \mid x)$ as functions of t and x.

An investigation on the effect of ignoring the cure status is done which compares this estimator with the corresponding estimator proposed in Section 3.2 (Lopez et al. 2017b). The conclusion was that it was not straightforward to evaluate the exact gain of bias when considering the cure status information in [11].

3.3 Limitations

Some limitations were made when choosing what methods to use for this project. Firstly, only mixture cure models were considered for fitting the data, while non-mixture cure models were not explored. Non-mixture models were mainly introduced for cancer research and follows the pattern of survival for cancer cells. Additionally, the concept of relative survival, discussed in Chapter 2, was not incorporated into the methods we eventually used. This omission can be motivated to the relatively short follow-up period of up to 138 days. Since the relative survival does not exhibit significant differences across age groups within such a limited timeframe, their inclusion would likely have a minimal impact on the results.

4

Results

This chapter will provide a thorough description of the dataset of the Master's thesis as well as the models created considering the age or gender of the patients. We refer to Appendix A for further results and Appendix B for the R code of the implementation.

4.1 The dataset

The dataset used in this study comprises a total of 2074 hospitalized patients from the south metropolitan area of Barcelona. The data collection took place after the first wave of the pandemic, specifically from the months of March and April 2020. It is important to note that the data was not collected in real-time alongside the event of the patients. Consequently, the end of the study is given by the last recorded time in the dataset.

4.1.1 Variable description

The dataset is organized in the form of a dataframe, consisting of eight variables. A description of these variables is provided in the following table:

Variable	**Description**
id	The ID of the patient.
time	The time in days to either death or discharge.
death	1 = Death, 0 = Discharge
sex	1 = Man, 2 = Woman
cvasc	1 = No cardiacvascular history, 0 = Cardivascular history
age	The age of the patient in years
charlson	The patient's Charlson Comorbity Index
safi	The patient's SaFi ratio.

Table 4.1: Descriptions of the variables in the dataset.

Each patient in the dataset is assigned a unique identifier, referred to as the patient's ID. The variable "time" denotes the duration, measured in days, that the patient spent in the hospital until either being discharged or experiencing death.

The variable "death" is an indicator, taking a value of 1 if the patient passed away during his/her hospital stay, and 0 if they were successfully discharged.

The variable "sex" indicates the gender of the patient, where a value of 1 corresponds to male patients and a value of 2 represents female patients. Additionally, the variable "cvasc" provides information regarding the patient's cardiovascular history, with a binary value of either 0 (indicating no cardiovascular history) or 1 (indicating that the patient has a history of cardiovascular conditions).

The patient's age at the time of admission to the hospital is reperesented by the variable "age". Furthermore, the variable "charlson" represents the Charlson comorbidity index which is a weighted measure used to predict the risk of death within one year of hospitalization for patients with specific comorbidity conditions. It is categorized into different levels, where lower values indicate a lower risk of mortality and higher values indicate a higher risk.

Lastly, the variable "safi" represents the patient's SaFi ratio. The SaFi ratio is calculated as the ratio between the oxygen saturation (SpO2), typically ranging from 95% to 100%, and the fraction of inhaled oxygen (FiO2) for a person, which is 0.21 at atmospheric air. The values of "safi" generally fall within the interval [300, 476] mmHg, with higher values indicating a better overall condition for the patient.

Table 4.2 presents descriptive statistics for the minimum, maximum, and median values based on age quantiles. The age quantiles are generated using the 25%, 50%, and 75% quartiles to ensure an equal number of observations in each category. The table considers the group of individuals who died. Similarly, Table 4.3 provides the same descriptive statistics, but for the group of patients who were discharged.

Quantile	**Time**			**Gender**		**Age**		
	Min	**Max**	**Median**	**Man**	**Woman**	**Min**	**Max**	**Median**
1	4	46	15	6	2	34	49	44
2	1	55	16.5	21	5	50	59	57
3	1	58	16.5	41	19	60	69	65
4	1	74	9	74	50	70	96	77

Table 4.2: Summary of the survival time, gender and age distribution for the *death* group per age quantile.

Quantile	**Time**			**Gender**		**Age**		
	Min	**Max**	**Median**	**Man**	**Woman**	**Min**	**Max**	**Median**
1	1	131	7	318	208	19	49	42
2	1	120	8	303	180	50	59	55
3	1	110	9	348	214	60	69	54
4	1	138	19	210	177	70	96	74

Table 4.3: Summary of survival time, gender and age distribution for the *discharge* group per age quantile.

4.1.2 Preparation of the data

The method for estimation of the cure fraction for data with known cures is based on three groups; the uncensored which experienced the event, the known to be cured and the censored patients which has neither died nor been cured before the follow-up-time. Recall that the observations are denoted

$$\{(T_i, \delta_i, \boldsymbol{X}_i, \xi_i \nu_i) : i = 1, ..., n\},$$

where δ_i the indicator of death, ν_i the indicator of cure and ξ_i the indicator of known cure status. Due to the specific format of our data where the data is collected after the patients observation period ended, the dataset described in Table 4.1 consists solely of two groups; the patients that were discharged and the patients who died. The death group will be modeled as the uncensored group while the discharged will be modeled as the known cures. Hence, the two groups present in the data are represented as:

1. $(T_i, \delta_i = 1, \boldsymbol{X}_i, \xi_i \nu_i = 0)$ - The subjects who died (uncensored).
2. $(T_i, \delta_i = 0, \boldsymbol{X}_i, \xi_i \nu_i = 1)$ - The subjects who were discharged (known cures).

In order to apply the methodology presented in Section 3.2.4 for constructing models to predict the cure rate, we needed to incorporate an artificial follow-up time. This involved assuming that the study concluded after a specified duration of k days. Consequently, observations within the timeframe of k days retained their respective outcomes, while observations beyond this threshold were treated as censored individuals. Regardless of the previous outcome, all individuals who surpassed the follow-up time of k were treated as censored. Applying this artificial follow-up time we end up with the third group the method requests:

3. $(T_i, \delta_i = 0, \boldsymbol{X}_i, \xi_i \nu_i = 0)$ - The subjects who neither died or was discharged before follow-up time k (censored).

A vector of follow-up times was used in the analysis which consisted of thirteen distinct time points. Hence $k \in \{20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80\}$, where the chosen follow-up times were chosen to examine the effect of varying time intervals on the results.

4.2 The models

This section includes three different models, two incorporating only the covariate age and one the covariate gender. They were built by applying the methods explained in Section 3.2.4 with the an extension of the `npcure` package. Models created for the cardiovascular history, Charlson comorbidity index and for the SaFi ratio are presented in Appendix A.

4.2.1 A model incorporating gender as a covariate

First of all, we present a model that takes into account the gender of the patients. Figure 4.1 displays the estimated cure rates for women and men at thirteen different follow-up times. Additionally, the true cure rates for each gender are provided to facilitate result evaluation. An initial conclusion from the plot is that the women have a higher cure rate than the men. Moreover, we see that the larger follow-up time the closer the estimate gets to the true cure rate since the model has been provided more information from the data. The estimated cure rate converges to the true cure rate at follow-up time of 45 days for the women and 60 days for the men.

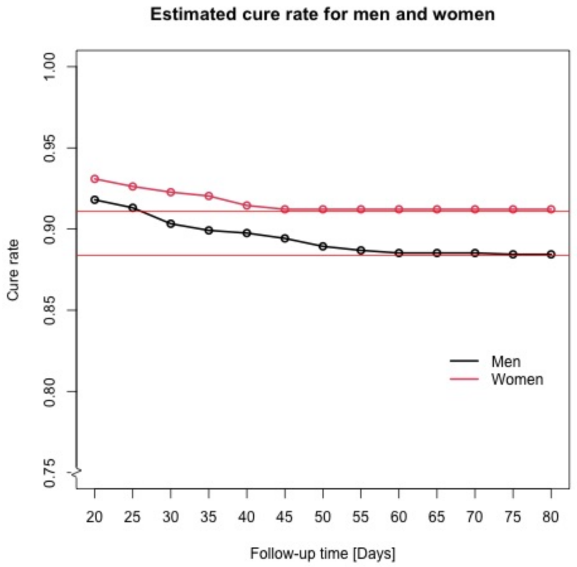
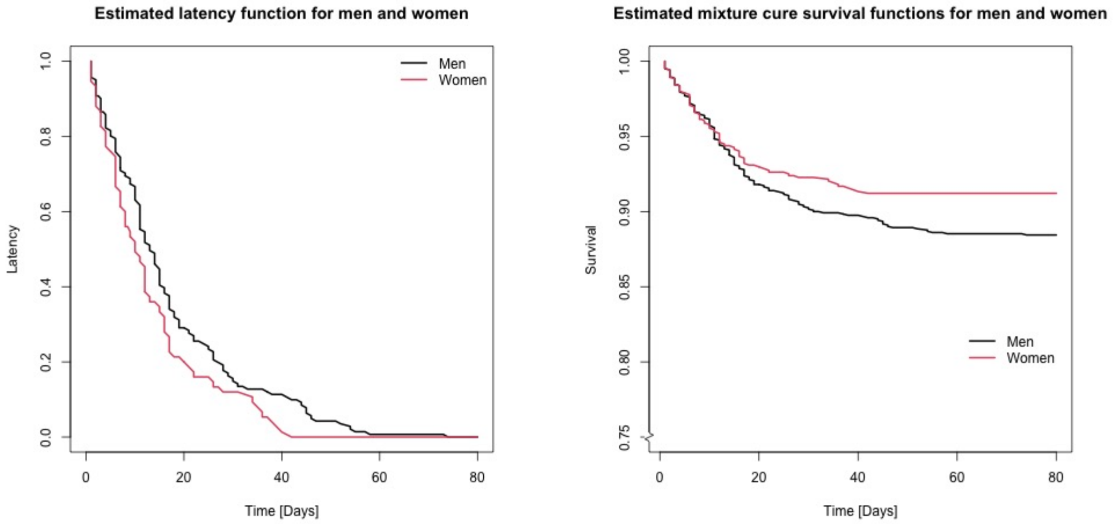


Figure 4.1: Estimated cure rates for men and women with respect to the follow-up times.



(a)

(b)

Figure 4.2: Estimated latencies and mixture cure survival functions for men and women at a follow-up time of 80 days.

Furthermore, the latency for the genders are presented in Figure 4.2a. It is crucial to note that the latency, which is displayed in Figure 4.2a, represents the survival pattern exclusively for the susceptible individuals. In this context, the term "susceptible" refers to individuals who are at risk of experiencing the event of interest, hence the proportion of the population who eventually will die in the hospital. Furthermore, the latency curve shows a clear difference between the genders, with the curve for women consistently positioned below that of men. This disparity suggests that the survival probabilities for females are slightly lower compared to males within the susceptible population.

In Figure 4.2b, we explore the mixture cure survival functions for both genders. Notably, the curves exhibit a plateau around 50 days, indicating the estimated cure rate for respective gender. Moreover, we notice that the female curve is steadily above the male curve, implying that the overall survival is higher amongst the women.

4.2.2 A model incorporating age as covariate

Secondly, a model incorporating age as a covariate was developed. To account for the continuity of the age variable and to have enough values to construct an accurate model, the values were categorized into quantiles as explained earlier. The table below presents the values for each respective quantile:

Quantile	Min	Max
1	19	49
2	50	59
3	60	69
4	70	96

Table 4.4: The age of values for each quantile.

Figure 4.3 presents the resulting plot of the estimated cure rates along with the true values of the respective cure rate. The analysis reveals that the first quantile exhibits a true cure rate close to 1, while the fourth quantile shows a rate around 0.75. Similar to the gender-based model, this pattern indicates that the estimated cure rates approach the true values as the follow-up period becomes sufficiently long. Notably, all four groups converge to the cure rate within the range of 50 to 60 days of follow-up time.

Furthermore, a plot depicting the latency for the different age quantiles was generated and is presented in Figure 4.4a. The latencies appear to be quite similar for the groups with the fourth quantile almost constantly staying below the other curves. However, an observation is that the latency function for the first quantile appears less smooth compared to the other three. It exhibits a staircase-like pattern with fewer and larger steps, indicating a lower number of deceased individuals in the first quantile group. This conclusion agrees with the high cure rate exhibited in Figure 4.3.

Next, we examine the mixture cure survival functions associated with the quantiles illustrated in Figure 4.4b. All the curves exhibit characteristics of improper survival

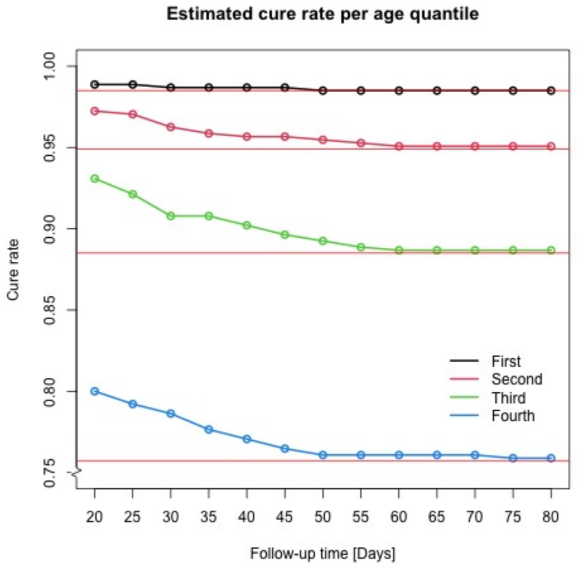
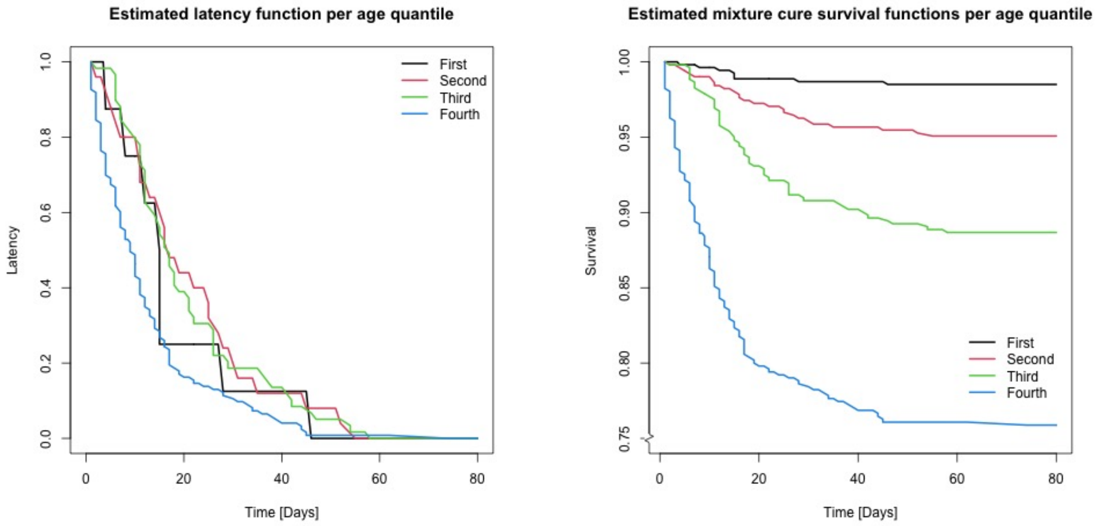


Figure 4.3: Estimated cure rates per age quantile with respect to the follow-up times.



(a)

(b)

Figure 4.4: Estimated latencies and mixture cure survival functions per age quantile at a follow-up time of 80 days.

functions, with each curve gradually reaching a plateau, indicating a potential cure around 50 days. This finding aligns with our earlier observation of convergence in follow-up time, as depicted in Figure 4.3. Notably, the fourth quantile displays a notably lower curve, ultimately converging to a value of 0.75. This figure serves as evidence that the age of the patients influences the resulting survival outcomes in the study.

Additionally, we conducted a model to evaluate the cure rate for the specific ages 39, 59, 64, and 76 years. Here we estimate the conditional survival curve for the fixed ages using the estimator in Equation (3.12) together with the bootstrap bandwidths explained in Section 3.2.4.1. Due to the complexity of this algorithm this simulation were very computationally heavy.

These particular ages were selected based on their frequency in the dataset, in order to achieve the most accurate result. However, as depicted in Figure 4.5, the model does not exhibit as good of a fit as the model incorporating the four age quantiles. For example, while the true cure rate for patients at age 39 was 1, the estimated cure rate falls slightly short at approximately 0.95. Similarly, the estimated cure rate for age 76 is underestimated, while the resulting cure rate for age 64 is the only estimate that converges to the true value after a sufficient follow-up time.

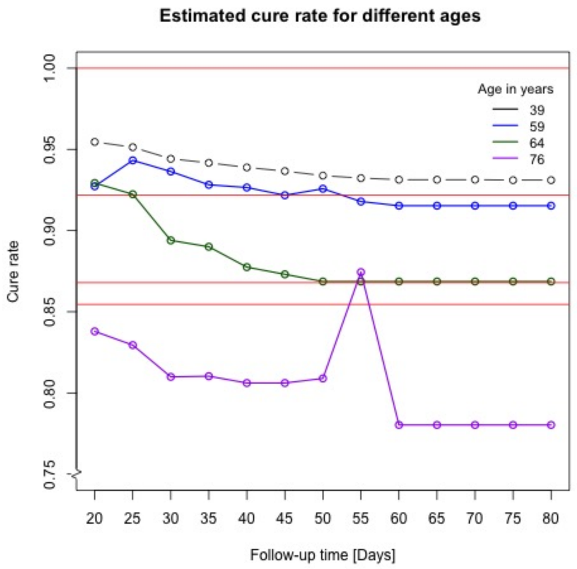
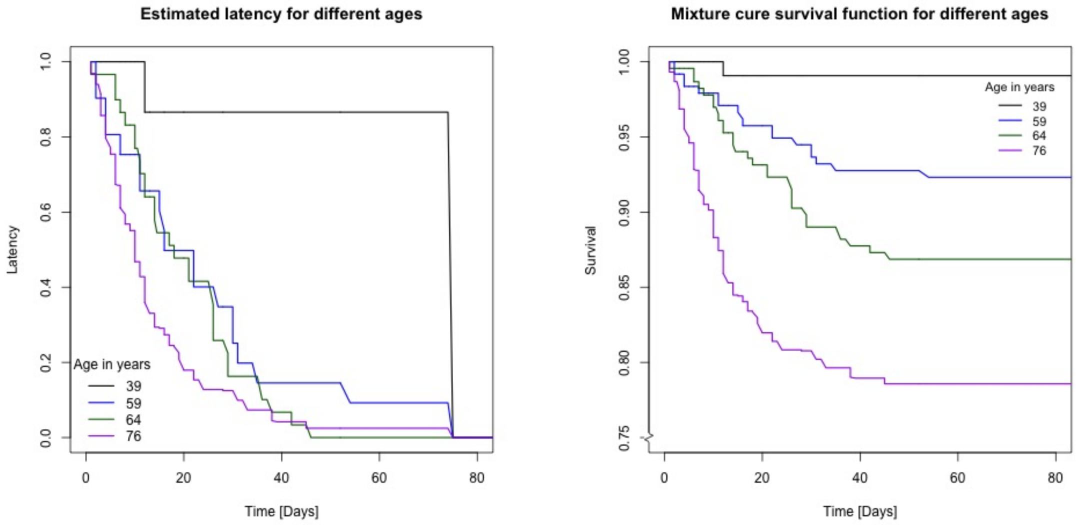


Figure 4.5: Estimated cure rates for the ages of 39, 59, 64 and 76 years with respect to the follow-up times.

Furthermore, when examining the latency plot shown in Figure 4.6a, we observe a similarity to the corresponding plot for the quantiles in Figure 4.4a for most ages. However, there is a noticeable difference in the latency pattern for the age of 39. This deviation can likely be attributed to the limited number of observations available before 80 days for this age group.

Lastly, the mixture cure survival functions plotted in Figure 4.6b all appear as improper functions, similar to the previous models. One possible reason for this outcome, when previous plots in this model has shown inaccurate results, is that it also incorporate patients from the discharged group making the sample size larger than for the latency. This discrepancy raises the question of why the estimates of the cure rates were bad, considering that the discharged group is taken into account in Figure 4.3 as well. The investigation and analysis of this issue will be discussed in detail in Section 5.



(a)

(b)

Figure 4.6: Estimated latencies and mixture cure survival functions for the ages of 39, 59, 64 and 76 years at a follow-up time of 80 days.

4.2.3 Model predictions for follow-up time 30 days

In this section, we present the estimates for a follow-up time of 30 days to assess the feasibility of predicting patient outcomes with a commonly used follow-up time of a month. To evaluate the accuracy of these predictions, the estimated cure rates are compared with the true cure rates for each quantile. The differences between the estimated and the true cure rates are summarized in Table 4.5 for the gender-based model and Table 4.6 for the age-based model. We see that the differences in cure rates between genders are relatively consistent, while there is more variation observed across the different age quantiles.

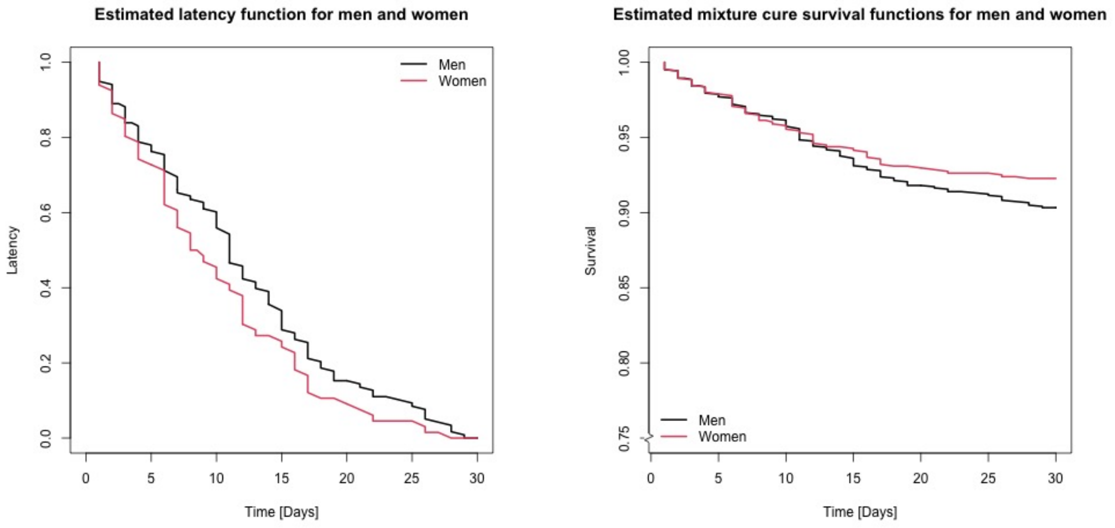
Gender	$(1-\hat{\pi})$	$(1-\pi)$	Difference
Women	0.922	0.911	1.273 %
Men	0.903	0.884	2.215 %

Table 4.5: Estimated cure rates for men and women at a follow-up time of 30 days.

Quantile	$(1-\hat{\pi})$	$(1-\pi)$	Difference
1	0.985	0.987	0.19 %
2	0.949	0.963	1.42 %
3	0.885	0.908	2.51 %
4	0.757	0.786	3.68 %

Table 4.6: Estimated cure rates per age quantile at a follow-up time of 30 days.

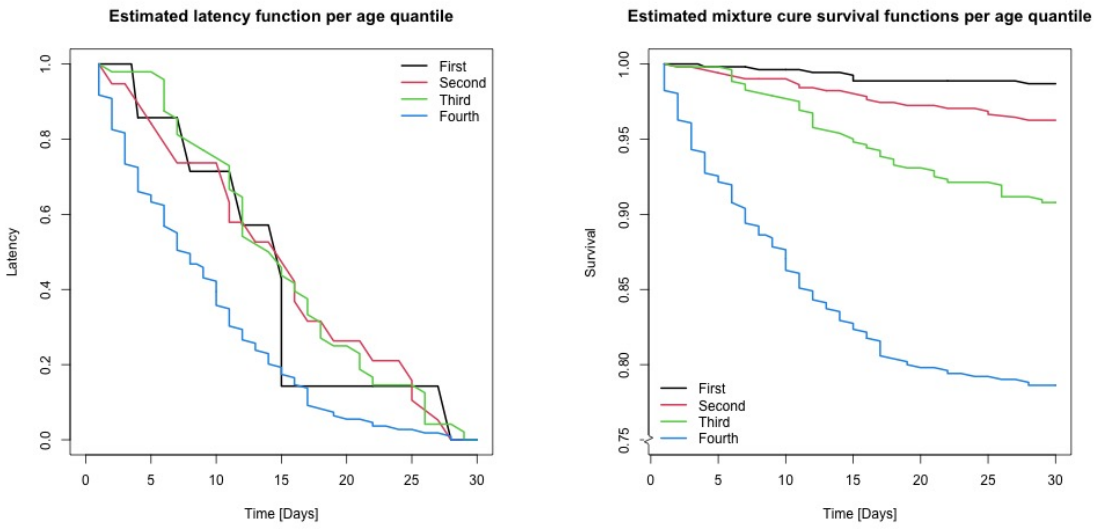
The latencies depicted in Figure 4.7a and 4.8a exhibit a noticeable difference compared to the previously analyzed latencies as they lack the tail of the curves. Similarly, in the mixture cure survival functions shown in Figure 4.7b and 4.8b, it is challenging to identify a clear plateau in any curve, which makes it difficult to determine the presence of a cured proportion in the data. A detailed discussion of these results will be presented in the subsequent chapter.



(a)

(b)

Figure 4.7: Estimated latencies and mixture cure survival functions for men and women at a follow-up time of 30 days.



(a)

(b)

Figure 4.8: Estimated latencies and mixture cure survival functions per age quantile at a follow-up time of 30 days.

5

Discussion

In this chapter, we will perform a comprehensive analysis of the methods and results presented in this report. Our examination will begin by discussing the outcomes outlined in Chapter 4. This will be followed with an explanation of the limitations identified in Chapter 3 and possible future applications.

5.1 Discussion of the results

Due to the specific format of the dataset of the study, several of the mentioned estimation methods in Chapter 3 did not provide any accurate results. This limitation arose because these methods were not designed to handle cases where cure was observed. Instead, individuals who were discharged from the study had to be treated as right-censored, even though the true outcome was already known. This introduced some bias to the estimation since these models treat them as observations who could either die or get discharged when this was not actually a possibility.

However, thanks to the recently presented methods (Safari et al. 2023) described in Section 3.2.4 a method more suitable for our kind of data was available. As explained in Chapter 4 an artificial follow-up time was set to investigate whether the model could predict the cure rate at different values of the follow-up time. The model presented estimates quite close to the true cure rate. For a larger follow-up time more information is available since more people either experienced the event or was cured (discharged), implying a more accurate estimate.

It is worth noting that as the cure rate decreases, meaning a higher number of deaths, the precision of the estimate reduces. This raises the question of whether the method performs worse when there are fewer known cases of cure. In the dataset analyzed for this thesis, the number of individuals who were cured is relatively large, with the lowest observed cure rate being approximately 75 %. An idea of a future work is to see how the estimator changes with the amount of known cures in the dataset.

We compare the estimated latency plots for the age quantiles at two follow-up times: 80 days (Figure 4.4a) and 30 days (Figure 4.8a). It is important to note that these plots only consider subjects who have died and not the cured individuals, resulting in a significantly smaller number of observations at the follow-up time of 30 days. Specifically, there were 218 individuals who died before 81 days, whereas only 188 individuals died before the 31st day. It is also interesting to compare the two mixture cure survival functions of the two follow-up times. In the plot in Figure 4.4b the plateau that follows from the cured individuals are obvious and clearly converges to the true cure rate. However, it is challenging to determine whether the survival curves in Figure 4.8b will eventually reach plateaus, indicating the presence of cured individuals. Therefore, it is not appropriate to draw conclusions about the existence of cured individuals in the data only based on these graphs.

After examining Table 4.5 and 4.6, it can be concluded that the accuracy of the estimated cure rate appears to be remarkably high. This leads us to believe that despite the absence of visibly cured individuals in the mixture cure survival functions, a reliable estimation of the cure rate after 30 days can still be obtained. However, determining whether this is sufficient evidence to assert that the models are already effective after just 30 days is difficult to determine.

The method proposed in [11] offered two kinds of estimations: one for the estimation of the unconditional survival (from Equation (3.13)), which we applied for all categorical variables, and one for the conditional survival function. The conditional survival curve (from Equation (3.12)) was estimated for four specific ages within the continuous variable, with a sufficiently large sample size. Both methods were applied to analyze the effects of age on the outcome. The result from the unconditional estimation of the cure rate was presented in Figure 4.5. The estimate for the specific ages was not as accurate compared to the estimation based on the age quantiles. Specifically, the estimated cure rate for the age of 39 years significantly deviated from the true value, a cure rate of one which implied that 100 % of the patients got discharged. For the age of 59 years, the estimate remained close to the true value across all follow-up times and the estimate for the age of 64 years converged to the true value after approximately 50 days. At last, for the age of 76 years the estimate was notably inaccurate with a strange behavior of the curve for the cure rate. This discrepancy in estimation accuracy for the youngest and oldest age groups can likely be assigned to the limited sample size, which may not have provided sufficient data for an accurate estimation in these cases.

The behavior of the conditional estimator regarding the estimated cure rate is not consistently decreasing, unlike the cure rates from the unconditional method. The reason behind this behavior remains uncertain. It is worth noting that the model is presented with a larger amount of data for longer follow-up times which could potentially influence the estimated cure rate. However, further investigation is required to fully understand the factors contributing to this deviation. Hence, another question that came up was how the conditional estimator of the ages would behave if a larger sample-size was given. Would the estimates be more accurate or would the unconditional estimation still provide a better result?

5.2 Limitations and future work

After doing a thorough scope review within the broad area of cure models some limitations were made. The primary focus of this thesis is to examine the suitability of mixture cure models for estimating the cure rate and understanding the population dynamics in the Covid-19 dataset. In Chapter 2, an extension of the mixture cure models was introduced, incorporating background survival. However, due to the short survival times observed in the dataset, this aspect was not taken into account during model development. This could have been a factor to consider if the dataset consisted of for instance long-term Covid patients and the objective was to estimate the cure rate for these patients.

Similar reasoning applies to the theory presented regarding non-mixture cure models. These models are specifically designed for modeling cancer and are customized to capture the biological features of cancer cell growth. Nevertheless, a potential topic for future research would be to explore whether these models can provide insights in understanding the complex nature of the Covid-19 disease.

The parametric and semiparametric estimation methods for both the incidence and latency were presented in Chapter 3. However, these also only consider the two groups; the susceptible subpopulation and the people not experiencing the event. Thus, for a dataset where a cured proportion is suspected, but not known, due to the large amount of right-censored observations these estimation methods could be considered.

6 Conclusion

The main goal of this Master's thesis was to explore whether cure models could help hospitals plan and manage resources during similar crises as the recent pandemic. There is an abundance of data after the last pandemic which can be used to apply appropriate models and analyze the respective outcome. Hence, these investigations could lead to a better preparedness and broader knowledge with several areas concerning pandemic crises. In response to the question posed in the introduction on how statistical models derived from Covid-19 data can provide valuable knowledge to prevent shortages of critical resources in potential future pandemics or similar situations, the results obtained from fitting Covid-19 data to cure models offer insightful answers.

First of all, we can conduct that the cure rate can be estimated with a high degree of accuracy, if the follow-up time is sufficiently large. This enables us to predict the percentage of the patients discharged from the hospital. Additionally, the estimation of latency, or the predicted time until death given that the patient eventually will die, provides information about the behavior of progress of the disease as well as it provides a prediction of time to death for the uncured population.

Therefore, we can conclude that Covid-19 data can be effectively fitted to cure models. This thesis has examined models incorporating individuals which are known to be cured, applying methodologies that have recently been introduced. Given that the known cures in this context were represented by discharged individuals, distinguishing a cure was straightforward. However, it is crucial to note that the conclusion regarding the applicability of these models to Covid-19 data does not imply that individuals are actually cured from Covid-19.

The topic of cure models is extensive and continuously evolving, with ongoing research in this field. This thesis provides evidence for the applicability of cure models in analyzing Covid-19 data, opening up for several questions and a large curiosity about modeling different types of Covid-19 data using various cure models. The findings presented here pave the way for future investigations and advancements in understanding and modeling the dynamics of Covid-19 and its potential implications.

Bibliography

[1] M. Marani et al. "Intensity and frequency of extreme novel epidemics". In: *PNAS* 118.35 (2021). DOI: 10.1073/pnas.2105482118. URL: https://www.pnas.org/doi/10.1073/pnas.2105482118.

[2] R. Maller and X. Zhou. *Survival analysis with long-term survival.* Wiley series in probability and statistics. John Wiley & Sons Ltd, 1996. ISBN: 0471962015.

[3] P. Lambert. "Modeling the cure fraction in survival studies". In: *Stata Journal* 7.3 (2007), pp. 351–375. DOI: 10.1177/1536867X0700700304. URL: https://www.researchgate.net/publication/24096759_Modeling_the_cure_fraction_in_survival_studies.

[4] D. H. Kutal and L. Qian. "A Non-Mixture Cure Model for Right-Censored Data with Fréchet Distribution". In: *Stats* 1.1 (2018), pp. 176–188. DOI: 10.3390/stats1010013. URL: https://www.mdpi.com/2571-905X/1/1/13.

[5] M. Cvancarova et al. "Proportion cured models applied to 23 cancer sites in Norway". In: *Int J Cancer* 132.7 (2013), p. 1700. DOI: 10.1002/ijc.27802. URL: https://pubmed.ncbi.nlm.nih.gov/22927104/.

[6] Yifan Peng and Binbing Yu. *Cure Models: Methods, Applications, and Implementation.* 1st. Chapman and Hall/CRC, 2021. URL: https://doi.org/10.1201/9780429032301.

[7] A. López-Cheda, A. Jácome, and R. Cao. "Nonparametric latency estimation for mixture cure models". In: *TEST* 26.3 (2017). DOI: 10.1007/s11749-016-0515-1. URL: https://link.springer.com/article/10.1007/s11749-016-0515-1.

[8] R. Peláez et al. "Nonparametric estimation of the conditional survival function with double smoothing". In: *Journal of Nonparametric Statistics* 34.4 (2022), pp. 1063–1090. DOI: 10.1080/10485252.2022.2102631. URL: https://doi.org/10.1080/10485252.2022.2102631.

[9] A. López-Cheda, M. Jácome, and I. López-de-Ullibarri. "npcure: An R Package for Nonparametric Inference in Mixture Cure Models". In: *The R journal* 13.1 (2021). DOI: 10.32614/rj-2021-027. URL: https://ruc.udc.es/dspace/handle/2183/29290.

[10] W. C. Safari, I. López-de-Ullibarri, and M. A. Jácome. "A product-limit estimator of the conditional survival function when cure status is partially known". In: *Biom J* 63.5 (2021), pp. 984–1005. DOI: 10.1002/bimj.202000173. URL: https://pubmed.ncbi.nlm.nih.gov/33646606/.

[11] W. C. Safari, I. López-de-Ullibarri, and M. A. Jácome. "Latency function estimation under the mixture cure model when the cure status is available".

In: *Lifetime data analysis* (2023). DOI: `10.1007/s10985-023-09591-x`. URL: `https://link.springer.com/article/10.1007/s10985-023-09591-x`.

A

Further results

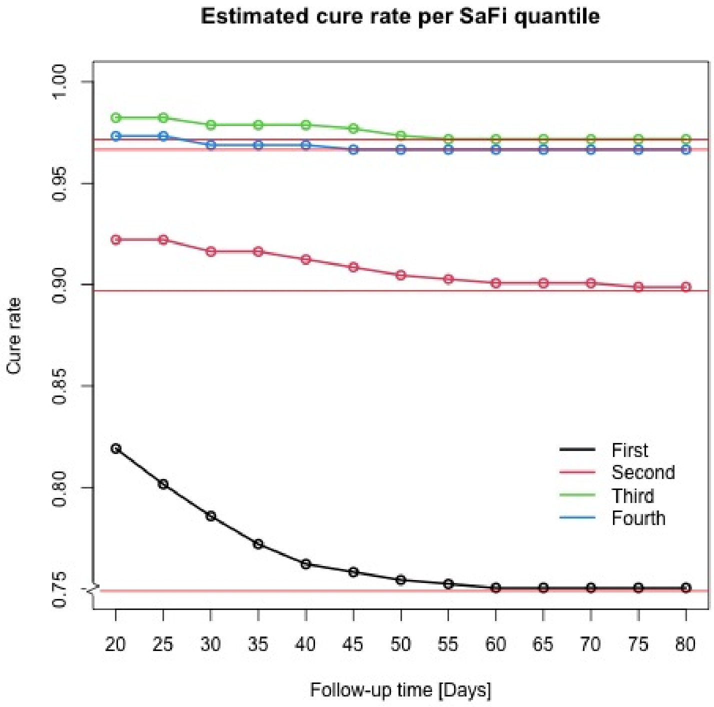
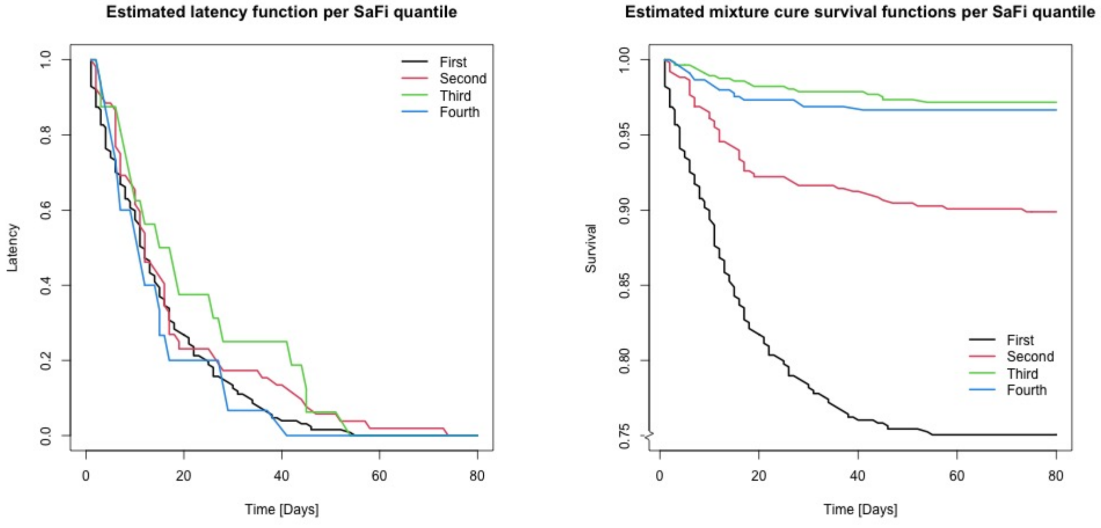


Figure A.1: Estimated cure rates per SaFi quantile with respect to the follow-up times.

SaFi ratio	$(1-\hat{\pi})$	$(1-\pi)$	Difference
100	0.688	1	31.2 %
200	0.734	1	26.6 %
300	0.653	0.6	8.83 %
400	0.863	0.842	2.43 %

Table A.1: Estimated cure rates for the SaFi ratios of 100, 200, 300 and 400 at a follow-up time of 30 days.



(a) **(b)**

Figure A.2: Estimated latencies and mixture cure survival functions per SaFi quantile at a follow-up time of 80 days.

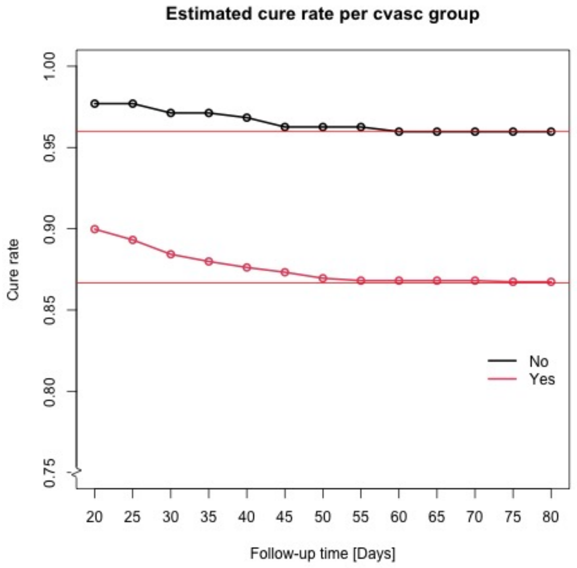
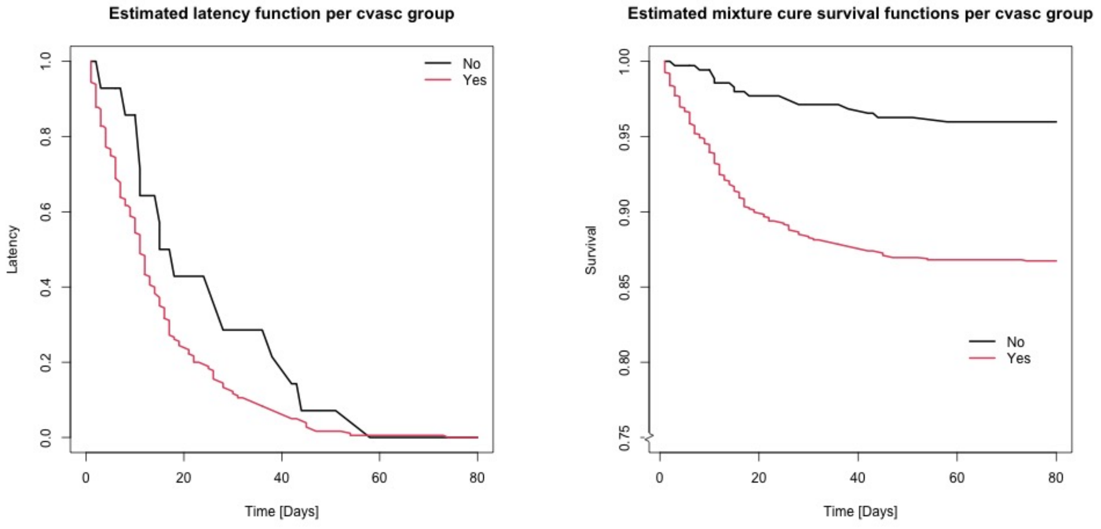


Figure A.3: Estimated cure rates per cardiovascular history group with respect to the follow-up times.



(a) **(b)**

Figure A.4: Estimated latencies and mixture cure survival functions per cardiovascular history group at a follow-up time of 80 days.

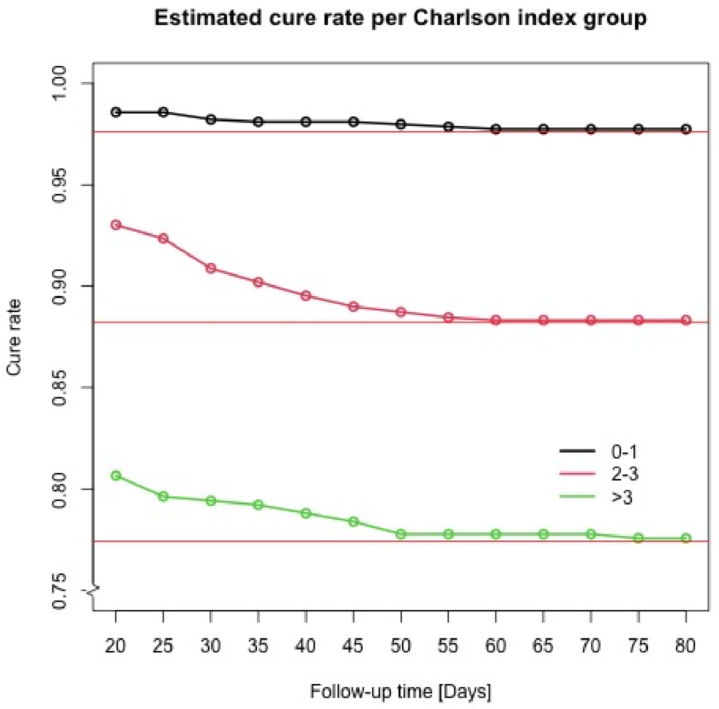
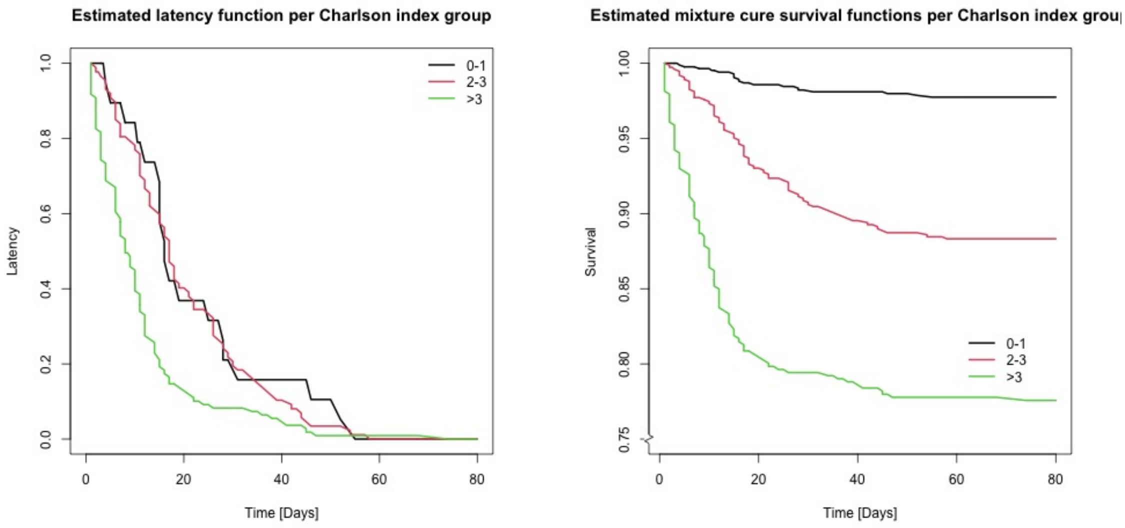


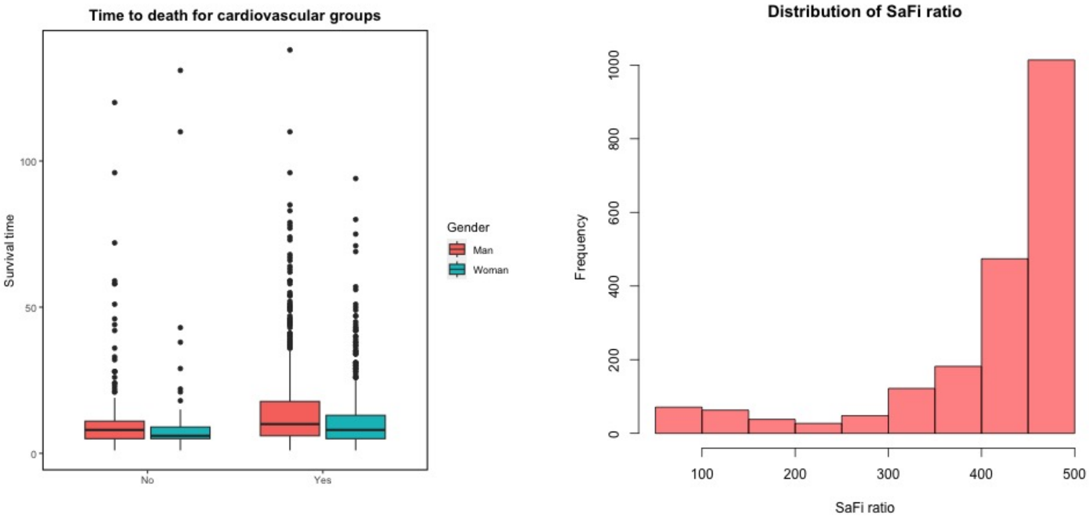
Figure A.5: Estimated cure rates per Charlson index group with respect to the follow-up times.



(a)

(b)

Figure A.6: Estimated latencies and mixture cure survival functions per Charlson index group at a follow-up time of 80 days.



(a)

(b)

Figure A.7: Description of distribution of the variables cardiovascular history and SaFi ration.

B

R code

B.1 Packages

```r
library(survival)
library(dplyr)
library(npcure)
library(tidyr)
library(plotrix)
library(jpeg)
library(data.table)
library(DescTools)
library(MASS)
library(microbenchmark)
library(foreach)
library(doParallel)
library(doRNG)
library(doSNOW)
library(readr)
```

B.2 Data preparation

```r
load('CureModelDataCovs.RData')
dataCMod <- dataCMod %>% mutate_if(is.character, as.factor)

data_noCnans <-dataCMod %>% drop_na(cvasc)

dataCMod$charlson_group <- cut(dataCMod$charlson, breaks = c
   (0,2,4,13), include.lowest = TRUE, right = FALSE)
levels(dataCMod$charlson_group) <- c("0-1", "2-3", ">3")

data_noSnans <-dataCMod %>% drop_na(safi)
data_noSnans$safi_cut <- cut(data_noSnans$safi,
                             breaks=c(quantile(data_noSnans$safi,
                                probs = seq(0, 1, by = 0.25))),
                             labels=c('1st', '2nd', '3rd', '4th'),
                                include.lowest = TRUE)

dataCMod$age_group <- cut(dataCMod$age,
                          breaks=c(quantile(dataCMod$age, probs =
                             seq(0, 1, by = 0.25))),
                              labels=c('1st', '2nd', '3rd', '4th'),
                                 include.lowest = TRUE)

# Calculating true cure rate
true_cure_rate_men= sum(dataCMod$sex == "Man" & dataCMod$death ==
   0)/sum(dataCMod$sex == "Man")
true_cure_rate_women = sum(dataCMod$sex == "Woman" & dataCMod$death
    == 0)/sum(dataCMod$sex == "Woman")
true_cure_rate_cvasc= sum(data_noCnans$cvasc == "Yes" & data_
   noCnans$death == 0)/sum(data_noCnans$cvasc == "Yes")
true_cure_rate_nocvasc = sum(data_noCnans$cvasc == "No" & data_
   noCnans$death == 0)/sum(data_noCnans$cvasc == "No")
true_cure_rate_1 = sum(dataCMod$charlson_group == "0-1" & dataCMod$
   death == 0)/sum(dataCMod$charlson_group == "0-1")
true_cure_rate_2 = sum(dataCMod$charlson_group == "2-3" & dataCMod$
   death == 0)/sum(dataCMod$charlson_group == "2-3")
true_cure_rate_3 = sum(dataCMod$charlson_group == ">3" & dataCMod$
   death == 0)/sum(dataCMod$charlson_group == ">3")

true_cure_rate_1q = sum(data_noSnans$safi_cut == "1st" & data_
   noSnans$death == 0)/sum(data_noSnans$safi_cut == "1st")
true_cure_rate_2q = sum(data_noSnans$safi_cut == "2nd" & data_
   noSnans$death== 0)/sum(data_noSnans$safi_cut == "2nd")
true_cure_rate_3q = sum(data_noSnans$safi_cut == "3rd" & data_
   noSnans$death == 0)/sum(data_noSnans$safi_cut == "3rd")
true_cure_rate_4q = sum(data_noSnans$safi_cut == "4th" & data_
   noSnans$death == 0)/sum(data_noSnans$safi_cut == "4th")

true_cure_rate_1qa = sum(dataCMod$age_group == "1st" & dataCMod$
   death == 0)/sum(dataCMod$age_group == "1st")
true_cure_rate_2qa = sum(dataCMod$age_group == "2nd" & dataCMod$
   death== 0)/sum(dataCMod$age_group == "2nd")
true_cure_rate_3qa = sum(dataCMod$age_group == "3rd" & dataCMod$
   death == 0)/sum(dataCMod$age_group == "3rd")
true_cure_rate_4qa = sum(dataCMod$age_group == "4th" & dataCMod$
   death == 0)/sum(dataCMod$age_group == "4th")
```

B.3 Function for computing cure rate, survival functions, and latency

```
compute_cure_survival <- function(variable, follow_up, data) {
    S_list <- vector(mode = 'list', length = length(levels(data[[
       variable]])))
    p_list <- time_list <- S0_list <- vector(mode = 'list', length
       = length(levels(data[[variable]])))
    cure_rate_list <- vector(mode = 'list', length = length(levels(
       data[[variable]])))
    count <- 0

    for (i in follow_up) {
        count <- count + 1
        temp_data <- data
        temp_data$death[temp_data$death == 1 & temp_data$futime > (
           i-1)] <- 0
        temp_data$futime[temp_data$futime > (i-1)] <- i
        temp_data$knowncure <- with(temp_data, ifelse(death == 0 &
           futime <=    (i-1), 1, 0))

    # Extract the dataframe for the unconditional estimation
    dfr1 <- temp_data[, c(variable, "futime", "death", "knowncure")
       ]
    for (j in 1:length(levels(temp_data[[variable]]))) {
        S_list[[j]] <- survfitcurePK_un(dataset = dfr1[dfr1[[
           variable]] == levels(temp_data[[variable]])[j], 2:4])
        # p: probability of experiencing the final outcome
        p_list[[j]] <- S_list[[j]][[2]]
        # Time
        time_list[[j]] <- S_list[[j]][[3]]

        # S0(t): Survival function of the individuals experiencing
           the event
        S0_list[[j]] <- ((S_list[[j]][[1]] - (1 - p_list[[j]])) / p
           _list[[j]])
        cure_rate_list[[j]][count] <- 1 - p_list[[j]][1]
    }
  }

  return(list(time_list = time_list, p_list = p_list, S_list = S_
     list, S0_list = S0_list, cure_rate_list = cure_rate_list))
}
```

B.4 Function for generating plots

```
generate_plots <- function(variable_name, cure_rate_list, true_cure
   _rates, S_list, S0_list) {

    legend_labels <- switch(variable_name,
                age_group = c("First", "Second", "Third", "Fourth")
                    ,
                sex = c("Men", "Women"),
                charlson_group = c("0-1", "2-3", ">3"),
                cvasc = c("No", "Yes"),
                safi_cut = c("First", "Second", "Third", "Fourth")
    )

    title <- switch(variable_name,
                age_group = "Estimated cure rate per age quantile",
                sex = "Estimated cure rate for men and women",
                charlson_group = "Estimated cure rate per Charlson
                   index group",
                cvasc = "Estimated cure rate per cvasc group",
                safi_cut = "Estimated cure rate per SaFi quantile"
  )

  # Plot cure rate
  jpeg(paste0(variable_name, "_cure.jpeg"))
```

```
plot(follow_up, cure_rate_list[[1]], type = "o", main = title,
      ylab = "Cure rate", xlab = "Follow-up time [Days]", ylim =
         c(0.75, 1), cex=1, lwd = 2, xaxt = "n")
abline(h = true_cure_rates[1], col = "red")
axis(1, at = follow_up, labels = follow_up)
for (i in 2:length(cure_rate_list)) {
      lines(follow_up, cure_rate_list[[i]], col = i, type = "o",
         lwd = 2)
      abline(h = true_cure_rates[i], col = "red")
}
legend(70,0.83, legend = legend_labels, bty = "n",
       col = 1:length(cure_rate_list), lty = 1, cex = 1, lwd = 2)
axis.break(axis = 2, breakpos = 0.75, style = "zigzag")
dev.off()

title2 <- switch(variable_name,
                age_group = "Estimated latency function per age
                   quantile",
                sex = "Estimated latency function for men and
                   women",
                charlson_group = "Estimated latency function per
                   Charlson index group",
                cvasc = "Estimated latency function per cvasc
                   group",
                safi_cut = "Estimated latency function per SaFi
                   quantile"
)
# Plot latency
jpeg(paste0(variable_name, "_latency.jpeg"))
par(mar = c(5, 5, 4, 3))
plot(S_list[[1]][[3]], S0_list[[1]], type = "l", xlab = "Time [
   Days]",            ylab = "Latency", main = title2, col = "
   black", lwd = 2, xlim =             c(0, 80))
for (i in 2:length(S_list)) {
      lines(S_list[[i]][[3]], S0_list[[i]], type = "l", col = i,
         lwd = 2)
}
legend("topright", legend = legend_labels, bty = "n",
       col = 1:length(cure_rate_list), lty = 1, cex = 1, lwd = 2)
dev.off()

title3 <- switch(variable_name,
                age_group = "Estimated mixture cure survival
                   functions per age quantile",
                sex = "Estimated mixture cure survival functions
                   for men and women",
                charlson_group = "Estimated mixture cure survival
                    functions per Charlson index group",
                cvasc = "Estimated mixture cure survival
                   functions per cvasc group",
                safi_cut = "Estimated mixture cure survival
                   functions per SaFi quantile"
)

# Plot mixture cure survival
jpeg(paste0(variable_name, "_mixsurv.jpeg"))
```

```
  par(mar = c(5, 5, 4, 3))
  plot(S_list[[1]][[3]], S_list[[1]][[1]], type = "l", ylim = c
     (0.75, 1),           ylab = "Survival", xlab = "Time [Days]",
     main = title3, col =               "black", lwd = 2)
  for (i in 2:length(S_list)) {
    lines(S_list[[i]][[3]], S_list[[i]][[1]], col = i, lwd = 2)
  }
  legend_pos <- c("topright", "topright", "bottomright", "
     bottomright")
  legend(60,0.825, legend = legend_labels, bty = "n",
         col = 1:length(cure_rate_list), lty = 1, cex = 1, lwd = 2)
  axis.break(axis = 2, breakpos = 0.75, style = "zigzag")
  dev.off()
}
```

B.5 Retrieve results

```
variables <- c("age_group", "sex", "charlson_group", "cvasc", "safi
   _cut")
follow_up <- c(20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80)

# Results for each variable
results <- vector(mode = "list", length = length(variables))
datavec <- list(dataCMod, dataCMod, dataCMod, data_noCnans, data_
   noSnans)
# Iterate over variables
results_list <- list()
for (i in seq_along(variables)) {
    variable <- variables[i]
    data <- datavec[[i]]

# Compute cure rate, survival functions, and latency for the
   variable
temp_results <- compute_cure_survival(variable, follow_up, data)
time_list <- temp_results$time_list
p_list <- temp_results$p_list
S_list <- temp_results$S_list
S0_list <- temp_results$S0_list
cure_rate_list <- temp_results$cure_rate_list

# Assign the components individually to results
variable_lists <- list(
        time_list = time_list,
        p_list = p_list,
        S_list = S_list,
        S0_list = S0_list,
        cure_rate_list = cure_rate_list
)

results_list[[variable]] <- variable_lists

# Access the corresponding results for the variable
variable_results <- results_list[[variable]]
cure_rate_list <- variable_results$cure_rate_list
S_list <- variable_results$S_list
S0_list <- variable_results$S0_list

# Determine the true cure rate based on the variable
true_cure_rate <- switch(variable,
        age_group = c(true_cure_rate_1qa, true_cure_rate_2qa, true_
           cure_rate_3qa, true_cure_rate_4qa),
        sex = c(true_cure_rate_men, true_cure_rate_women),
        charlson_group = c(true_cure_rate_1, true_cure_rate_2, true
           _cure_rate_3),
        cvasc = c(true_cure_rate_nocvasc, true_cure_rate_cvasc),
        safi_cut = c(true_cure_rate_1q, true_cure_rate_2q, true_
           cure_rate_3q, true_cure_rate_4q)
)

# Call the function to generate plots for the variable
generate_plots(variable, cure_rate_list, true_cure_rate, S_list, S0
   _list)
}
```