A computationally efficient nonparametric sampling (NPS) method of time to event for individual-level models

David Garibay, M.P.P.[1]

Hawre Jalal, MD, Ph.D.[2]

Fernando Alarid-Escudero, Ph.D.[3 4]

2024-04-05

Purpose: Individual-level simulation models often require sampling times to events, however efficient parametric distributions for many processes may often not exist. For example, time to death from life tables cannot be accurately sampled from existing parametric distributions. We propose an efficient nonparametric method to sample times to events that does not require any parametric assumption on the hazards. **Methods:** We developed a nonparametric sampling (NPS) approach that simultaneously draws multiple time-to-event samples from a categorical distribution. This approach can be applied to univariate and multivariate processes. The probabilities for each time interval are derived from the time interval-specific constant hazards. The times to events can then be used directly in individual-level simulation models. We compared the accuracy of our approach in sampling time-to-events from common parametric distributions, including exponential, Gamma, and Gompertz. In addition, we evaluated the method's performance in sampling age to death from US life tables and sampling times to events from parametric baseline hazards with time-dependent covariates. **Results:** The NPS method estimated similar expected times to events from 1 million draws for the three parametric distributions, 100,000 draws for the homogenous cohort, 200,000 draws from the heterogeneous cohort, and 1 million draws for the parametric distributions with time-varying covariates, all in less than a second. **Conclusion:** Our method produces accurate and computationally efficient samples for time-to-events from hazards without requiring parametric assumptions. This approach can substantially reduce the computation time required to simulate individual-level models.

Introduction

Discrete-event simulation (DES) models simulate processes as discrete sequences of events that occur over time.[1] These models rely on sampling the time of different events. For example, if events have a constant rate or hazard of occurrence, the time of their occurrence can be sampled from an exponential distribution. In DES models, time-to-event data following a nonconstant hazard could be sampled from parametric distributions.[2] However, some events cannot be easily described by parametric distributions. For example, life tables, or events following hazards that are a function of time-varying covariates, such as smoking histories or tumor size, do not always follow standard parametric distributions. An alternative is to use a nonhomogeneous Poisson

[1] Health Research Consortium (CISIDAT), Cuernavaca, Morelos, Mexico.

[2] School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, ON, CA.

[3] Department of Health Policy, Stanford University School of Medicine, Stanford, CA, USA.

[4] Center for Health Policy, Freeman Spogli Institute, Stanford University, Stanford, CA, USA.

point process (NHPPP) which assumes that the rate of events follows a Poisson process that can vary over time.[3] There are different implementations of algorithms for sampling from NHPPP, which require either numerical integration or rejection sampling.[4]

In this brief report, we propose a nonparametric sampling (NSP) implementation of NHPPP that is both generalizable and computationally efficient. The method assumes that time to event is drawn from a nonparametric categorical distribution. We illustrate the NPS method using 5 examples highlighting its accuracy, flexibility and computational efficiency. Additionally, we provide an open-source implementation in `R` and `Python` to facilitate wider adoption.

Constructing the categorical distribution

The steps to implement the NPS method are described in Box 1 and shown in Figure 1. In summary, the approach involves six steps: First, obtaining the interval-specific discrete-time hazard, h_t, for the complete time interval of the analysis. If the cumulative hazard is available in continuous time, $H(t)$, h_t within a time interval Δt can be derived from $H(t+\Delta t) - H(t)$. Second, computing the cumulative discrete-time hazard, H_t. Third, deriving the discrete-time cumulative distribution function, F_t, from the cumulative hazard. Fourth, obtaining the probability mass function, p_t, from the cumulative distribution function. Fifth, sampling the times to events employing a categorical distribution, using the interval-specific probabilities, and defining each time interval as a category. And lastly, approximating time to event in continuous time. Below we provide further details on these steps.

Let $T = t$ be a random variable denoting the time-to-event following a piecewise constant hazard h_t within a time interval Δt, where $t = 0, \dots, Z$, and Z is the last time interval by which the event can occur. Thus, the cumulative hazard function at time t, H_t, is obtained from

$$H_t = \sum_{x=0}^{t} h_x \qquad (1)$$

The cumulative distribution function (CDF) of T at time t, F_t, is

$$F_t = 1 - \exp(-H_t \Delta t), \qquad (2)$$

where Δt represents representing the time interval, defined above. For example, if the hazards are on a yearly scale and we want to sample monthly time-to-event data, we use, $\Delta t = \frac{1}{12}$, and when the samples are in years, we use $\Delta t = 1$.

We derive the probability of an event happening within the t-th interval $[t, t+\Delta t)$ by the difference in the CDF in Equation 2 as

$$p_t = F_{t+\Delta t} - F_t \qquad (3)$$

To conduct a non-parametric sampling (NPS) of the time interval at which the event can occur, we define X as the time interval at which the event can occur and assume it follows a categorical distribution where each time interval is considered a category. Thus,

$$X \sim \text{Cat}[p_0, p_1, \dots, p_Z],$$

with a probability mass function $f(X = x|\mathbf{p}) = p_t$, where $\mathbf{p} = (p_0, p_1, \dots, p_Z)$, $p_t \geq 0$ is the probability of the event occurring at the t-th time interval $[t, t + \Delta t)$ and $\sum_{t=0}^{Z} p_t = 1$. Most statistical software provide built-in functions to sample from a categorical distribution. For example, in `R` is the `sample` function, and in `Python` is the `numpy.random.choice` function.

Multivariate categorical distribution

We expand the previous approach to sample values for multiple random variables simultaneously by defining a multivariate categorical distribution as

$$\mathbf{X} = [X^1, X^2, \dots, X^K] \sim \text{Cat}_K[\mathbf{p}^1, \mathbf{p}^2, \dots, \mathbf{p}^K],$$

where $X^k = x^k$ is the k-th random variable with a vector of probabilities p^k of having the events in each of the time intervals defined as

$$\mathbf{p}^k = \left[p_0^k, p_1^k, \dots, p_Z^k\right]$$

Common statistical software has no built-in functions to sample from a multivariate categorical distribution. However, we provide the code of the multivariate categorical distribution in `R` and `Python` in the Supplementary material.

Approximating continuous time-to-event

An approximation error occurs when approximating continuous time-to-event by using a discrete-time approach.[7] Since the NPS samples for the exact time categories that were initially defined while dividing the time interval, the method does not contemplate the possibility of events happening in between any two categories. This generates a systematic bias which could be reduced by adding a random variable $Y \sim U[0, \Delta t]$ to X, assuming that the time to event within each Δt interval is equally likely to occur within the interval. Thus, the random variable of the time to event with the correction, is $X_c = X + Y$. The steps to sample time to event using an NPS are shown in Box 1 and illustrated in Figure 1.

Accounting for covariates

Hazards could be a function of either of time-independent covariates, such as sex, race, or birth cohort, or time-dependent covariates, such as smoking histories, exposure to environmental risk factors or tumor size. In this section, we demonstrate the use the NPS method to sample times to events from hazards as functions of time-independent and time-dependent covariates.

Time-independent covariates

Let the i-th individual time to an event T_i follow a time-dependent hazard, $h_i(t) = f_i(x_i; \beta)$, over a time interval $[0, Z]$, as a function of a time-independent covariate, x_i, that can take any functional form and vary between individuals, and a set of coefficients β. We assume a proportional hazards approach of the effect of the covariates on the hazard to demonstrate how to sample time to events the NPS method. That is, $h_i(t) = f(x_i; \beta) = h_0(t)e^{x_i\beta}$, where $h_0(t)$ is

the time-dependent baseline hazard, x_i is the covariate for the i-th individual and β is the log-hazard ratio of the proportional effect of the covariate x_i on $h_0(t)$.

Time-dependent covariates

We now consider that the covariate can vary over time $x_i(t)$ and can take any functional form, which results in a time-varying hazard $h_i(t) = f_i(x_i(t); \beta)$, over a time interval $[0, Z]$. The time-dependent covariate could be the same across all individuals (e.g., all experiencing the same mean tumor growth over time) or vary by individuals (e.g., everyone having their own smoking history). To use the NPS method to sample from hazards with time-varying covariates, we generate or pre-specify the time-dependent covariate and compute the corresponding hazard. For example, Figure 2 shows a time-dependent Weibull hazard. We use the multivariate categorical distribution to sample time to events for multiple individuals with different covariate paths.

Examples

Below, we provide 5 examples to illustrate the implementation of the NPS method for different processes. The R code for these examples and the function of the multivariate categorical distribution is provided in a GitHub repository (https://github.com/DARTH-git/NPS_time_to_event).

Example 1: Time to event from parametric hazards

We used the NPS method for drawing times to events from various commonly used parametric distributions, such as exponential, gamma, and log-normal. We derived the piece-wise constant hazard, h_t, as described in step 1 in Figure 1 and Box 1 and applied Equation 2 and Equation 3. We sampled 10,000 times to event, computed the mean across all samples, and repeated this 1,000 times to compute the overall mean across all simulations. We then compared the expected time to event obtained from our method, with and without the approximation to continuous-time interval, to the analytic expected time from the parametric distributions. We also computed the mean execution time, and their 95% interquantile range (IQR, see Table 1), from 100 iterations using a computer with 2.3GHz Quad-Core Intel Core i7 with 32GB memory.

Example 2: Sampling age to death from a homogeneous cohort

We sampled the age to death for 100,000 individuals in a hypothetical cohort from the US population in 2015.[8] We estimated the life expectancy by taking the average across the 100,000 samples with the continuous-time approximation. The probability mass function (PMF) for the age to death obtained from the NPS methods closely follows the PMF from the life table (Figure 3). The estimated life expectancy from the NPS method is 78.53 years, which is close to the life expectancy obtained from the life tables of 78.37 years. We also calculated the mean execution time, repeating the sampling process 100 times (Table 1).

Example 3: Drawing age to death from a heterogeneous cohort

We used the multivariate categorical distribution to sample ages to death for 100,000 males and females from sex-specific life tables for the U.S. population in 2015, with the continuous-time approximation defined above. The sex-specific PMF from the NPS method and the exact PMF

from life tables are shown in Figure 4. The NPS method's estimated life expectancy id 76.22 and 80.93 years for males and females, respectively. The life expectancy obtained from the life tables was 75.93 and 80.76 years for males and females, respectively. The mean execution time, repeating the sampling process 100 times is 349.03 milliseconds (Table 1).

Example 4: Drawing time to event from hazards with time-dependent covariates

We used a proportional hazard setup with a time-dependent covariate that increases linearly over time, $x_i(t) = \alpha_0 + \alpha_1 t$, obtaining $h_i(t) = h_0(t)e^{(x_i(t)\beta)} = h_0(t)e^{((\alpha_0+\alpha_1 t)\beta)}$. We compared the accuracy of the method in sampling time-to-events from parametric exponential (rate = 0.1) and Gompertz (shape = 0.1, scale = 0.001) baseline hazards, considering a linear time-varying covariate ($\alpha_0 = 0$, and $\alpha_1 = 1$) with a log-hazard ratio ($\beta = 1.02$) against those obtained using direct sampling (DS) from the inverse cumulative density functions obtained analytically.[9,10]

The NPS method produced similar expected time to events for the two distributions compared to the DS method, from 1 million draws: exponential (8.61 NPS vs. 8.52 DS), Gompertz (35.98 NPS vs. 35.48 DS), and Weibull (8.79 NPS vs. 8.02, DS). Their mean execution time in milliseconds, repeating the sampling process 100 times, is 35.65, 55.23, and 57.34, respectively.

Example 5: Drawing time to event from hazards with time-dependent covariates following random paths

We specify a time-varying covariate $x_i(t) = \alpha_0 + \alpha_1 y_i(t)$ assuming $y_i(t)$ follows a Gaussian random walking process $y_i(t) = y_i(t-1) + \epsilon_i$, where $\epsilon_i \sim Normal(\mu = 0, \sigma = 0.5)$ and generated 1,000 random paths over 100 years (Figure 5). We assume a Weibull baseline hazard, $h_0(t) = Weibull(shape = 1.3, scale = 30.1)$, obtaining $h_i(t) = h_0(t)e^{(x_i(t)\beta)} = h_0(t)e^{\left((\alpha_0+\alpha_1 y_i(t))\beta\right)}$. We used the multivariate categorical distribution to sample times to events from the individual-level Gaussian random walk processes and estimated an expected time to event of 27.88 years. Repeating the sampling process 100 times, the average sampling time was 3.91 milliseconds.

Discussion

We developed a nonparametric approach method to sampling times to events with high computation efficiency. The NPS method uses a categorical distribution, which discretizes the hazard of events over a fixed and finite time period, assuming a piecewise hazard. We illustrated the NPS method with five examples that show common situations encountered when building DES models and provided their mean execution times. NPS can be used to sampling age to death from age-, sex-, race-, and year-specific life tables, and smoking histories. In this case, only a nonparametric estimator of the survival curve or cumulative hazard is provided. Another example is considering hazards that include time-independent or time-dependent covariates.

The NPS method accurately approximates the expected time to events from parametric distributions and can generate times to events from hazards for which no parametric distributions can be accurately fitted. Once the probability distributions are derived from the observed hazards, the sampling process is computationally efficient and can be easily repeated multiple times.

Our approach does not provide criteria to determine the optimal time interval length and it is up to the user to define it. This may pose a limitation, because selecting an excessively wide interval can result in distributions that do not resemble the observed hazard, such as those with extremely swift changes in their levels. However, this is a focus for future research. Besides, since this approach uses a nonparametric categorical distribution, it does not provide summarized descriptive information about the original data.

We proposed a method that can efficiently generate time to events from any process as long as we have the hazard over time. Additionally, this method can sample from multiple different hazards simultaneously with the multivariate categorical distribution, which we provide as an `R` and `Python` in the Supplementary material.

Financial disclosure:

Dr. Alarid-Escudero was supported by the grant U01CA253913, and Drs. Alarid-Escudero and Jalal were supported by the grant U01CA265750 from the National Cancer Institute (NCI) as part of the Cancer Intervention and Surveillance Modeling Network (CISNET). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funding agencies had no role in the design of the study, interpretation of results, or writing of the manuscript. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

Aknowledgements:

We thank Rowan Iskandar for his valuable contributions to the code for the multivariate categorical sampling. We thank Karen Kuntz, Thomas Trikalinos and Yuliia Sereda for their feedback and ideas related to this work.

References

1. Karnon J, Stahl J, Brennan A, Caro JJ, Mar J, Möller J. Modeling using discrete event simulation: A report of the ISPOR-SMDM modeling good research practices task force–4. Medical Decision Making [Internet]. 2012;32(5):701–11. Available from: https://doi.org/10.1177/0272989X12455462

2. Arrospide A, Ibarrondo O, Blasco-Aguado R, Larrañaga I, Alarid-Escudero F, Mar J. Using age-specific rates for parametric survival function estimation in simulation models. Medical Decision Making [Internet]. 2024;0(0):0272989X241232967. Available from: https://doi.org/10.1177/0272989X241232967

3. Cinlar E. Introduction to Stochastic Processes. Englewood Cliffs, NJ: Prentice-Hall; 1975.

4. Trikalinos TA, Sereda Y. Nhppp: Simulating nonhomogeneous poisson point processes in r [Internet]. 2024. Available from: https://arxiv.org/abs/2402.00358

5. Elbasha EH, Chhatwal J. Theoretical foundations and practical applications of within-cycle correction methods. Medical Decision Making. 2016;36(1):115–31.

6. Elbasha EH, Chhatwal J. Myths and misconceptions of within-cycle correction: a guide for modelers and decision makers. PharmacoEconomics. 2016;34(1):13–22.

7. Hunink MGGM, Weinstein MC, Wittenberg E, Drummond MF, Pliskin JS, Wong JB, et al. Decision Making in Health and Medicine [Internet]. 2nd ed. Cambridge: Cambridge University Press; 2014. Available from: http://ebooks.cambridge.org/ref/id/CBO9781139506779

8. Max Planck Institute for Demographic Research (Germany), University of California-Berkeley (USA) and French Institute for Demographic Studies (France). HMD. Human mortality database. https://www.mortality.org/Country/Country?cntr=USA; 2021.

9. Austin PC. Generating survival times to simulate cox proportional hazards models with time-varying covariates. Statistics in Medicine [Internet]. 2012;31(29):3946–58. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5452

10. Julius S. Ngwa DMC Howard J. Cabral, Cupples LA. Generating survival times with time-varying covariates using the lambert w function. Communications in Statistics - Simulation and Computation [Internet]. 2022;51(1):135–53. Available from: https://doi.org/10.1080/03610918.2019.1648822

Tables

Table 1: Comparison of expected time to events and mean sampling time, in milliseconds, from 100 iterations of N samples each between the non-parametric sampling (NPS) method and parametric distributions or life table estimates. NPS-U: Non-parametric sampling uncorrected; NPS-C: Non-parametric sampling corrected by adding a uniformly distributed random number.

Distribution	NPS-U	NPS-C	Analytic solution
Exponential (rate = 0.1; N = 10,000)			
Expected value	9.51	10.00	10.00
Mean execution time (95% IQR)	0.35[0.31, 0.47]	0.53[0.43, 0.72]	
Gamma (rate = 0.1, shape = 4; N = 10,000)			
Expected value	39.47	39.98	40.00
Mean execution time (95% IQR)	0.88[0.39, 4.84]	0.80[0.55, 1.10]	
Log-normal (μ = 3.5, σ = 0.15; N = 10,000)			
Expected value	32.99	33.49	33.49
Mean execution time (95% IQR)	0.34[0.30, 0.49]	0.83[0.47, 1.76]	
Life tables - Homogeneous cohort; N = 100,000			
Expected value	78.04	78.53	78.37
Mean sampling time (95% IQR)	5.07[4.63, 5.75]	7.51[6.54, 9.51]	
Life tables - Heterogeneous cohort; N = 200,000			
Expected execution (females)	80.32	80.93	80.76
Expected execution (males)	75.71	76.22	75.93
Mean sampling time (95% IQR)	377.25[227.60, 662.33]	349.03[241.87, 555.16]	

Boxes

1. Get the cumulative continuous-time hazard $H(t) = \int_0^t h\,(x)dx$.
2. Compute the discrete-time hazard h_t for a time interval Δt as $H(t + \Delta t) - H(t)$ and derive the cumulative discrete-time hazard, $H_t = \sum_{x=0}^{t} h_x$.
3. Calculate the cumulative distribution function $F_t = 1 - \exp(H_t \Delta t)$.
4. Calculate the category-specific sample probabilities $p_t = F_{t+\Delta t} - F_t$
5. Sample the time to event, $X = x$, from a categorical distribution.
   a. If one or multiple samples are taken from the same process, $X \sim \text{Cat}[p_0, p_1, \ldots, p_Z]$, use the univariate categorical distribution through the `sample` function in base `R` or `numpy.random.choice` in `Python`.
   b. If multiple samples are taken from different processes, $\mathbf{X} = [X^1, X^2, \ldots, X^K] \sim \text{Cat}_K[\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^K]$, use the multivariate categorical distribution through the nps_nhppp function, implemented for `R` and `Python`, provided in the Supplementary material.
6. Approximate to continuous time by adding a random variable $Y \sim U[0, \Delta t]$ to all sampled elements.

Box 1: Steps to draw a time-to-event using a nonparametric sampling (NPS) approach.

Figures

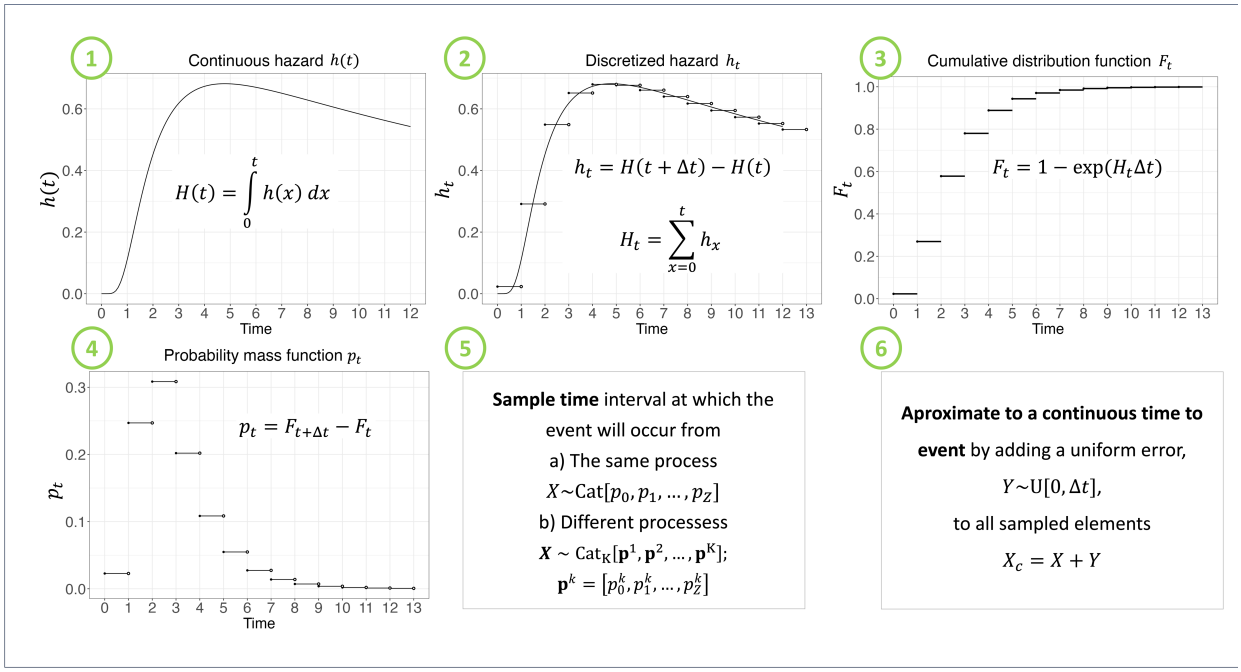

Figure 1: Steps to sample time-to-events using a nonparametric sampling (NPS) approach.

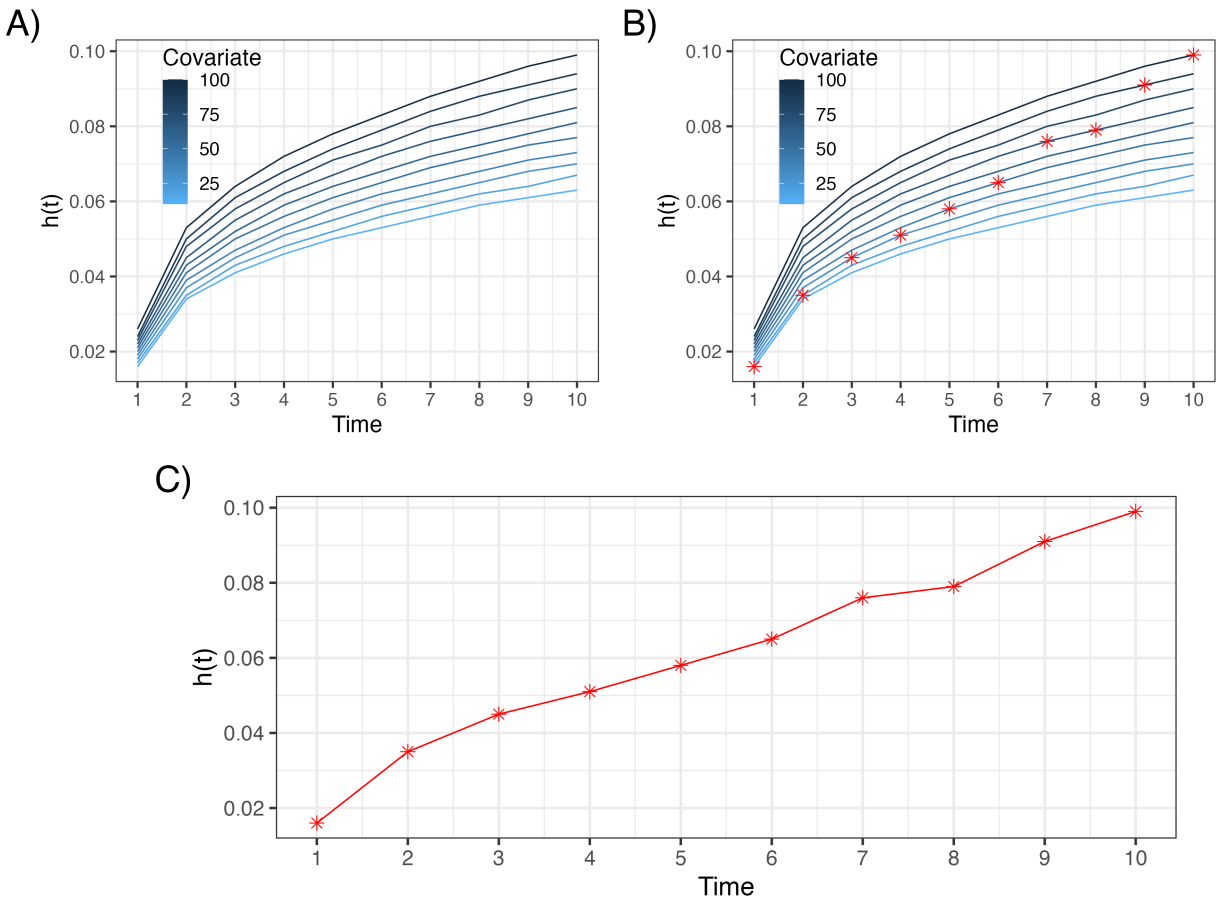

Figure 2: A) Time-dependent hazard, $h(t)$, for different values of a covariate; B) example of a covariate path; C) Corresponding path of the $h(t)$.

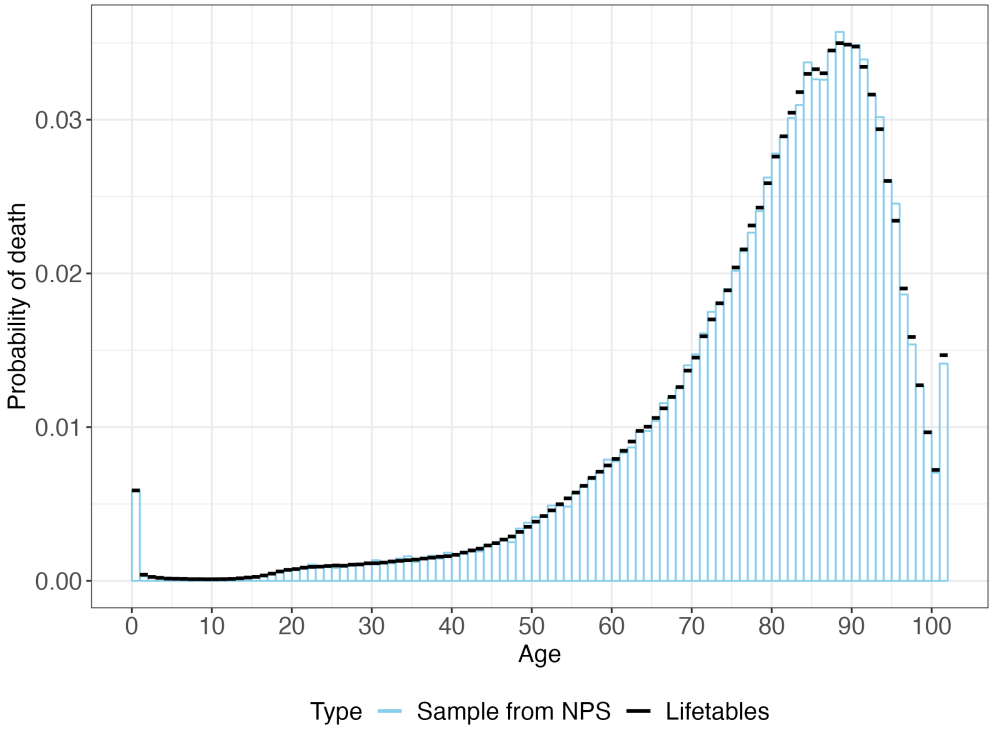

Figure 3: Probability mass function (PMF) of dying within a year of age in the total U.S. population in 2015.

Figure 4: PMF of dying within a year of age by sex, U.S. population in 2015.

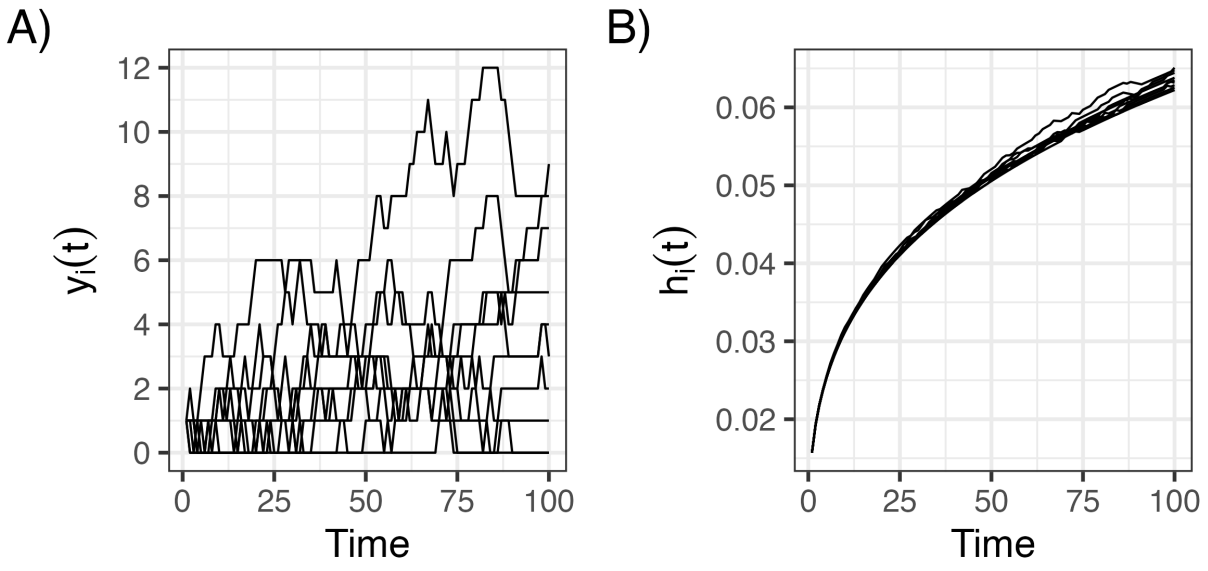

Figure 5: A) Individual-specific trajectories. B) Individual-specific time-dependent hazards. Sample of 10 individuals