Tutorial

Minghan Yang

2024-09-18

Table of contents

1 Survival Analysis Basics

Survival analysis is often used in areas such as biology, medicine, and engineering. For cancer studies, one typical topic is the survival of cancer patients after a diagnosis of cancer.

The data used in survival analysis is often sample of **times to event** from a population $(t_1, t_2, \dots, t_n)$. in our project, we use the breast cancer patient data from hospitals in Rotterdam.

- Survival Time: The duration of time from a starting point (for example, diagnosis of a disease) to the occurrence of an event (e.g. death).
- Vital status (censoring indicators): $\delta_1, \delta_2, \dots, \delta_n$.

  $\delta_i = 1$: death,

  $\delta_i = 0$: alive / right-censored.

In many studies, some subjects may not experience the event before the study ends, leading to censored data. Survival analysis methods can appropriately handle these censored observations. It helps estimate survival rates at different time points, which is vital for predicting and understanding the progression of diseases, patient outcomes, and the effectiveness of treatments.

1.1 Censoring

Censoring occurs when the exact time of the event is not known for some subjects. This happens for various reasons, such as the study ending before the event occurs, the subject withdrawing from the study, or loss to follow-up. For

example, if a study ends before a patient dies or if a patient leaves the study early, their data is *censored*. If death, failure of a machine or relapse of a disease happens, then the data is an *event*.

1. **Right Censoring**

Right censoring occurs when the event of interest has not happened by the end of the study period for some subjects. We only know that the event time exceeds a certain value.

Right-censoring is a common feature in survival data and must be appropriately handled to avoid bias in the analysis. Survival analysis techniques, like the Kaplan-Meier estimator and Cox proportional hazards model, are specifically designed to deal with right-censored data, allowing researchers to make valid inferences about the time-to-event process even when not all events are observed.

2. **Left Censoring**

Left censoring happens when the event of interest has already occurred before the subject is observed or enters the study. We only know that the event time is less than a certain value.

If a subject is left-censored at time t, it means the event time T is less than t.

3. **Interval Censoring**

Interval censoring occurs when the event time is known to lie within a specific interval but the exact time is unknown.

4. **Administrative Censoring**

Administrative censoring refers to the situation where all subjects are censored at a predefined time point due to the end of the study.

5. **Random Censoring**

Random censoring occurs when the censoring times are random and independent of the event times. For example, subjects dropping out of a study at random times due to various reasons such as relocation or withdrawal.

1.2 Functions

1. **Lifetime distribution function**

$$F(t) = \mathbb{P}(T < t), x \geq 0. \tag{1}$$

This function describes the probability of failing up age of t or of having a life span of at most t. $F(t)$ is a monotone and increasing function. $0 \leq F(t) \leq 1$. The derivative of $F(t)$ is the corresponding PDF $f(t)$.

2. **Survival function**

The survival function is another representative of lifetime distribution. It is defined as

$$S(t) = \mathbb{P}(T > t) = 1 - F(t) = \mathbb{P}(T > t) \tag{2}$$

The survival function is also known as the reliability function, indicating the probability of surviving a time of t. It is the probability of exceeding time t. The study of survival function is crucial in survival analysis.

Properties:

- The survival function $S(t)$ is monotone and decreasing over $[0, \infty)$. Furthermore, $S(t)$ satisfies $S(0) = 1$, $S(\infty) = 0$.
- The survival function is related to $F(t)$: $S(t) = 1 - F(t)$.
- The survival function appropriately handles right-censored data. Censored observations contribute information up until the censoring time but not beyond it. This allows the survival function to provide unbiased estimates even when not all subjects have experienced the event by the end of the study.

3. **Hazard function**

Hazard function is defined to be

$$h(t) = \lim_{dt \to 0} \frac{\mathbb{P}[t \leq T \leq t + dt \mid T \geq t]}{dt}.$$

This describes the instantaneous rate at which events occur, given no previous event. It is the probability that an event occurs in a very small time interval, given survival up to the start of the interval.

The hazard function is the ratio of the probability of the event occurring at time t to the probability of surviving up to time t.

$$h(t) = \frac{f(t)}{S(t)} \tag{3}$$

Proof:

$$\begin{aligned} h(t) &= \lim_{dt \to 0} \frac{\mathbb{P}[t \leq T \leq t + dt | T \geq t]}{dt} \\ &= \lim_{dt \to 0} \frac{\mathbb{P}[t \leq T \leq t + dt, T \geq t]}{dt \cdot \mathbb{P}(T \geq t)} \\ &= \lim_{dt \to 0} \frac{\mathbb{P}[t \leq T \leq t + dt]}{dt \cdot \mathbb{P}(T \geq t)} \\ &= \lim_{dt \to 0} \frac{\mathbb{P}[t \leq T \leq t + dt]}{dt} \cdot \frac{1}{\mathbb{P}(T \geq t)} \\ &= f(t) \cdot \frac{1}{S(t)} \quad \blacksquare \end{aligned}$$

Properties:

- $f(0) = h(0)$
- $f(t) \geq h(t), \forall t > 0$, because $S(t) \leq 1, \forall t > 0$
- $h(t)$ is *not* a density function because it is not normalised.
- Any function $h(t)$ is a hazard rate function iff:
  i) $h(t) \geq 0, \forall t \geq 0$
  ii) $\int_0^\infty h(t)\, dt = \infty$
- The survival function can be expressed in terms of the hazard function:

$$S(t) = \exp\left\{ -\int_0^t h(u)\, du \right\} \tag{4}$$

Proof:

$$h(t) = \frac{f(t)}{S(t)} = \frac{F'(t)}{1 - F(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}(\ln S(t))$$

$$\int_0^t h(u)\, du = -[\ln S(u)]_0^t = -\ln S(t) + \ln S(0) = -\ln S(t) \quad \text{, as } S(0) = 1$$

$$\therefore S(t) = \exp\left\{ -\int_0^t h(u)\, du \right\} \quad \blacksquare$$

4. **Cumulative hazard function**

The function

$$H(t) = \int_0^t h(s)\, ds \tag{5}$$

is known as the cumulative hazard function.

Properties:

- $H(0) = 0$
- $\lim_{t \to \infty} H(t) = \infty$
- $H(t)$ is non-decreasing
- $H(t) = -\ln S(t) = -\ln [1 - F(t)]$

5. **Likelihood Function**

When $\delta_i = 0$, i.e. alive/right-censored, $L_i = S(t_i \mid \theta)$.

The likelihood function is given by

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(t_i \mid \theta)^{\delta_i} S(t_i \mid \theta)^{1-\delta_i} \\ &= \prod_{i=1}^{n} [h(t_i \mid \theta) \cdot S(t_i \mid \theta)]^{\delta_i} S(t_i \mid \theta)^{1-\delta_i}, \quad \text{because } h(t) = f(t)/S(t) \\ &= \prod_{i=1}^{n} h(t_i \mid \theta)^{\delta_i} \cdot \exp\{-H(t_i \mid \theta)\}, \quad \text{because } S(t) = \exp\{-H(t)\} \end{aligned}$$

Therefore, the log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \delta_i \log h(t_i \mid \theta, Y_0) - \sum_{i=1}^{n} H(t_i \mid \theta, Y_0). \tag{6}$$

This allows for calculating the maximum likelihood estimates of the parameters θ after a hazard function is specified.

2 Hazard-Response Model

2.1 Model Formulation

ODEs are used to characterise many physical systems, and using the system of ODEs to define the hazard function adds dynamics and interpretability. Here, we use the competitive Lotka-Volterra model to describe the competition between the hazard and response (from therapy, interventions and immune system). This model assumes the competing relationship between the hazard function $f(t)$ and the response $q(t)$, which is related to the immune system and interventions recieved at a population level (Christen and Rubio 2024+). The hazard-response model is defined through the below system of ODEs.

$$\begin{cases} h'(t) = \lambda h(t)\left(1 - \frac{h(t)}{\kappa}\right) - \alpha q(t) h(t), & h(0) = h_0 \\ q'(t) = \beta q(t)\left(1 - \frac{q(t)}{\kappa}\right) - \alpha q(t) h(t), & q(0) = q_0 \\ H'(t) = h(t), & H(0) = 0 \end{cases}$$

where $\lambda > 0$, $\alpha \geq 0$, $\beta > 0$, $\kappa > 0$, $h_0 > 0$, $q_0 > 0$. The competition between hazard and response is modeled through $\alpha q(t) h(t)$, and the growth of the two are given by the logistic growth part $\lambda h(t)\left(1 - \frac{h(t)}{\kappa}\right)$.

When $\alpha = 0$, there is no competition, and the hazard function will grow and reach a carrying capacity of κ as the time tends to infinity. When $\alpha > 0$, both the hazard and the response will receive negative effect from the term $-\alpha q(t) h(t)$.

2.2 Fixing the initial conditions

To numerically solve the system of ODEs, we need to specify the initial conditions h_0 and q_0. Based on prior knowledge, within 1 month ($\Delta t = 1/12$), approximately 1 in 1000 breast cancer patient die. So we use the approximation

$$h_0 = h(0) = -\frac{S'(0)}{S(0)} \approx -\frac{S'(\Delta t)}{S(\Delta t)} \approx -\frac{S(\Delta t) - S(0)}{\Delta t S(\Delta t)} \approx 0.01.$$

For q_0, we assume that treatment usually does not start at the beginning of the follow-up. It is likely that the treatment take some time to start, so the effect of reducing the hazard function should be small at the beginning, and we choose the value $q_0 = 10^{-6}$.

2.3 Logarithmic formulation of the hazard-response model

By modeling the logarithmic hazard function, we ensure that the hazard function itself remains positive for all t, which is essential as negative hazards do not make practical sense. The logarithmic scale can improve the numerical stability of the computations, especially when solving differential equations.

$$\begin{cases} \tilde{h}'(t) = \lambda\left(1 - \frac{\exp\{\tilde{h}(t)\}}{\kappa}\right) - \alpha\exp\{\tilde{q}(t)\}, & \tilde{h}(0) = \log\{h_0\} \\ \tilde{q}'(t) = \beta\left(1 - \frac{\exp\{\tilde{q}(t)\}}{\kappa}\right) - \alpha\exp\{\tilde{h}(t)\}, & \tilde{q}(0) = \log\{q_0\} \\ H'(t) = \exp\{\tilde{h}(t)\}, & H(0) = 0 \end{cases}$$

where $\tilde{h}(t) = \log\{h(t)\}$, and $\theta = (\lambda, \kappa, \alpha, \beta)$.

2.4 Codes

The hazard-response model is defined through the below function

```
# Hazard-Response ODE model
function HazResp(dh, h, p, t)
    # Model parameters
    lambda, kappa, alpha, beta = p

    # ODE System
    dh[1] = lambda * h[1] * (1 - h[1] / kappa) - alpha * h[1] * h[2] # hazard
    dh[2] = beta * h[2] * (1 - h[2] / kappa) - alpha * h[1] * h[2] # response
    dh[3] = h[1] # cumulative hazard
    return nothing
end
```

For stability, we calculate the Jacobian manually to avoid computational instability caused by numerical approximation of the derivatives.

$$J_H R(t, h, q, H, \theta) = \begin{pmatrix} \lambda - \frac{2\lambda h}{\kappa} - \alpha q & -\alpha h & 0 \\ -\alpha q & \beta - \frac{2\beta h}{\kappa} - \alpha h & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

where $\theta = (\lambda, \kappa, \alpha, \beta)$.

```
# Jacobian for Hazard-Response model
function jacHR(J, u, p, t)
    # Parameters
    lambda, kappa, alpha, beta = p
    # state variables
    h = u[1]
    q = u[2]

    # Jacobian
    J[1, 1] = lambda * (1 - 2 * h / kappa) - alpha * q
    J[1, 2] = -alpha * h
    J[1, 3] = 0.0
    J[2, 1] = -alpha * q
    J[2, 2] = beta * (1 - 2 * q / kappa) - alpha * h
    J[2, 3] = 0.0
    J[3, 1] = 1.0
    J[3, 2] = 0.0
    J[3, 3] = 0.0
    nothing
end
```

The logarithmic hazard-response model is defined through the below function

```
# Logarithmic Hazard-Response ODE model
function HazRespL(dlh, lh, p, t)
    # Model parameters
    lambda, kappa, alpha, beta = p
```

```
    # ODE System
    dlh[1] = lambda * (1 - exp(lh[1]) / kappa) - alpha * exp(lh[2]) # log hazard
    dlh[2] = beta * (1 - exp(lh[2]) / kappa) - alpha * exp(lh[1]) # log response
    dlh[3] = exp(lh[1]) # cumulative hazard
    return nothing
end
```

Its corresponding Jacobian is calculated to be:

$$J_H R(t, h, q, H, \theta) = \begin{pmatrix} -\frac{\lambda}{\kappa}\exp\{\tilde{h}(t)\} & -\alpha\exp\{\tilde{q}(t)\} & 0 \\ -\alpha\exp\{\tilde{h}(t)\} & -\frac{\beta}{\kappa}\exp\{\tilde{q}(t)\} & 0 \\ \exp\{\tilde{h}(t)\} & 0 & 0 \end{pmatrix}$$

```
# Jacobian for Logarithmic Hazard-Response model
function jacHRL(J, u, p, t)
    # Parameters
    lambda, kappa, alpha, beta = p
    # state variables
    lh = u[1]
    lq = u[2]

    # Jacobian
    J[1, 1] = -lambda * exp(lh) / kappa
    J[1, 2] = -alpha * exp(lq)
    J[1, 3] = 0.0
    J[2, 1] = -alpha * exp(lh)
    J[2, 2] = -beta * exp(lq) / kappa
    J[2, 3] = 0.0
    J[3, 1] = exp(lh)
    J[3, 2] = 0.0
    J[3, 3] = 0.0
    nothing
end
```

Therefore we can define the ODE function for the logarithmic hazard-response model with explicit Jacobian and initial conditions:

```
using DifferentialEquations
# Hazard-Response model with explicit Jacobian
HRJL = ODEFunction(HazRespL; jac=jacHRL)

# Initial conditions (h,q,H)
lu0 = [log(1.0e-2), log(1.0e-6), 0.0]
```

3 Bayesian Methods and Turing

As mentioned in Section 1, the log-likelihood can be calculated given the hazard function:

$$\ell(\theta) = \sum_{i=1}^{n} \delta_i \log h(t_i \mid \theta, Y_0) - \sum_{i=1}^{n} H(t_i \mid \theta, Y_0). \tag{7}$$

For the hazard-response model, an analytical solution is not available. With an ODE solver from `DifferentialEquations.jl`, we will be able to solve for the hazard function, the response function, and the cumulative hazard function numerically (Rackauckas and Nie 2017).

```
#  log likelihood function (log h and log q)
log_likL = function (par)
    # Parameters for the ODE
    odeparams = exp.(par)

    sol = solve(ODEProblem(HRJL, lu0, [0.0, tmax], odeparams); alg_hints=[:stiff])
    #  sol = solve(ODEProblem(HRJL, lu0, tspan0[i, :], odeparams[i, :]),Tsit5())
    OUT = sol(df.time)

    # Terms in the log log likelihood function
    ll_haz = sum(OUT[1, status])
    ll_chaz = sum(OUT[3, :])

    ll = ll_haz - ll_chaz
return ll
end
```

In our analysis, we choose `Gamma(2,2)` to be the priors of our parameters. This is because `Gamma(2,2)` is weakly informative, meaning it is broad enough to allow the data to inform the posterior distribution without being overly restrictive. It has a mean of 4 and variance of 8, which provides a reasonable range for the parameters in the absence of strong prior knowledge. With the help of `Turing.jl` (Ge, Xu, and Ghahramani 2018), we can apply MCMC samplers for posterior distributions of the parameters.

```
using Distributions, Turing
distprior = Gamma(2,2)

@model function bayesian_model(log_likL)
    # prior (defined on the positive parameters)
   odeparams ~ filldist(distprior, 4)
        params = log.(odeparams)

        Turing.@addlogprob!(log_likL(params))
    end
```

```
bayesian_model (generic function with 2 methods)
```

4 Predictive Hazard

With the posterior distribution for the parameter θ, we can obtain the posterior distribution of the hazard function $h_\theta(t)$. Given a MCMC sample of the parameter $\theta^{(1)}, \dots, \theta^{(M)}$, the posterior hazard functions are $h(t|\theta^{(1)}), \dots, h(t|\theta^{(M)})$.

According to the relationship $h(t) = \frac{f(t)}{S(t)}$, the expected hazard should $\int \frac{f(t|\theta)}{S(t|\theta)}\pi(\theta|data)d\theta$.

On the other hand, the predictive distribution $f(t)$ for a future patient is

$$f(t|data) = \int_\Theta f(t|\theta)\pi(\theta|data)d\theta = \int_\Theta h(t|\theta)\exp(-H(t|\theta))\pi(\theta|data)d\theta.$$

Then we can calculate the posterior predictive hazard function, which is the function of interest if we hope to prognose the hazard of a future patient

$$h(t|data) = \frac{f(t|data)}{1 - \int_0^t f(r|data)dr}.$$

However, the analytical solution of $h(t|data)$ is not obtainable, so we use posterior samples of parameters to obtain

a Monte Carlo approximation of it, which is shown to be

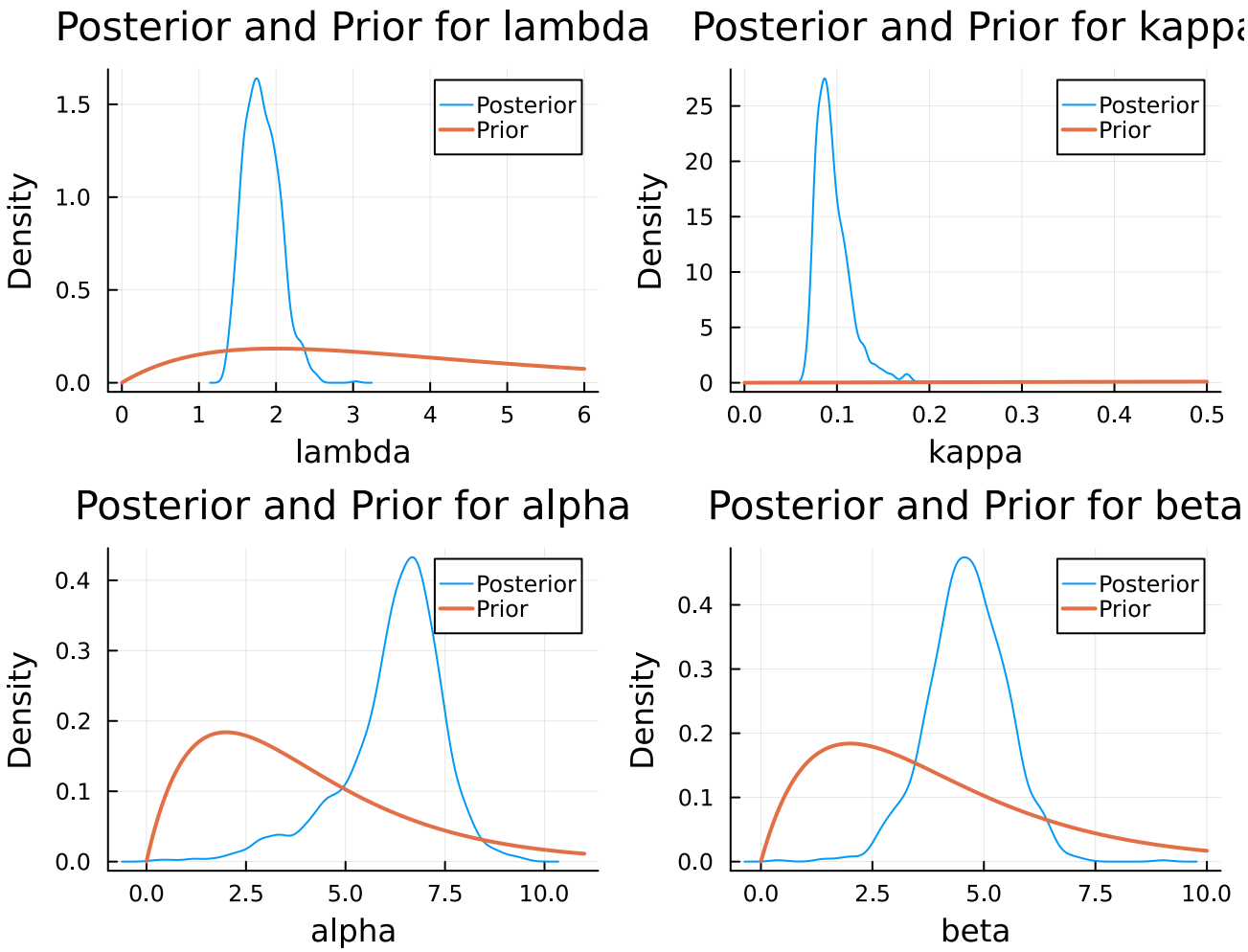
$$h(t|data) \approx \frac{\frac{1}{M}\sum_{j=1}^{M} h(t|\theta^{(j)}) \exp\{-H(t|\theta^{(j)})\}}{\frac{1}{M}\sum_{j=1}^{M} \exp\{-H(t|\theta^{(j)})\}},$$

and the predictive survival function could be approximated by

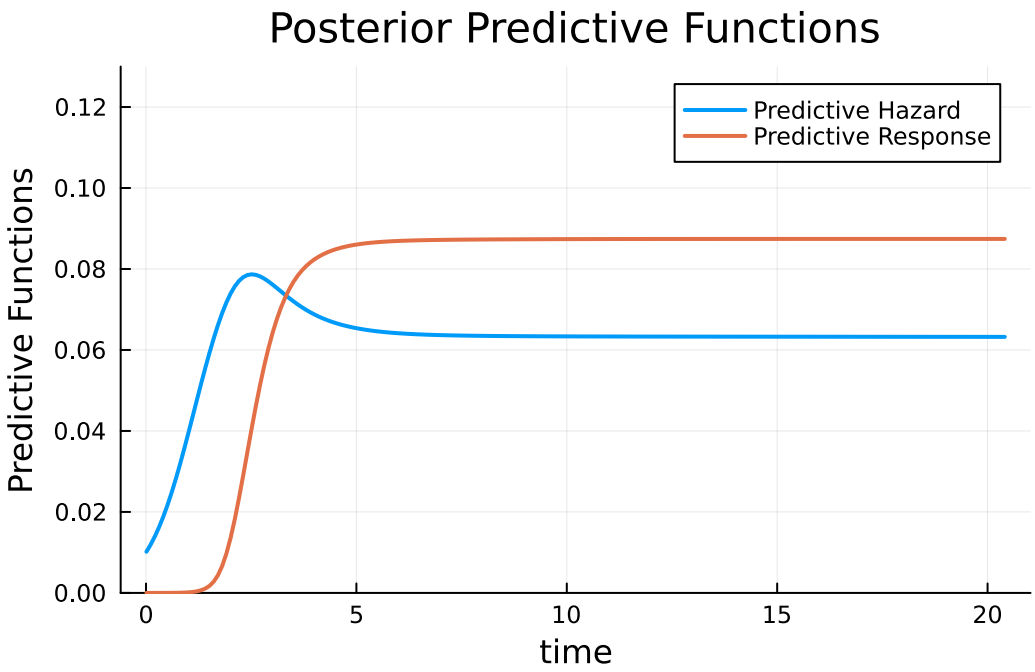
$$S(t|data) \approx \sum_{j=1}^{M} \exp\{-H(t|\theta^{(j)})\}.$$

5 Some Results

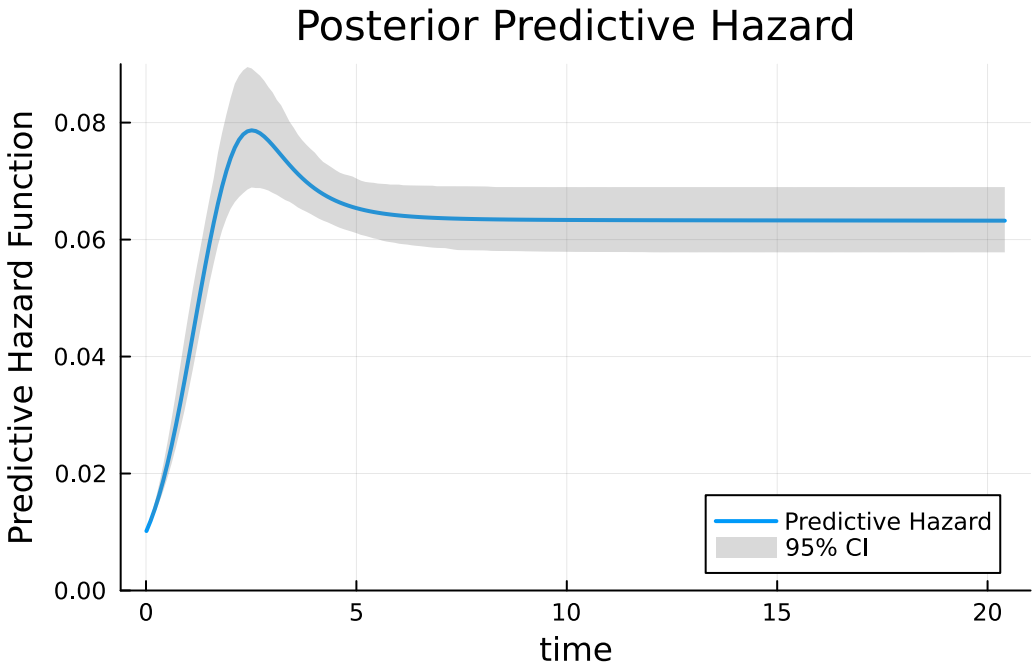
The marginal posterior distributions of the parameter $\theta = (\lambda, \kappa, \alpha, \beta)$ with priors are shown below:



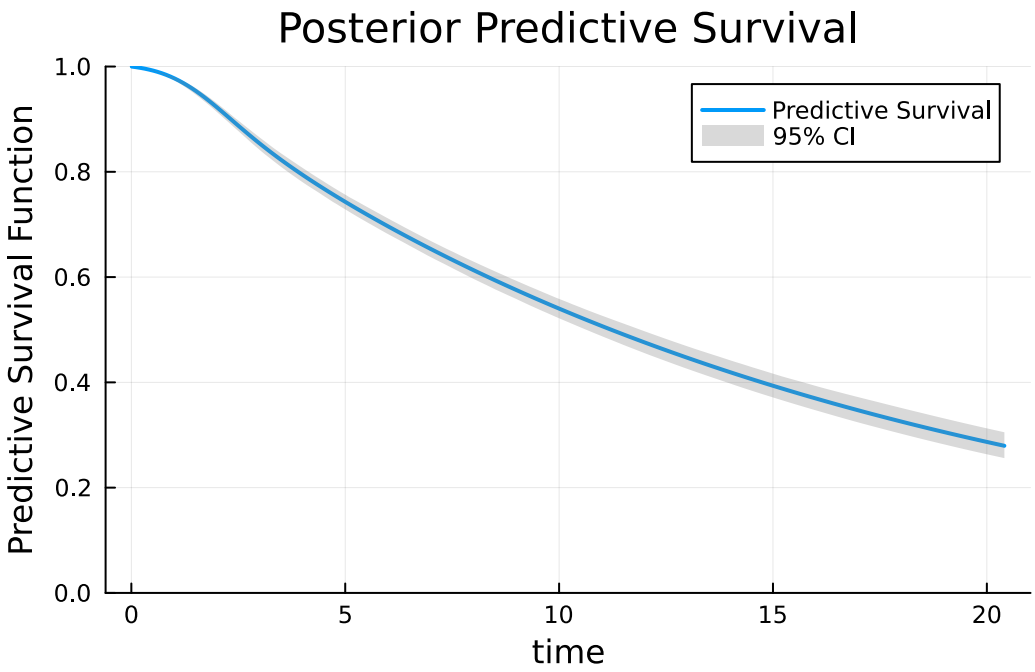
The posterior predictive hazard function and response function.



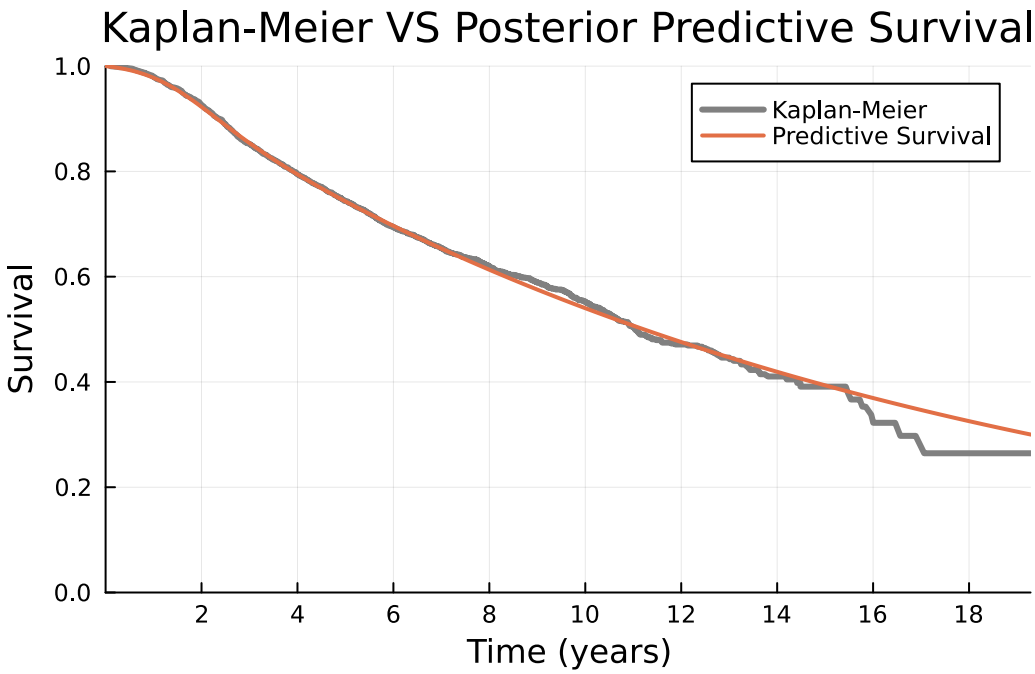
Posterior predictive hazard with 95% credible interval.



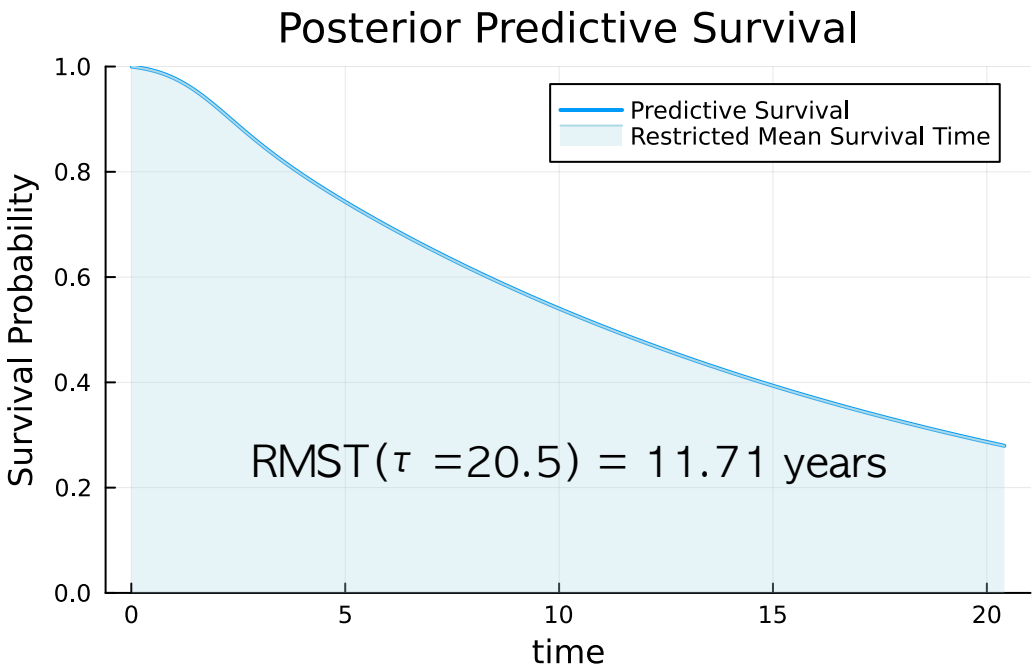
Posterior predictive survival with 95% credible interval.



We can also compare the predictive survival function with the Kaplan-Meier estimator.



We can also explore many other interpretation of the results for communication, for example the Restricted Mean Survival Time (RMST) which measures the averaged time that an individual in the population is expected to survive within a certain time frame (Belot et al. 2019).



6 Future Work

The current method apply to the analysis for population level. People might be interested in individual hazard predictions, then more covariated will be needed. It is reasonable that the hazard for individual is likely to be associated with the age, gender, deprivation level of that individual. A possible next step is exploring the incorporation of the covariates into the model. It might also be interesting to explore the application of this method to other topics outside medical statistics.

7 Summary

The hazard function is a central concept in the analysis of survival data. In this work, we try to dynamically model the hazard function with the help of autonomous ordinary differential equations (ODEs). By using ODEs, we are able to improve the interpretability and flexibility of the model. With Bayesian methods and the help of MCMC, we are able to obtain some quantitative insights into the time evolution of the hazard function. To illustrate the practical application and interpretability, we present its application to the `rotterdam` dataset of breast cancer patient data and arrive at some interesting results. This case study shows that the hazard-response model facilitates the interpretation of competing processes involving the mortality hazard and population-level responses, potentially connected to clinical treatments and natural immune mechanisms.

Belot, A., A. Ndiaye, M. A. Luque-Fernandez, D. K. Kipourou, C. Maringe, F. J. Rubio, and B. Rachet. 2019. "Summarizing and Communicating on Survival Data According to the Audience: A Tutorial on Different Measures Illustrated with Population-Based Cancer Registry Data." 53–65. https://doi.org/10.2147/CLEP.S173523.

Christen, J. A., and F. J. Rubio. 2024+. "Dynamic Survival Analysis: Modelling the Hazard Function via Ordinary Differential Equations." **Statistical Methods in Medical Research*. In Press.*, 2024+. https://doi.org/10.1177/09622802241268504.

Ge, Hong, Kai Xu, and Zoubin Ghahramani. 2018. "Turing: A Language for Flexible Probabilistic Inference," 1682–90. http://proceedings.mlr.press/v84/ge18b.html.

Rackauckas, Christopher, and Qing Nie. 2017. "DifferentialEquations.jl–a Performant and Feature-Rich Ecosystem for Solving Differential Equations in Julia." *Journal of Open Research Software* 5 (1).