Methodological developments to describe the association between socioeconomic inequalities and cancer survival with an illustration using French population-based data

Aurélien Belot

Cancer survival group, Non-Communicable Disease Epidemiology,
Faculty of Epidemiology and Population Health,
London School of Hygiene and Tropical Medicine

aurelien.belot@lshtm.ac.uk

Collaborative work with:
- Hadrien Charvat (National Cancer Center, Tokyo, Japan),
- Laurent Remontet, Nadine Bossard, Laurent Roche (Service de Biostatistique, Lyon, France),
- Olivier Dejardin, Guy Launoy (U1086 INSERM, Cancers and Preventions, Caen, France),
- Bernard Rachet (Cancer Survival Group, LSHTM)

Fundings: ANR (MESURE/CENSUR Working Survival Group ANR-09-BLAN-0357/ANR-12-BSV1-0028) and IRESP (AAR2013-13)

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Outline

Introduction
- Context
- Objectives

Mixed-effect hazard-based regression models
- Excess hazard regression model
- Shared frailty model
- Mixed-effect excess hazard regression model
- Likelihood function and estimation procedure
- Simulation study

R Package mexhaz

Illustration

Context 1/2

Describe the association between the socio-economic status and the cancer-specific hazard using population-based cancer registry data

Context 1/2

Describe the association between the socio-economic status and the cancer-specific hazard using population-based cancer registry data

- No cause of death information

Context 1/2

Describe the association between the socio-economic status and the cancer-specific hazard using population-based cancer registry data

- No cause of death information
- Socio-economic level of patients assessed by an ecological measure (area of residence)

Context 1/2

Describe the association between the socio-economic status and the cancer-specific hazard using population-based cancer registry data

- No cause of death information
- Socio-economic level of patients assessed by an ecological measure (area of residence)
- Hierarchical structure of the data
  - Level 1: individual's time-to-event
  - Level 2: cluster (area of residence, hospital, ...)

Context 1/2

Describe the association between the socio-economic status and the cancer-specific hazard using population-based cancer registry data

- No cause of death information
- Socio-economic level of patients assessed by an ecological measure (area of residence)
- Hierarchical structure of the data
    - Level 1: individual's time-to-event
    - Level 2: cluster (area of residence, hospital, ...)

⇒ Assumption of independence between individual's survival times is violated for individuals living in the same area (same level of deprivation, but also local medical practice, environmental factors...)

Context 1/2

Describe the association between the socio-economic status and the cancer-specific hazard using population-based cancer registry data

- No cause of death information
- Socio-economic level of patients assessed by an ecological measure (area of residence)
- Hierarchical structure of the data
  - Level 1: individual's time-to-event
  - Level 2: cluster (area of residence, hospital, ...)

⇒ Assumption of independence between individual's survival times is violated for individuals living in the same area (same level of deprivation, but also local medical practice, environmental factors...)

⇒ Correct statistical inference requires that the hierarchical structure of the data be taken into account.

Context 2/2

- Cancer-specific hazard without the cause of death?
  ⇒ excess hazard regression models
- Correlated data / hierarchical structure?
  ⇒ mixed effect models (multilevel models) provide a satisfying and convenient theoretical framework by introducing a random effect at the cluster level.

Mixed effect models have been developed in the context of overall survival
But lack of tools/development in the context of net survival/excess hazard regression models

Objectives

Methodological

- To propose an approach to fit an excess hazard regression model with a random effect, allowing for non linear and time-dependent effects of covariates

Objectives

Methodological

- To propose an approach to fit an excess hazard regression model with a random effect, allowing for non linear and time-dependent effects of covariates
- To evaluate the performances of the proposed approach in an extensive simulation study

Objectives

Methodological

- To propose an approach to fit an excess hazard regression model with a random effect, allowing for non linear and time-dependent effects of covariates
- To evaluate the performances of the proposed approach in an extensive simulation study
- To make available the approach in an user-friendly software (R-package)

Objectives

Methodological

- To propose an approach to fit an excess hazard regression model with a random effect, allowing for non linear and time-dependent effects of covariates
- To evaluate the performances of the proposed approach in an extensive simulation study
- To make available the approach in an user-friendly software (R-package)

Epidemiological

- To describe the association between socioeconomic context and cancer survival using French population-based cancer registry data
- To provide some methodological guidelines

Excess hazard regression model 1/2

Classical method used to analyse population-based cancer registry data

The overall mortality hazard λ is split into an excess mortality hazard (due to cancer) λ_E and an expected (or population) mortality hazard λ_P [Estève 1990]:

$$\boxed{\lambda(t, \mathbf{x}, \mathbf{z}) = \lambda_E(t, \mathbf{x}) + \lambda_P(a + t, y + t, \mathbf{z})}$$

Where

- Covariates $\mathbf{x}$: age at diagnosis a, deprivation, stage at diagnosis, sex, year of diagnosis y, ...
- Variables defining the population mortality hazard in the life-table: age $a + t$, year $y + t$ and $\mathbf{z}$ (sex, region, deprivation, ...)

Excess hazard regression model 2/2

$$\boxed{\lambda(t,\mathbf{x},\mathbf{z}) = \lambda_E(t,\mathbf{x}) + \lambda_P(a+t, y+t, \mathbf{z})}$$

- The quantity λ_P is considered known
- The quantity to estimate is λ_E

Many different models have been proposed: more flexible and allowing time-dependent effects using splines [Bolard 2002, Giorgi 2003, Lambert 2005, Nelson 2007, Remontet 2007, Pohar-Perme 2009, ...]

But nothing has been done to fit an for excess hazard model on correlated data, without losing flexibility (parametric hazard, or piecewise step function [Dupont 2013])

The classical shared frailty hazard-based regression model

In survival analysis, random effect is usually called "frailty"

The frailty, u, can be viewed as a random variable that acts multiplicatively on the baseline hazard [Duchateau 2008, Wienke 2011].

$$\lambda(t; \mathbf{x}_{ij}, u_i) = \lambda_0(t) u_i \exp({}^t\mathbf{x}_{ij}\beta)$$

Each geographical unit i has a frailty value u_i $[= exp(w_i)]$ which is shared by all individuals j observed in unit i

The classical shared frailty hazard-based regression model

In survival analysis, random effect is usually called "frailty"

The frailty, u, can be viewed as a random variable that acts multiplicatively on the baseline hazard [Duchateau 2008, Wienke 2011].

$$\lambda(t; \mathbf{x}_{ij}, u_i) = \lambda_0(t) u_i \exp({}^t\mathbf{x}_{ij}\boldsymbol{\beta})$$

Each geographical unit i has a frailty value u_i [$= exp(w_i)$] which is shared by all individuals j observed in unit i

Usual assumptions:

- Parametric distribution for T (Weibull, piecewise constant,...)
- Gamma distribution for the frailty u

Mainly due to practical reasons (analytical expression of the marginal likelihood)

$\Rightarrow$ No tool for flexible (excess) hazard

Mixed-effect Excess hazard regression model

The flexible model proposed

$$\boxed{\lambda_E(t, \mathbf{x}_{ij}) = \lambda_0(t; \boldsymbol{\xi}) \cdot \exp(\beta_1 x_1 + f(x_2; \boldsymbol{\beta}_2) + g(t; \boldsymbol{\beta}_3) x_3 + w_i)}$$

Where

- λ_0 is the baseline hazard modelled with (exp of) B-splines (or piecewise step function or Weibull),
- β_1 the linear and proportional (fixed) effect of x_1,
- f and g are flexible functions (B-splines) allowing for non-linear and non-proportional effects for x_2 and x_3 (defined with $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$), respectively,
- w_i is the random effect of cluster i, assumed to follow a normal distribution with mean 0 and standard deviation σ

Likelihood function: overview

1. Likelihood of one observation j in cluster i
2. Conditional Likelihood for cluster i
3. Marginal Log-Likelihood for cluster
4. Total Log-likelihood

Likelihood function: overview

1. Likelihood of one observation *j* in cluster *i*

2. Conditional Likelihood for cluster *i*

3. Marginal Log-Likelihood for cluster

4. Total Log-likelihood

Conditional Likelihood for cluster *i*

For one observation *j* in cluster *i*: $\{t_{ij}, \delta_{ij}, \mathbf{x}_{ij}\}$

$$\mathrm{L}_{ij}^{C}(\boldsymbol{\beta}|w_i) =$$
$$\exp\left\{-\Lambda_E(t_{ij}, \mathbf{x}_{ij}, w_i) - \Lambda_P(a_{ij} + t_{ij}, \mathbf{z}_{ij})\right\}\left\{\lambda_E(t_{ij}, \mathbf{x}_{ij}, w_i) + \lambda_P(t_{ij}, \mathbf{z}_{ij})\right\}^{\delta_{ij}}$$

- Gauss-Legendre quadrature to approximate the cumulative excess hazard $\Lambda_E(t_{ij}, \mathbf{x}_{ij}, w_i) = \int_0^t \lambda(u, \mathbf{x}_{ij}, w_i)\, \mathrm{d}u$
- Last term of the exponential can be omitted (does not depend on the $\boldsymbol{\beta}$s)

For cluster *i*:

$$\mathrm{L}_{i}^{C}(\boldsymbol{\beta}|w_i) = \prod_{j=1}^{n_i}\left\{\mathrm{L}_{ij}^{C}(\boldsymbol{\beta}|w_i)\right\}$$

Likelihood function: overview

1. Likelihood of one observation *j* in cluster *i*
2. Conditional Likelihood for cluster *i*
3. Marginal Log-Likelihood for cluster
4. Total Log-likelihood

Marginal Likelihood for cluster i

We assume a normal distribution for the random effect, with mean=0 and variance=σ^2, $\phi(w,0,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{w^2}{2\sigma^2}\right\}$

For cluster i

$$\mathrm{L}_i^M(\boldsymbol{\beta},\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathrm{L}_i^C(\boldsymbol{\beta}|w) \exp\left\{-\frac{w^2}{2\sigma^2}\right\} \mathrm{d}w$$

- Problem : How to evaluate this likelihood ?
- A solution is to use the Gauss-Hermite Quadrature (GHQ)

Definition

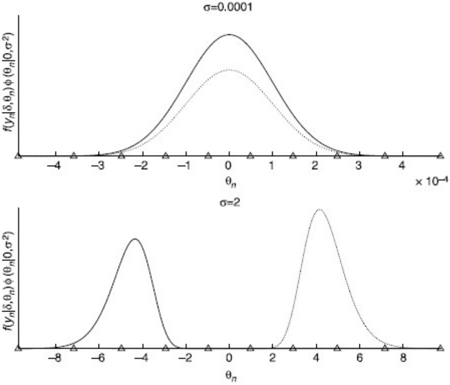
Gauss-Hermite Quadrature

$$\int_{-\infty}^{\infty} f(v)\exp\{-v^2\}\,\mathrm{d}v \approx \sum_{k=1}^{Q} \rho_k^H \cdot f(x_k^H)$$

- Nodes $= x_k^H$
- Weights $= \rho_k^H$

The nodes and weights depend only on Q (not on the integrand f...)

Illustration of the GHQ



Tuerlinckx F et al., British Journal of Mathematical and Statistical Psychology, 2006

A refinement of the GHQ : the **adaptive** GHQ

Basic idea:

- The quadrature locations are rescaled and translated so that they cover the region where the integrand varies most, i.e. around its mode
- To transform the integrand to obtain a new quadrature formula in which the new nodes and the corresponding weights depend on the integrand (and so on the cluster i)

The **adaptive** GHQ 1/2

Apply the LAPLACE approximation to :

$$g_i(w,\boldsymbol{\beta},\sigma) = \mathrm{L}_i^C(\boldsymbol{\beta}|w)\phi(w,0,\sigma) \quad \Rightarrow \quad \left\{ \begin{array}{l} \mu_i \\ \sigma_i \end{array} \right.$$

We have :

$$\mathrm{L}_i^M(\boldsymbol{\xi},\boldsymbol{\beta},\sigma) = \int_{-\infty}^{\infty} \underbrace{\frac{g_i(w,\boldsymbol{\beta},\sigma)}{\phi(w,\mu_i,\sigma_i)}}_{f_i^{\mathrm{A}}(w,\boldsymbol{\beta},\sigma)} \phi(w,\mu_i,\sigma_i)\,\mathrm{d}w$$

Using the GHQ, we approximate :

$$\mathrm{L}_i^M(\boldsymbol{\xi},\boldsymbol{\beta},\sigma) \approx \sum_{k=1}^{Q} \rho_k^N(\mu_i,\sigma_i) \cdot f_i^{\mathrm{A}}\left(x_k^N(\mu_i,\sigma_i),\boldsymbol{\beta},\sigma\right)$$

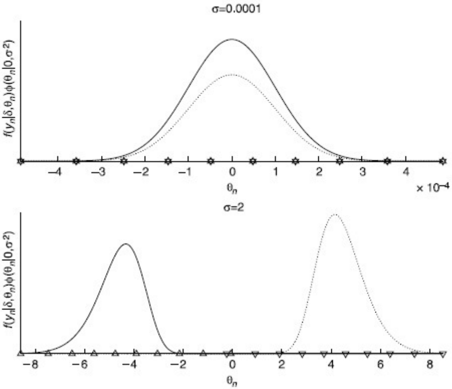
The **adaptive** GHQ 2/2

The modified nodes and weights are given (as functions of the original ones) by:

$$\begin{cases} x_k^N(\mu_i, \sigma_i) = \mu_i + \sigma_i\sqrt{2} \cdot x_k^H \\ \rho_k^N(\mu_i, \sigma_i) = \rho_k^H \cdot \sigma_i\sqrt{2\pi}\exp\{(x_k^H)^2\} \end{cases}$$

More details in Liu & Pierce [14] and Pinheiro & Bates [15]

Illustration of the Adaptive GHQ



More accurate approximation than GHQ and it needs less quadrature points

Tuerlinckx F et al., British Journal of Mathematical and Statistical Psychology, 2006

Likelihood function: overview

1. Likelihood of one observation *j* in cluster *i*
2. Conditional Likelihood for cluster *i*
3. Marginal Log-Likelihood for cluster
4. Total Log-likelihood

Finally

Log-likelihood for cluster *i*

$$\ell_i^M(\boldsymbol{\beta},\sigma) \approx \log\left\{\sum_{k=1}^{Q} \rho_k^N(\mu_i,\sigma_i) \cdot f_i^{\mathrm{A}}\left(x_k^N(\mu_i,\sigma_i),\boldsymbol{\beta}\right)\right\}$$

Total Log-likelihood

$$\ell(\boldsymbol{\beta},\sigma) \approx \sum_{i=1}^{D} \ell_i^M(\boldsymbol{\beta},\sigma)$$

To estimate the parameters $(\widehat{\boldsymbol{\beta}},\widehat{\sigma})$, use a standard optimisation routine (such as the Newton-Raphson algorithm) on the quantity $\ell(\boldsymbol{\beta},\sigma)$
More details in Charvat *et al.*, StatMed 2016 [1]

Overview of the different simulated scenarios 1/2

Aim: to evaluate the performances of the proposed approach in different scenarios, in terms of its ability to estimate

- the baseline excess hazard
- the fixed effects of covariates defined **both** at the individual level and at the cluster level (including time-dependent effect)
- the variance of the random effect

Overview of the different simulated scenarios 2/2

In scenarios A and B, the **impact of the design** (number of clusters and number of patients by cluster) and the **level of the variance** of the random effect were studied

- scenario A: Balance-Design: N patients by cluster is fixed
- scenario B: UnBalance-Design: N patients by cluster is variable

In scenario C, we studied the ability of our approach to model **non proportional effect** ((NPH)) of covariates (with unbalanced design)

In scenario D, we checked the robustness of our approach in case of **miss-specified distribution of the random effect** (with unbalanced design)

Simulation study (I)

Design of the 1000 simulated dataset, with 1000 patients in each

- Age (25% [30, 65], 35% [65, 75], 40% [75, 85], with an uniform law in each age-class)
- Sex (Binomial distribution with P(sex=man)=0.5
- Cluster (the cluster ID (D = 10, 20, 50, 100))
- Deprivation Index (DI) defined at the cluster level (Normal(0,sd=1.5))

In scenarios A, Balance-Design: the number of patients by cluster is **exactly** equal to 10, 20, 50 or 100

In scenarios B, UnBalance-Design: the number of patients by cluster is **variable and equal, on average**, to 10, 20, 50 or 100 (one additional simulated condition with 800 clusters and 10 patients on average).

Simulation study (II)

- To simulate the time to death due to cancer T_1
  $\lambda_E(t, \mathrm{Age}_{ij}, \mathrm{Sex}_{ij}, \mathrm{DI}_i) =$
  $\lambda_0(t) \exp\{\beta_{\mathrm{Age}} \mathrm{Age}_{ij} + \beta_{\mathrm{Sex}} \mathrm{Sex}_{ij} + \beta_{\mathrm{DI}} \mathrm{DI}_i + w_i\}$
  - Weibull baseline hazard $\lambda_0(t) = \lambda \rho t^{\rho-1}$ ($\lambda = 0.25$; $\rho = 0.7$)
  - Age effect ($\beta_{\mathrm{Age}} = 0.05$ for 1 year increase)
  - Sex effect ($\beta_{\mathrm{Sex}} = 1$, Men vs. women)
  - DI effect ($\beta_{\mathrm{DI}} = 0.02$ for 1 unit increase)
  - Random effect w_i: Normal distribution with mean 0 and standard deviation $\sigma = 0.25$ or 0.5 or 1

Simulation study (II)

- To simulate the time to death due to cancer T_1
  $\lambda_E(t, \mathrm{Age}_{ij}, \mathrm{Sex}_{ij}, \mathrm{DI}_i) =$
  $\lambda_0(t) \exp\{\beta_{\mathrm{Age}}\mathrm{Age}_{ij} + \beta_{\mathrm{Sex}}\mathrm{Sex}_{ij} + \beta_{\mathrm{DI}}\mathrm{DI}_i + w_i\}$
  - Weibull baseline hazard $\lambda_0(t) = \lambda\rho t^{\rho-1}$ ($\lambda = 0.25$; $\rho = 0.7$)
  - Age effect ($\beta_{\mathrm{Age}} = 0.05$ for 1 year increase)
  - Sex effect ($\beta_{\mathrm{Sex}} = 1$, Men vs. women)
  - DI effect ($\beta_{\mathrm{DI}} = 0.02$ for 1 unit increase)
  - Random effect w_i: Normal distribution with mean 0 and standard deviation $\sigma = 0.25$ or 0.5 or 1
- To simulate the time to death due to other causes T_2 : yearly piecewise exponential law using mortality rates from the population lifetable
- $\Rightarrow$ Final time $T = \min(T_1, T_2)$, with the corresponding vital status δ

Simulation study (III)

For the scenario **NPH**, two different Weibull baseline hazards for men and women:

- Times to cancer-death in men = Weibull (shape=0.7, scale=0.25)
- Times to cancer-death in women = Weibull (shape=0.8, scale=0.18).

$\Rightarrow$ the Hazard Ratio between Men vs. Women is time-dependent

For the scenario **Robustness** The random effect was drawn from a normal distribution with $\sigma = 0.5$ **but**

- with mean=-1 for the first half of the clusters, and
- with mean=1 for the other half

$\Rightarrow$ standard deviation of the resulting distribution is $\sqrt{(}1.25) \approx 1.12$.

Simulation study (IV)

The model used to analyse the data

- in scenarios balance- and unbalance- Design and Robustness

$$\lambda_E(t, \mathrm{Age}_{ij}, \mathrm{Sex}_{ij}, \mathrm{DI}_i) =$$
$$\lambda_0(t)\exp\{\beta_{\mathrm{Age}}\mathrm{Age}_{ij} + \beta_{\mathrm{Sex}}\mathrm{Sex}_{ij} + \beta_{\mathrm{DI}}\mathrm{DI}_i + w_i\}$$

With $\lambda_0(t)$ modelled either as a Weibull or using a cubic B-spline (1 knot at 1 year)

- in scenarios NPH

$$\lambda_E(t, \mathrm{Age}_{ij}, \mathrm{Sex}_{ij}, \mathrm{DI}_i) =$$
$$\lambda_0(t)\exp\{\beta_{\mathrm{Age}}\mathrm{Age}_{ij} + \beta_{\mathrm{Sex}}(t)\mathrm{Sex}_{ij} + \beta_{\mathrm{DI}}\mathrm{DI}_i + w_i\}$$

With $\lambda_0(t)$ and $\beta_{\mathrm{Sex}}(t)$ modelled using a cubic B-spline (1 knot at 1 year)

Overview of simulation results

Scenarios balance-Design, unbalance-Design and NPH

- Fixed-effect estimates of individual-level covariates unbiased and CP $\approx$ 95% whatever number and size of clusters, the level of heterogeneity simulated and the level of unbalance
- Same performances with B-spline instead of Weibull for the baseline hazard
- With small number of clusters (10 or 20), bias and CP less than 95% for cluster-level covariate (β_{DI}) and std.dev (σ) of the random effect
- RMSEs for β_{DI} and σ $\searrow$ when the number of clusters $\nearrow$
- Time-dependent effects correctly estimated

Scenario Robustness

- Fixed effect estimates of individual-level covariates unbiased and CP $\approx$ 95%
- Bias and bad CP for cluster-level covariate
- Bad CP for σ

R-package mexhaz

A R-package was developed: mexhaz, Mixed-effect EXcess HAZard model (available on the CRAN website https://cran.r-project.org/)
The mexhaz package allows

- to fit flexible hazard regression model
  - with/without introducing λ_P (i.e. to estimate overall or excess hazard)
  - with different baseline hazards: piecewise step function, Weibull or B-splines
  - with non-linear and/or time-dependent effect(s) of covariate(s)
  - with/without a random effect defined at the cluster level

R-package mexhaz

A R-package was developed: mexhaz, Mixed-effect EXcess HAZard model (available on the CRAN website https://cran.r-project.org/)

The mexhaz package allows

- to fit flexible hazard regression model
  - with/without introducing λ_P (i.e. to estimate overall or excess hazard)
  - with different baseline hazards: piecewise step function, Weibull or B-splines
  - with non-linear and/or time-dependent effect(s) of covariate(s)
  - with/without a random effect defined at the cluster level
- to predict the hazard and the corresponding survival
  - at several time points for one vector of covariates
  - for several vectors of covariates at one time point

R-package mexhaz

A R-package was developed: mexhaz, Mixed-effect EXcess HAZard model (available on the CRAN website https://cran.r-project.org/)

The mexhaz package allows

- to fit flexible hazard regression model
  - with/without introducing λ_P (i.e. to estimate overall or excess hazard)
  - with different baseline hazards: piecewise step function, Weibull or B-splines
  - with non-linear and/or time-dependent effect(s) of covariate(s)
  - with/without a random effect defined at the cluster level
- to predict the hazard and the corresponding survival
  - at several time points for one vector of covariates
  - for several vectors of covariates at one time point
- to plot the hazard and the corresponding survival

R-package mexhaz - Example of code

- **Estimation**

```
Mod1 <- mexhaz(formula=Surv(time=timesurv, event=vstat)~
agecr+depindex+IsexH+nph(agecr), data=simdatn1,
base="exp.bs", degree=3, knots=c(1,5), expected="popmrate",
random="clust")
```

R-package mexhaz - Example of code

- **Estimation**

```
Mod1 <- mexhaz(formula=Surv(time=timesurv, event=vstat)~
agecr+depindex+IsexH+nph(agecr), data=simdatn1,
base="exp.bs", degree=3, knots=c(1,5), expected="popmrate",
random="clust")
```

- **Prediction** at several time points for one vector of covariates

```
Pred_Mod1 <- predict(Mod1, time.pts=seq(0.1,10,by=0.1),
data.val=data.frame(agecr=0,depindex=0.5,IsexH=1),
conf.int="delta")
```

R-package mexhaz - Example of code

- **Estimation**

```
Mod1 <- mexhaz(formula=Surv(time=timesurv, event=vstat)~
agecr+depindex+IsexH+nph(agecr), data=simdatn1,
base="exp.bs", degree=3, knots=c(1,5), expected="popmrate",
random="clust")
```

- **Prediction** at several time points for one vector of covariates

```
Pred_Mod1 <- predict(Mod1, time.pts=seq(0.1,10,by=0.1),
data.val=data.frame(agecr=0,depindex=0.5,IsexH=1),
conf.int="delta")
```

- **Plot**

```
plot(Pred_Mod1, which="hazard")
```

Conclusions

We proposed an approach to fit a flexible excess hazard model, allowing for a random effect defined at the cluster level and time-dependent and/or non-linear effects of covariates [Charvat *StatMed* 2016]

- Numerical integration techniques:
  - Adaptive Gauss-Hermite Quadrature to calculate the cluster-specific marginal likelihood
  - Gauss-Legendre quadrature for the cumulative hazard
- Flexible functions (B-splines) used for the baseline and the time-dependent effects
- Good performances shown by simulation
- R-package available on the CRAN website

Illustration: Assessing the relationship between socio-economic environment and cancer survival in a French region

- Measure the socio-economic environment using a relevant and reproducible index on the whole population

Illustration: Assessing the relationship between socio-economic environment and cancer survival in a French region

- Measure the socio-economic environment using a relevant and reproducible index on the whole population
- Isolate cancer-specific mortality hazard

Illustration: Assessing the relationship between socio-economic environment and cancer survival in a French region

- Measure the socio-economic environment using a relevant and reproducible index on the whole population
- Isolate cancer-specific mortality hazard
- Enable possibly complex association (non-linear and/or time-dependent)

Illustration: Assessing the relationship between socio-economic environment and cancer survival in a French region

- Measure the socio-economic environment using a relevant and reproducible index on the whole population
- Isolate cancer-specific mortality hazard
- Enable possibly complex association (non-linear and/or time-dependent)
- Deal with hierarchical structure of the data (socio-economic environment is an ecological variable) for correct inference

Our choices

- Measure of the socioeconomic environment?
  ⇒ the European Deprivation Index (EDI), built to be reproducible [16]

Our choices

- Measure of the socioeconomic environment?
  ⇒ the European Deprivation Index (EDI), built to be reproducible [16]
- Impact on the whole population?
  ⇒ Population-based cancer registry data (exhaustive source of data)

Our choices

- Measure of the socioeconomic environment?
  ⇒ the European Deprivation Index (EDI), built to be reproducible [16]
- Impact on the whole population?
  ⇒ Population-based cancer registry data (exhaustive source of data)
- Cancer-specific mortality hazard without cause of death, and with possible complex effects?
  ⇒ Flexible Parametric Excess-hazard Model [4], with time-dependent and/or non-linear effects [8] combined with a model-building strategy [17]

Our choices

- Measure of the socioeconomic environment?
  ⇒ the European Deprivation Index (EDI), built to be reproducible [16]
- Impact on the whole population?
  ⇒ Population-based cancer registry data (exhaustive source of data)
- Cancer-specific mortality hazard without cause of death, and with possible complex effects?
  ⇒ Flexible Parametric Excess-hazard Model [4], with time-dependent and/or non-linear effects [8] combined with a model-building strategy [17]
- Correlated data / hierarchical structure?
  ⇒ Mixed-effect/multilevel models with a random effect defined at the cluster level (from which the socioeconomic environment was assessed) [1]

Material 1/2

- Data from Calvados and Manche population-based cancer registries
- Patients over 15 years, diagnosed between 1997 and 2010 and followed up to 30/06/2013
- 17 cancer sites analysed, separately in men and women
- R software and the package `mexhaz` we developped (Mixed-effect EXcess HAZard model)

Material 2/2

Indicators produced for each cancer-site combination

- Age-Standardised Net Survival (ASNS) predicted at 1, 5 and 10 years after diagnosis, by deprivation quintiles of the French population (ICSS weights)
- Variation with time since diagnosis of the Excess Mortality Hazard for 3 values of age and EDI (10^{th}, 50^{th} and 90^{th} percentiles)
- Excess Hazard Ratio for 1-unit increase of the EDI (may be non-linear and time-dependent)

Summary of the results

Age-Standardised Net Survival at 5 years

- In men, absolute difference (Dep 1 vs. Dep 5) $>$ 10% in Lip-Oral Cavity-Pharynx and melanoma. Around 5% in colon-rectum, bladder, kidney and prostate
- In women, absolute difference $>$ 10% in Lip-Oral Cavity-Pharynx. Around 5% in bladder, breast and melanoma

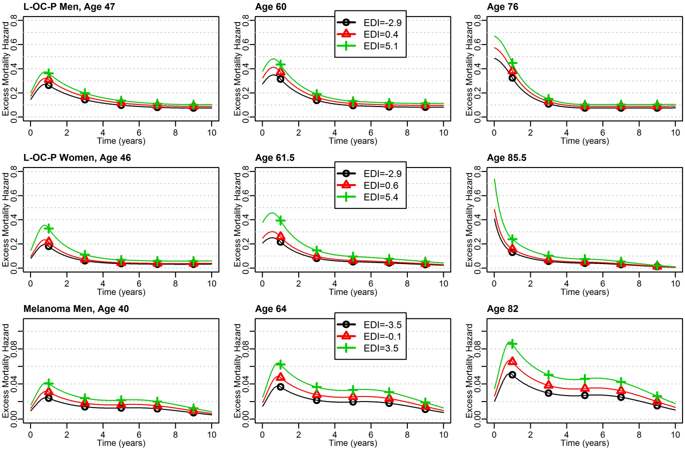
A linear and constant-in-time EDI's effect retained in most cases, except for

- Lip-Oral Cavity-Pharynx (NL effect in both sexes)
- Stomach (TD) and Pancreas (NL and TD) in men,
- Cervix uteri (NL and TD)

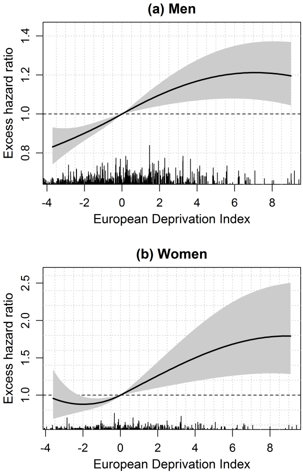
Paper available soon (hopefully) Belot *et al.* (*under review*) [2]

Variation with time since diagnosis of the excess mortality hazard according to EDI and age

For the 10^{th}, 50^{th} and 90^{th} percentiles of age and EDI distributions

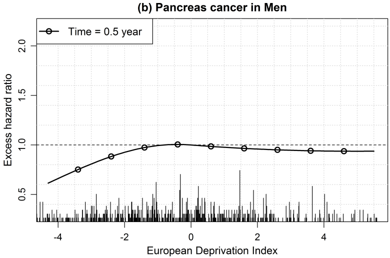


Non-linear effect of the EDI: Lip-Oral Cavity-Pharynx



Time-dependent and non-linear effect of the EDI
Example for pancreas cancer in men

Excess Hazard Ratio for EDI (Ref: EDI=0) at 6 months:

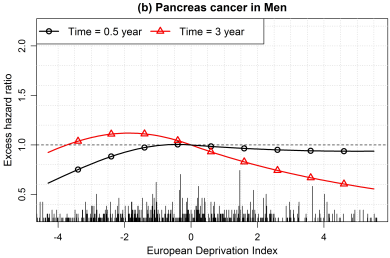


1-year ASNS by deprivation quintiles of the EDI (Q1-Q5):

Q1	Q2	Q3	Q4	Q5
36 [33;40]	26 [24;28]	23 [21;25]	24 [22;26]	25 [22;28]

Time-dependent and non-linear effect of the EDI Example for pancreas cancer in men

Excess Hazard Ratio for EDI (Ref: EDI=0) at 3 years:

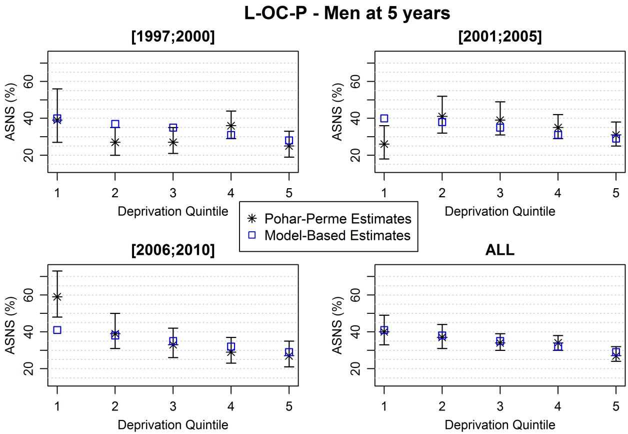


5-year ASNS by deprivation quintiles of the EDI (Q1-Q5):

Q1	Q2	Q3	Q4	Q5
8 [6;11]	4 [3;5]	4 [3;5]	5 [4;6]	7 [5;9]

Goodness of fit
Example for Lip-Oral Cavity-Pharynx

Predicted vs Non-Parametric ASNS (Pohar-Perme estimator [10])



Summary of our guidelines 1/3

Data

- Use data from a source that provides an unbiased picture of the whole population, such as population-based registries data
- Use an appropriate ecological deprivation measure, which can be (i) replicated in other countries (for comparison purposes); and, (ii) based on as small geographical unit as possible

Summary of our guidelines 2/3

Method

- Define the excess mortality hazard as your main quantity of interest
- Use flexible parametric multivariable regression models, which enable modelling non-linear as well as time-dependent effects of prognostic factors (such as the deprivation index)
- Take account of the multilevel/hierarchical structure of the data to derive correct statistical inference
- Use a model-building strategy or an information criterion to eliminate spurious non-linear and time-dependent effects

Summary of our guidelines 3/3

Results

- Provide model-based predictions of the ASNSs by deprivation quintile and compare them to the non-parametric estimates (to check the goodness-of-fit of the model)
- Give additional and clinically relevant information from the modelling approach:
    - the change with time since diagnosis of the excess mortality hazard for different values of the deprivation index
    - the Excess Hazard Ratios for the effect of the EDI (eventually non-linear and/or time-dependent)
- Quantify the impact of clustering on the excess mortality hazard using the General Contextual Effect and (whenever possible) an intra-class correlation coefficient

Conclusions/Discussion

- Those guidelines provide an efficient way to assess the association between socioeconomic inequalities and cancer survival
- Using flexible parametric models allow producing additional and relevant clinical information: variation with time of the excess-mortality hazard (instantaneous picture), non-linear and time-dependent effects
- Feasible with the R-package `mexhaz` we developped for this purpose

Conclusions/Discussion

- Those guidelines provide an efficient way to assess the association between socioeconomic inequalities and cancer survival
- Using flexible parametric models allow producing additional and relevant clinical information: variation with time of the excess-mortality hazard (instantaneous picture), non-linear and time-dependent effects
- Feasible with the R-package `mexhaz` we developped for this purpose

- No deprivation-specific life-table available in France, so probably slight over-estimation of the EDI's effect
- Studying the interactions between covariables (e.g. EDI and age) remains a challenge

Ongoing Work and Perspectives

- Extension of the R-package for allowing different shapes for the TD effects (degree/knots) than the ones used for the baseline hazard
- Extension to more than one random effect
- Use of this methodology to disentangle individual socioeconomic position from contextual deprivation

References I

Charvat H, Remontet L, Bossard N, Roche L, Dejardin O, Rachet B, Launoy G, Belot A; and the CENSUR working survival group.
A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates
Stat Med.2016; DOI 10.1002/sim.6881

Belot A, Remontet L, Rachet B, Dejardin O, Charvat H, Bara S, Guizard AV, Roche L, Launoy G, Bossard N.
Describing the association between socioeconomic inequalities and cancer survival: methodological guidelines and illustration with population-based data
Clinical Epi. Under review

Dupont C, Bossard N, Remontet L, Belot A.
Description of an approach based on maximum likelihood to adjust an excess hazard model with a random effect
Cancer Epidemiol.2013;37:449-56

References II

Estève J, Benhamou E, Croasdale M, Raymond L.
Relative survival and the estimation of net survival: elements for further discussion
Stat Med.1990;9:529-38.

Bolard P, Quantin C, Abrahamowicz M, Esteve J, Giorgi R, Chadha-Boreham H, Binquet C, Faivre J.
Assessing time-by-covariate interactions in relative survival models using restrictive cubic spline functions
J Cancer Epidemiol Prev.2002;7:11322.

Giorgi R, Abrahamowicz M, Quantin C, Bolard P, Esteve J, Gouvernet J, Faivre J.
A relative survival regression model using B-spline functions to model non-proportional hazards
Stat Med.2003;22:276784.

References III

Lambert PC, Smith LK, Jones DR, Botha JL
Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects
Stat Med.2005;24:387185.

Remontet L, Bossard N, Belot A, Estève J.
An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies
Stat Med.2007;26:2214-28.

Nelson CP, Lambert PC, Squire IB, Jones DR.
Flexible parametric models for relative survival, with application in coronary heart disease
Stat Med.2007;26:5486-98.

Pohar-Perme M, Henderson R, Stare J.
An approach to estimation in relative survival regression
Biostat.2009;10:136-46.

References IV

Tuerlinckx F, Rijmen F, Verbeke G, De Boeck P.
Statistical inference in generalized linear mixed models: a review
Br J Math Stat Psychol.2006;59:225-55.

Duchateau L, Janssen P.
The frailty model
Springer,2008.

Wienke A.
Frailty Models in Survival Analysis
Chapman and Hall/CRC,2010.

Liu Q, Pierce DA.
A note on Gauss-Hermite quadrature
Biometrika.1994;81:624-9.

Pinheiro JC, Bates DM.
Approximations to the log-likelihood function in the non-linear mixed-effects model
J Comput Graph Stat.1995;4:12-35.

References V

Pornet C, Delpierre C, Dejardin O, Grosclaude P, Launay L, Guittet L, Lang T, Launoy G.
Construction of an adaptable European transnational ecological deprivation index: the French version
J Epidemiol Community Health.2012;66:982-9.

Wynant W. and Abrahamowicz M.
Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis
Stat Med.2014;33:3318-37.

Austin PC, Wagner P, Merlo J.
The median hazard ratio: a useful measure of variance and general contextual effects in multilevel survival analysis
Stat Med.2017;36:928-38

APPENDIX

LAPLACE approximation

Let g be a strictly positive, unimodal function with mode μ_g and let us define l such that $l(x) = \log\{g(x)\}$.

In a neighbourhood of μ_g:

$$l(x) \approx l(\mu_g) + (x - \mu_g) l'(\mu_g) + \frac{(x-\mu_g)^2}{2} l''(\mu_g)$$

- μ_g extremum $\Rightarrow l'(\mu_g) = 0$
- μ_g maximum $\Rightarrow l''(\mu_g) < 0$

$$g(x) \approx g(\mu_g) \underbrace{\exp\left\{\frac{(x-\mu_g)^2}{2} l''(\mu_g)\right\}}_{\propto\, \phi(x,\mu_g,\sigma_g)} \quad \text{with} \quad \sigma_g = \frac{1}{\sqrt{-l''(\mu_g)}}$$

Median Excess Hazard Ratio

Indicators produced for each cancer-site combination:
General Contextual Effect, Median Excess Hazard Ratio with and without adjusting on the EDI (median of HRs comparing 2 patients randomly selected from 2 clusters with higher vs. lower excess mortality)

Results
Important General Contextual effect in Lip-Oral Cavity-Pharynx in both sexes, and in men for prostate, melanoma and pancreas (for those sex-cancer, EDI explains an important part of the variability between clusters)

Strategy of analysis

Model-building strategy [Wynant 2014]

Separately for each gender,

- Start from the most complex multilevel excess hazard model
  - Non-linear and time-dependent effects for the continuous covariables age, year of diagnosis and EDI (quadratic B-splines, with knots located at 1 and 5 years for baseline and TD effects, and at 70-years, in 2000 and at 0 for NLIN effect of age, year and EDI, resp.)
  - a random effect defined at the cluster level (normal distribution with mean 0 and standard deviation σ)

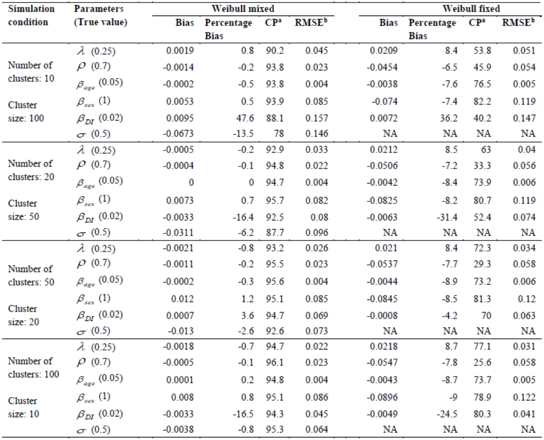
$$\boxed{\lambda_E(t,a,y,i \mid w) = \lambda_0(t) \cdot \exp(g(a) + h(t)a + j(y) + k(t)y + m(i) + n(t)i + w)}$$

- Backward Elimination procedure to successively eliminate spurious non-linear and time-dependent effects

Note: Compared to Wynant's proposal, we kept all the main effects

Simulation results

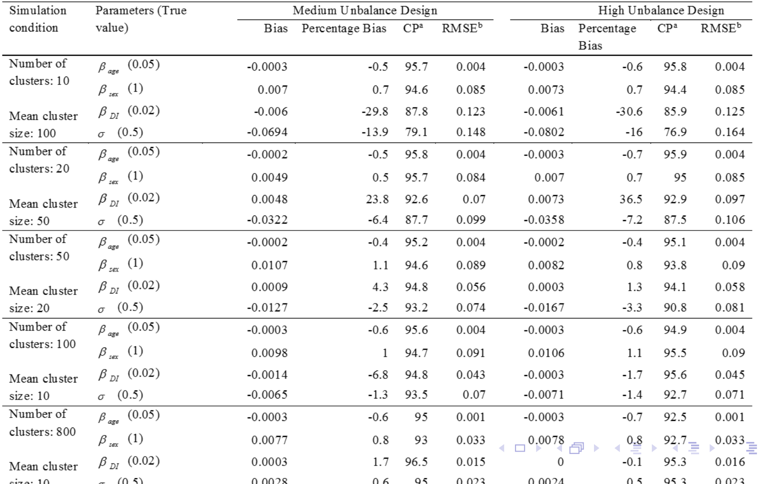
What about neglecting the hierarchical structure of the data ?



Simulation condition	Parameters (True value)	Weibull mixed				Weibull fixed			
		Bias	Percentage Bias	CP^{a}	RMSE^{b}	Bias	Percentage Bias	CP^{a}	RMSE^{b}
Number of clusters: 10 Cluster size: 100	λ (0.25)	0.0019	0.8	90.2	0.045	0.0209	8.4	53.8	0.051
	ρ (0.7)	-0.0014	-0.2	93.8	0.023	-0.0454	-6.5	45.9	0.054
	β_{age} (0.05)	-0.0002	-0.5	93.8	0.004	-0.0038	-7.6	76.5	0.005
	β_{sex} (1)	0.0053	0.5	93.9	0.085	-0.074	-7.4	82.2	0.119
	β_{DI} (0.02)	0.0095	47.6	88.1	0.157	0.0072	36.2	40.2	0.147
	σ (0.5)	-0.0673	-13.5	78	0.146	NA	NA	NA	NA
Number of clusters: 20 Cluster size: 50	λ (0.25)	-0.0005	-0.2	92.9	0.033	0.0212	8.5	63	0.04
	ρ (0.7)	-0.0004	-0.1	94.8	0.022	-0.0506	-7.2	33.3	0.056
	β_{age} (0.05)	0	0	94.7	0.004	-0.0042	-8.4	73.9	0.006
	β_{sex} (1)	0.0073	0.7	95.7	0.082	-0.0825	-8.2	80.7	0.119
	β_{DI} (0.02)	-0.0033	-16.4	92.5	0.08	-0.0063	-31.4	52.4	0.074
	σ (0.5)	-0.0311	-6.2	87.7	0.096	NA	NA	NA	NA
Number of clusters: 50 Cluster size: 20	λ (0.25)	-0.0021	-0.8	93.2	0.026	0.021	8.4	72.3	0.034
	ρ (0.7)	-0.0011	-0.2	95.5	0.023	-0.0537	-7.7	29.3	0.058
	β_{age} (0.05)	-0.0002	-0.3	95.6	0.004	-0.0044	-8.9	73.2	0.006
	β_{sex} (1)	0.012	1.2	95.1	0.085	-0.0845	-8.5	81.3	0.12
	β_{DI} (0.02)	0.0007	3.6	94.7	0.069	-0.0008	-4.2	70	0.063
	σ (0.5)	-0.013	-2.6	92.6	0.073	NA	NA	NA	NA
Number of clusters: 100 Cluster size: 10	λ (0.25)	-0.0018	-0.7	94.7	0.022	0.0218	8.7	77.1	0.031
	ρ (0.7)	-0.0005	-0.1	96.1	0.023	-0.0547	-7.8	25.6	0.058
	β_{age} (0.05)	0.0001	0.2	94.8	0.004	-0.0043	-8.7	73.7	0.005
	β_{sex} (1)	0.008	0.8	95.1	0.086	-0.0896	-9	78.9	0.122
	β_{DI} (0.02)	-0.0033	-16.5	94.3	0.045	-0.0049	-24.5	80.3	0.041
	σ (0.5)	-0.0038	-0.8	95.3	0.064	NA	NA	NA	NA

Simulation results

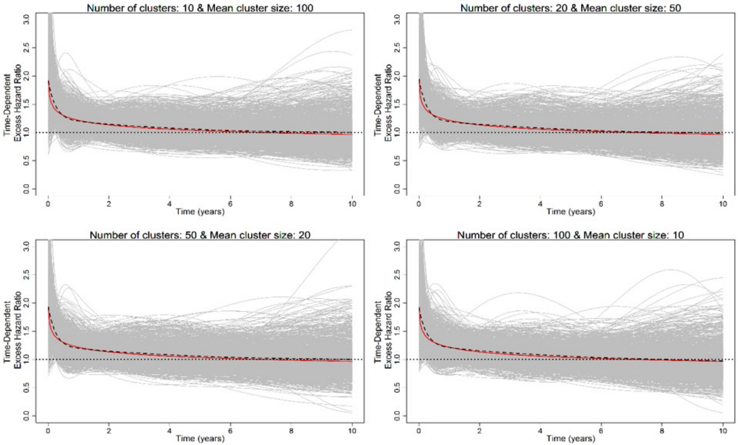
Scenario unbalance-Design



Simulation condition	Parameters (True value)	Medium Unbalance Design				High Unbalance Design			
		Bias	Percentage Bias	CP[a]	RMSE[b]	Bias	Percentage Bias	CP[a]	RMSE[b]
Number of clusters: 10	β_{age} (0.05)	-0.0003	-0.5	95.7	0.004	-0.0003	-0.6	95.8	0.004
	β_{sex} (1)	0.007	0.7	94.6	0.085	0.0073	0.7	94.4	0.085
Mean cluster size: 100	β_{DI} (0.02)	-0.006	-29.8	87.8	0.123	-0.0061	-30.6	85.9	0.125
	σ (0.5)	-0.0694	-13.9	79.1	0.148	-0.0802	-16	76.9	0.164
Number of clusters: 20	β_{age} (0.05)	-0.0002	-0.5	95.8	0.004	-0.0003	-0.7	95.9	0.004
	β_{sex} (1)	0.0049	0.5	95.7	0.084	0.007	0.7	95	0.085
Mean cluster size: 50	β_{DI} (0.02)	0.0048	23.8	92.6	0.07	0.0073	36.5	92.9	0.097
	σ (0.5)	-0.0322	-6.4	87.7	0.099	-0.0358	-7.2	87.5	0.106
Number of clusters: 50	β_{age} (0.05)	-0.0002	-0.4	95.2	0.004	-0.0002	-0.4	95.1	0.004
	β_{sex} (1)	0.0107	1.1	94.6	0.089	0.0082	0.8	93.8	0.09
Mean cluster size: 20	β_{DI} (0.02)	0.0009	4.3	94.8	0.056	0.0003	1.3	94.1	0.058
	σ (0.5)	-0.0127	-2.5	93.2	0.074	-0.0167	-3.3	90.8	0.081
Number of clusters: 100	β_{age} (0.05)	-0.0003	-0.6	95.6	0.004	-0.0003	-0.6	94.9	0.004
	β_{sex} (1)	0.0098	1	94.7	0.091	0.0106	1.1	95.5	0.09
Mean cluster size: 10	β_{DI} (0.02)	-0.0014	-6.8	94.8	0.043	-0.0003	-1.7	95.6	0.045
	σ (0.5)	-0.0065	-1.3	93.5	0.07	-0.0071	-1.4	92.7	0.071
Number of clusters: 800	β_{age} (0.05)	-0.0003	-0.6	95	0.001	-0.0003	-0.7	92.5	0.001
	β_{sex} (1)	0.0077	0.8	93	0.033	0.0078	0.8	92.7	0.033
Mean cluster size: 10	β_{DI} (0.02)	0.0003	1.7	96.5	0.015	0	-0.1	95.3	0.016
	σ (0.5)	0.0028	0.6	95	0.023	0.0024	0.5	95.3	0.023

Simulation results

Scenario NPH



Quick reminder on survival quantities 1/2

Survival at time t : $S(t) = P(T \geq t)$

Instantaneous mortality hazard :

$$\lambda(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t \,|\, T \geq t)}{\delta t} \right\}$$

Cumulative Mortality hazard : $\Lambda(t) = \displaystyle\int_0^t \lambda(u)\, \mathrm{d}u$

The following relationship holds :

$$S(t) = \exp\{-\Lambda(t)\} \qquad\qquad S(t) = 1 - \int_0^t \lambda(u) \cdot S(u)\, \mathrm{d}u$$

Quick reminder on survival quantities 2/2

For each patient j, we observe:

- the time to death (or of last known vital status) t_j
- the failure indicator δ_j
- possibly some covariates $\mathbf{x}_j$

The Log-Likelihood (assuming non informative censoring)

$$loglik = \prod_{j=1}^{N} S(t_j, \mathbf{x}_j)\lambda(t_j, \mathbf{x}_j)^{\delta_j}$$