Flexible Parametric Excess Hazard Regression Models: Application to population-based cancer registry data

Aurélien Belot

Cancer Survival Group
Non-communicable Disease Epidemiology
London School of Hygiene and Tropical Medicine

Granada, 27-28 March 2017

Intended Learning Outcomes

At the end of this lecture, you should be able to

- Define and explain what is a flexible parametric hazard model (FPM) and mention its advantages
- Interpret the results from a FPM applied in the relative survival setting
- Appreciate and interpret time-dependent effects
- Fit a flexible parametric hazard model (FPM) using R (package `mexhaz`) and Stata (command `strcs`)

Intended Learning Outcomes

At the end of this lecture, you should be able to

- Define and explain what is a flexible parametric hazard model (FPM) and mention its advantages
- Interpret the results from a FPM applied in the relative survival setting
- Appreciate and interpret time-dependent effects
- Fit a flexible parametric hazard model (FPM) using R (package `mexhaz`) and Stata (command `strcs`)

Intended Learning Outcomes

At the end of this lecture, you should be able to

- Define and explain what is a flexible parametric hazard model (FPM) and mention its advantages
- Interpret the results from a FPM applied in the relative survival setting
- Appreciate and interpret time-dependent effects
- Fit a flexible parametric hazard model (FPM) using R (package `mexhaz`) and Stata (command `strcs`)

Intended Learning Outcomes

At the end of this lecture, you should be able to

- Define and explain what is a flexible parametric hazard model (FPM) and mention its advantages
- Interpret the results from a FPM applied in the relative survival setting
- Appreciate and interpret time-dependent effects
- Fit a flexible parametric hazard model (FPM) using R (package `mexhaz`) and Stata (command `strcs`)

Intended Learning Outcomes

At the end of this lecture, you should be able to

- Define and explain what is a flexible parametric hazard model (FPM) and mention its advantages
- Interpret the results from a FPM applied in the relative survival setting
- Appreciate and interpret time-dependent effects
- Fit a flexible parametric hazard model (FPM) using R (package `mexhaz`) and Stata (command `strcs`)

Regression Model 1/2

Some general thoughts on “what is a model”:

- From Wikipedia: A statistical model is a class of mathematical model, which embodies a set of assumptions concerning the generation of some sample data, and similar data from a larger population. A statistical model represents, often in considerably idealized form, the data-generating process.
- A simplification or approximation of reality (Burnham and Anderson, 2002)
- A powerful tool for developing and testing theories by way of causal explanation, prediction, and description (Shmueli, 2010)

Regression Model 2/2

Different types of regression models depending on the main objective (G. Shmueli, Statistical Science 2010, “To Explain or to predict?”):

- **Descriptive modelling**: summarising or representing the data structure in a compact manner
- **Explanatory modelling**: applying statistical models to data to explain an association between variables and an outcome, and eventually testing causal explanations/hypotheses
- **Projection (Predictive modelling)**: applying a statistical model to data for the purpose of predicting future observations

In this session, focus on descriptive and explanatory modelling

Net survival and excess mortality hazard model

Classical methods used to analyse population-based cancer registry data

- Net survival: useful for comparison between countries in their ability to manage (broad sense) cancer patients, after eliminating other causes of death (potentially different between 2 countries)
- Non-parametric estimator of net survival exists, the Pohar-Perme estimator (same spirit as Nelson Aalen's estimator)
- Excess mortality hazard approach: allows to quantify the mortality due to cancer (still without knowing the cause of death)

In this session, focus on regression models for the excess mortality hazard

Net survival and excess mortality hazard model

Classical methods used to analyse population-based cancer registry data

- Net survival: useful for comparison between countries in their ability to manage (broad sense) cancer patients, after eliminating other causes of death (potentially different between 2 countries)
- Non-parametric estimator of net survival exists, the Pohar-Perme estimator (same spirit as Nelson Aalen's estimator)
- Excess mortality hazard approach: allows to quantify the mortality due to cancer (still without knowing the cause of death)

In this session, focus on regression models for the excess mortality hazard

Excess mortality hazard regression model 1/2

Overall mortality hazard $\lambda(t;\mathbf{x}_j)$: expressed as the sum of (i) an excess mortality hazard (due to cancer) λ_E and (ii) the population (expected) mortality hazard λ_P

Equation for the excess mortality hazard

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

Where

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

Excess mortality hazard regression model 1/2

Overall mortality hazard $\lambda(t;\mathbf{x}_j)$: expressed as the sum of (i) an excess mortality hazard (due to cancer) λ_E and (ii) the population (expected) mortality hazard λ_P

Equation for the excess mortality hazard

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

Where

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

Excess mortality 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 population mortality hazard λ_P is considered known (usually obtained from Office for national statistics in life-table format)
- **The quantity to estimate is λ_E**
- This excess hazard is associated with the net survival (from the classical relationship between hazard and survival) $S(t) = \exp\left(-\int_0^t \lambda(u)\right) \mathrm{d}u$

Existing regression models: a brief review

Different regression models have been developed during the last 30 years for fitting excess mortality hazard regression models

Additive decomposition of the overall mortality hazard

$\lambda_{obs} = \lambda_E + \lambda_P$ and $\lambda_E(t;x) = \lambda_0(t)\exp(\beta x)$

- Hakulinen et al., 1987 Biometrics: GLM implementation on grouped data, baseline step function, categorical variables
- Esteve et al., 1990 Stat Med: Maximum Likelihood estimation on individual data, baseline step function
- Dickman et al., 2004 Stat Med: GLM implementation (Poisson model with user-defined link function) of the Esteve et al. model on split data

Existing regression models: a brief review

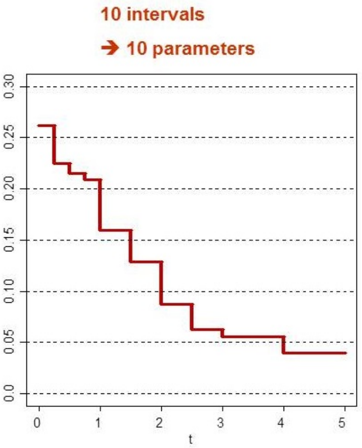
Different regression models have been developed during the last 30 years for fitting excess mortality hazard regression models

Additive decomposition of the overall mortality hazard

$\lambda_{obs} = \lambda_E + \lambda_P$ and $\lambda_E(t;x) = \lambda_0(t)\exp(\beta x)$

- Hakulinen et al., 1987 Biometrics: GLM implementation on grouped data, baseline step function, categorical variables
- Esteve et al., 1990 Stat Med: Maximum Likelihood estimation on individual data, baseline step function
- Dickman et al., 2004 Stat Med: GLM implementation (Poisson model with user-defined link function) of the Esteve et al. model on split data

Step or smooth function for the baseline excess mortality hazard



Existing flexible parametric regression models

Additive decomposition of the overall mortality hazard
$\lambda_{obs} = \lambda_E + \lambda_P$ and $\lambda_E(t;x) = \lambda_0(t)\exp(\beta x)$

- Bolard et al., 2002 JECP: Quadratic regression splines, time-dependent (TD) effects
- Giorgi et al., 2003 Stat Med: Quadratic B-splines, TD effects, package R (`RSurv`)
- Lambert et al., 2005 Stat Med: Fractional polynomials, TD effects
- Remontet et al., 2007 Stat Med: Regression splines, TD and non-linear (NLIN) effects ($f(t)*age+g(age)$) , package R (`flexrsurv`)
- Mahboubi et al., 2011 Stat Med: Regression splines, TD and NLIN effects ($f(t)*g(age)$), package R (`flexrsurv`)
- Charvat et al., 2016 Stat Med: Regression splines, TD and NLIN effects, random effects, package R (`mexhaz`)

Other existing flexible regression models

Additive decomposition of the overall mortality hazard $\lambda_{obs} = \lambda_E + \lambda_P$

Models assuming $\lambda_E(t;x) = \lambda_0(t)\exp(\beta x)$

- Pohar et al., Biostatistics 2009: EM algorithm, baseline left unspecified (Semi parametric excess hazard model)

Models assuming $\lambda_E(t;x) = \lambda_0(t) + \beta(t)x$

- Zahl et al., LDA 1998
- Cortese et al., Stat Med 2008

Models on the cumulative hazard scale

- Nelson et al., Stat Med 2007

Multiplicative decomposition of the overall mortality hazard $\lambda_{obs} = \lambda_E * \lambda_P$

- Andersen et al., 1985 Biometrics

Other existing flexible regression models

Additive decomposition of the overall mortality hazard $\lambda_{obs} = \lambda_E + \lambda_P$

Models assuming $\lambda_E(t;x) = \lambda_0(t)\exp(\beta x)$

- Pohar et al., Biostatistics 2009: EM algorithm, baseline left unspecified (Semi parametric excess hazard model)

Models assuming $\lambda_E(t;x) = \lambda_0(t) + \beta(t)x$

- Zahl et al., LDA 1998
- Cortese et al., Stat Med 2008

Models on the cumulative hazard scale

- Nelson et al., Stat Med 2007

Multiplicative decomposition of the overall mortality hazard $\lambda_{obs} = \lambda_E * \lambda_P$

- Andersen et al., 1985 Biometrics

Flexible parametric hazard model (FPM) 1/2

In this session, focus on Flexible parametric regression models for the excess mortality hazard modelled on the hazard scale, and assuming

Additive decomposition of the overall mortality hazard

$$\lambda_{obs} = \lambda_E + \lambda_P$$

Regression models of the form

$$\lambda_E(t;\mathbf{x}) = \lambda_0(t)\exp(\nu(t,\mathbf{x}))$$

Following the work published recently by Charvat et al. (Statistics in Medicine 2016, doi: 10.1002/sim.6881), with the associated R-package `mexhaz`

Flexible parametric hazard model (FPM) 2/2

Definition

$$\lambda_E(t;\mathbf{x}) = \lambda_0(t) \cdot \exp\Big(\sum_{a=1}^{A} \beta_a x_a + \sum_{b=1}^{B} f_b(t;\xi_b) x_b \Big)$$

- $\lambda_0(t)$ is the baseline excess hazard function
- The variables $x_a, (a = 1, \ldots, A)$ have a proportional effect (possibly non-linear if one specific x_a corresponds for example to the square of the original variable)
- The variables $x_b, (b = 1, \ldots, B)$ have a time-dependent effect modelled with flexible functional forms f_b
- Based on classical Maximum Likelihood theory

Flexible functional form: use of regression splines

- Flexible mathematical functions defined by piecewise polynomials (usually degree 2 or 3), which join at pre-specified points called knots
- Forced to have continuous 0^{th}, 1^{st} and 2^{nd} derivatives (ensure smoothness) for splines of degree 3
- Regression splines are linear in the regression coefficients, so we can use standard method of inference
- Regression splines can be incorporated into any regression model with a linear predictor

Flexible functional forms: Examples of regression splines

Spline of degree 3, with 1 knot at t=2 (truncated power basis)

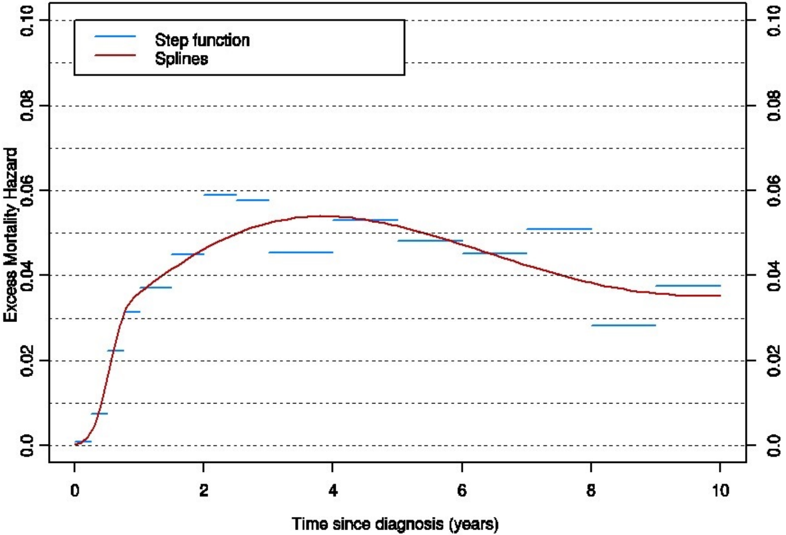
$$s(t) = a + bt + ct^2 + dt^3 + e(t-2)_+^3$$

where $(u)_+ = 0$ if $u \leq 0$ and $u_+ = u$ if $u > 0$

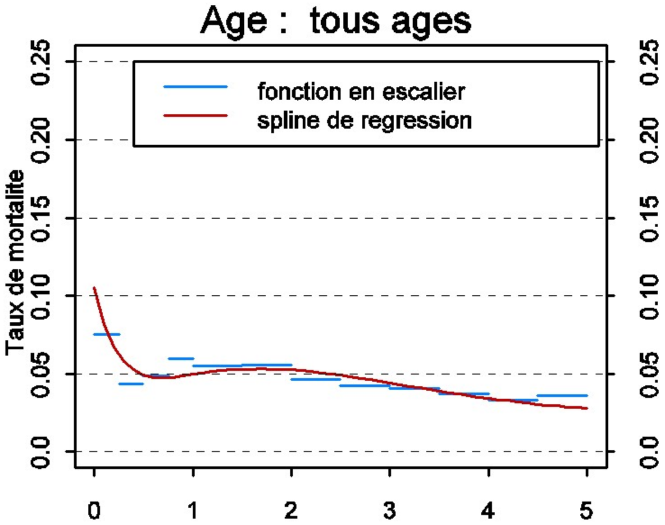
Spline of degree 2, with 2 knots at t=1 and 5 (truncated power basis)

$$s(t) = a + bt + ct^2 + d(t-1)_+^2 + e(t-5)_+^2$$

Flexible functional forms: Examples of regression splines



Flexible functional forms: Examples of regression splines



Flexible functional forms: regression splines

Restricted cubic regression splines: Regression splines that are forced to be linear before and after the boundary knots

General expression, for a restricted cubic regression splines with K knots

$$s(t) = \gamma_0 + \sum_{i=1}^{K-1} \gamma_i B_i(t)$$

where $B_1(t) = t$ and
$B_i(t), i = 2, ..., K-1$ define the basis, according to the knot k_i and on the first and last knot k_1 and k_K
For more details, see Durrleman and Simon, Stat Med 1989

Definition of the likelihood - General

Assuming non-informative right censoring, the contribution to the log-likelihood of individual j with observed data O_j (β denotes the vector of parameters to be estimated):

$$\begin{aligned} \mathrm{LL}_j(\beta; O_j) &= \log\left(S(t_j; \mathbf{x}_j)\right) + \delta_j \cdot \log(\lambda(t_j; \mathbf{x}_j)) \\ &= -\int_0^{t_j} \lambda(u; \mathbf{x}_j)\,\mathrm{d}u + \delta_j \cdot \log(\lambda(t_j; \mathbf{x}_j)) \end{aligned}$$

The **full log-likelihood** LL is defined as the sum of the individuals' contribution LL_j

$$\mathrm{LL}(\beta; O_j) = \sum_{j=1}^{N} \mathrm{LL}_j(\beta; O_j) \tag{1}$$

Maximized using an optimisation routine (e.g. Newton-Raphson method)

Definition of the likelihood - For excess hazard regression model

Individual's contribution

$$\mathrm{LL}_j^E(\beta; O_j) = -\int_0^{t_j} \left\{ \lambda_E(u; \mathbf{x}_j) + \lambda_P(a_j + u; y_j + u; \mathbf{z}_j) \right\} \mathrm{d}u +$$
$$\delta_j \cdot \log \left\{ \lambda_E(t_j; \mathbf{x}_j) + \lambda_P(a_j + t_j; y_j + t_j; \mathbf{z}_j) \right\}$$

Involves an integral of the overall hazard: use of numerical integration (Gauss Legendre quadrature in R-`mexhaz` and Stata-`strcs`)

The **full log-likelihood** is the sum of the individuals' contribution:

$$\mathrm{LL}^E(\beta; O_j) = \sum_{j=1}^{N} \mathrm{LL}_j^E(\beta; O_j)$$

Definition of the likelihood - For excess hazard regression model

Individual's contribution

$$\mathrm{LL}_j^E(\beta; O_j) = -\int_0^{t_j} \left\{ \lambda_E(u; \mathbf{x}_j) + \lambda_P(a_j + u; y_j + u; \mathbf{z}_j) \right\} \mathrm{d}u +$$
$$\delta_j \cdot \log \left\{ \lambda_E(t_j; \mathbf{x}_j) + \lambda_P(a_j + t_j; y_j + t_j; \mathbf{z}_j) \right\}$$

Involves an integral of the overall hazard: use of numerical integration (Gauss Legendre quadrature in R-`mexhaz` and Stata-`strcs`)

The **full log-likelihood** is the sum of the individuals' contribution:

$$\mathrm{LL}^E(\beta; O_j) = \sum_{j=1}^{N} \mathrm{LL}_j^E(\beta; O_j)$$

Definition of the likelihood - For excess hazard regression model

Individual's contribution

$$\mathrm{LL}_j^E(\beta;O_j) = -\int_0^{t_j} \left\{\lambda_E(u;\mathbf{x}_j) + \lambda_P(a_j+u;y_j+u;\mathbf{z}_j)\right\} \mathrm{d}u +$$
$$\delta_j \cdot \log\left\{\lambda_E(t_j;\mathbf{x}_j) + \lambda_P(a_j+t_j;y_j+t_j;\mathbf{z}_j)\right\}$$

Involves an integral of the overall hazard: use of numerical integration (Gauss Legendre quadrature in R-`mexhaz` and Stata-`strcs`)

The **full log-likelihood** is the sum of the individuals' contribution:

$$\mathrm{LL}^E(\beta;O_j) = \sum_{j=1}^{N} \mathrm{LL}_j^E(\beta;O_j)$$

Individual's contribution to the likelihood

Exercise: Give the mathematical expression of the individual's contribution to the log-likelihood for the following 3 observations, assuming the following model for the excess hazard:

$$\lambda_E(t;\mathbf{x}_j) = \lambda \cdot \exp(\beta_1\, agediag + \beta_2\, I(sex = M))$$

(Female, value sex=2 is the reference)
λ_P is the Population mortality hazard at the end of the follow-up

Id	Agediag	Sex	Time	Dead	λ_P
1	64	1	5	0	0.128
2	78	2	3.7	1	0.281
3	51	1	2.8	1	0.047

Individual's contribution to the likelihood

Solution for the first individual

Id	Agediag	Sex	Time	Dead	λ_P
1	64	1	5	0	0.128
2	78	2	3.7	1	0.281
3	51	1	2.8	1	0.047

$$\mathrm{LL}_1^E(\beta; O_1) = -\int_0^5 \left\{\lambda \cdot \exp(64\beta_1 + \beta_2) + 0.128\right\} \mathrm{d}u$$

$$\mathrm{LL}_1^E(\beta; O_1) = -5 \times \left\{\lambda \cdot \exp(64\beta_1 + \beta_2) + 0.128\right\}$$

Individual's contribution to the likelihood

Solution for the second individual

Id	Agediag	Sex	Time	Dead	λ_P
1	64	1	5	0	0.128
2	78	2	3.7	1	0.281
3	51	1	2.8	1	0.047

$$\mathrm{LL}_2^E(\beta; O_2) = -\int_0^{3.7} \left\{\lambda \cdot \exp(78\beta_1) + 0.281\right\} \mathrm{d}u + \log\left\{\lambda \cdot \exp(78\beta_1) + 0.281\right\}$$

$$\mathrm{LL}_2^E(\beta; O_2) = -3.7 \times \left\{\lambda \cdot \exp(78\beta_1) + 0.281\right\} \mathrm{d}u + \log\left\{\lambda \cdot \exp(78\beta_1) + 0.281\right\}$$

Individual's contribution to the likelihood

Solution for the third individual

Id	Agediag	Sex	Time	Dead	λ_P
1	64	1	5	0	0.128
2	78	2	3.7	1	0.281
3	51	1	2.8	1	0.047

$$\mathrm{LL}_3^E(\beta; O_3) = -\int_0^{2.8} \left\{\lambda \cdot \exp(51\beta_1 + \beta_2) + 0.047\right\} \mathrm{d}u + \log\left\{\lambda \cdot \exp(51\beta_1 + \beta_2) + 0.047\right\}$$

$$\mathrm{LL}_3^E(\beta; O_3) = -2.8 \times \left\{\lambda \cdot \exp(51\beta_1 + \beta_2) + 0.047\right\} \mathrm{d}u + \log\left\{\lambda \cdot \exp(51\beta_1 + \beta_2) + 0.047\right\}$$

How to build a regression model ?

Reminder: depends on the research question. To describe, to explain or to predict.

- The regression model needs to be adjusted for each life table variable to properly account for informative censoring
- Flexible functional forms for time-dependent and non-linear effects
- Interactions between variables (still an active research area)
- Information criterion as the Akaike Information Criteria may be used to choose the best fitting model (also the BIC)
- Another possibility to describe the association between (some) variable(s) and an outcome: model building strategy proposed by Wynant et al. (Wynant, Stat Med 2014)

How to build a regression model ?

Reminder: depends on the research question. To describe, to explain or to predict.

- The regression model needs to be adjusted for each life table variable to properly account for informative censoring
- Flexible functional forms for time-dependent and non-linear effects
- Interactions between variables (still an active research area)
- Information criterion as the Akaike Information Criteria may be used to choose the best fitting model (also the BIC)
- Another possibility to describe the association between (some) variable(s) and an outcome: model building strategy proposed by Wynant et al. (Wynant, Stat Med 2014)

How to build a regression model ?

Reminder: depends on the research question. To describe, to explain or to predict.

- The regression model needs to be adjusted for each life table variable to properly account for informative censoring
- Flexible functional forms for time-dependent and non-linear effects
- Interactions between variables (still an active research area)
- Information criterion as the Akaike Information Criteria may be used to choose the best fitting model (also the BIC)
- Another possibility to describe the association between (some) variable(s) and an outcome: model building strategy proposed by Wynant et al. (Wynant, Stat Med 2014)

How to build a regression model ?

Reminder: depends on the research question. To describe, to explain or to predict.

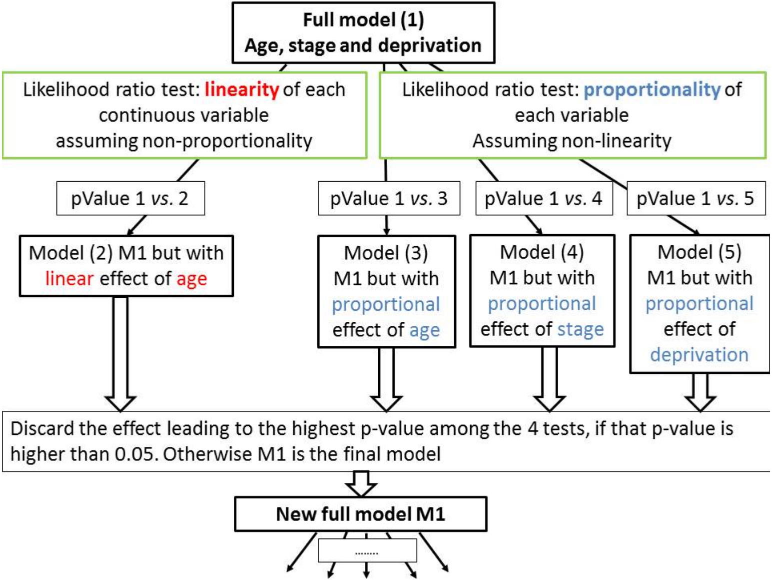
- The regression model needs to be adjusted for each life table variable to properly account for informative censoring
- Flexible functional forms for time-dependent and non-linear effects
- Interactions between variables (still an active research area)
- Information criterion as the Akaike Information Criteria may be used to choose the best fitting model (also the BIC)
- Another possibility to describe the association between (some) variable(s) and an outcome: model building strategy proposed by Wynant et al. (Wynant, Stat Med 2014)

How to build a regression model ?

Reminder: depends on the research question. To describe, to explain or to predict.

- The regression model needs to be adjusted for each life table variable to properly account for informative censoring
- Flexible functional forms for time-dependent and non-linear effects
- Interactions between variables (still an active research area)
- Information criterion as the Akaike Information Criteria may be used to choose the best fitting model (also the BIC)
- Another possibility to describe the association between (some) variable(s) and an outcome: model building strategy proposed by Wynant et al. (Wynant, Stat Med 2014)

How to build a regression model ?



Illustration

Data

- Men diagnosed in 2000-2002 with colon cancer in England
- Variables available:
  - age
  - stage (4 categories)
  - deprivation (5 categories)

Aim: To describe the association between age at diagnosis and the excess mortality hazard

First explanatory model:

$$\lambda_E(t;x) = \lambda_0(t)\exp\Big(\sum_{i=2}^{4}\alpha_i stage_i + \sum_{i=2}^{5}\beta_i dep_i + \sum_{i=2}^{5}\gamma_i agecat_i\Big)$$

where $\lambda_0(t)$ is the (exponential of a) B-spline of degree 3, with 1 knot located at 1 year

Illustration

Data

- Men diagnosed in 2000-2002 with colon cancer in England
- Variables available:
  - age
  - stage (4 categories)
  - deprivation (5 categories)

Aim: To describe the association between age at diagnosis and the excess mortality hazard

First explanatory model:

$$\lambda_E(t;x) = \lambda_0(t)\exp\left(\sum_{i=2}^{4}\alpha_i stage_i + \sum_{i=2}^{5}\beta_i dep_i + \sum_{i=2}^{5}\gamma_i agecat_i\right)$$

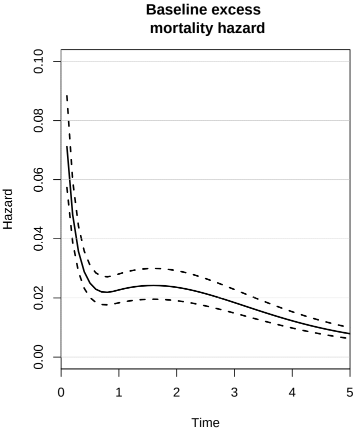
where $\lambda_0(t)$ is the (exponential of a) B-spline of degree 3, with 1 knot located at 1 year

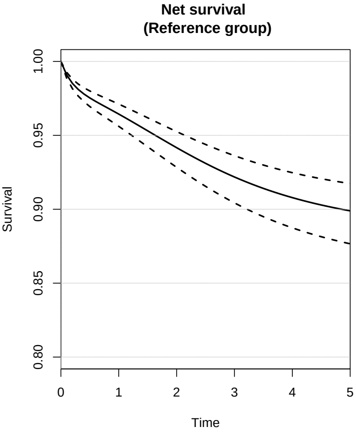
R syntax using the `mexhaz` package

R-code for the first model:

```
mexhaz(Surv(timey, dead) ~ Idep2+Idep3+Idep4+Idep5
                          + IstageB+IstageC+IstageD
                          + Iagegrp1545+Iagegrp4555
                          + Iagegrp5565+Iagegrp75pp,
        data=temp, base= "exp.bs", degree=3, knots=c(1),
        verbose = 500, expected="rate")
```

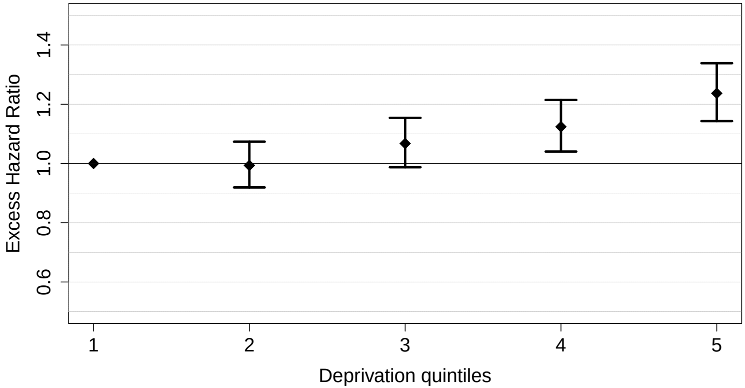
Results using the first model (Age groups)





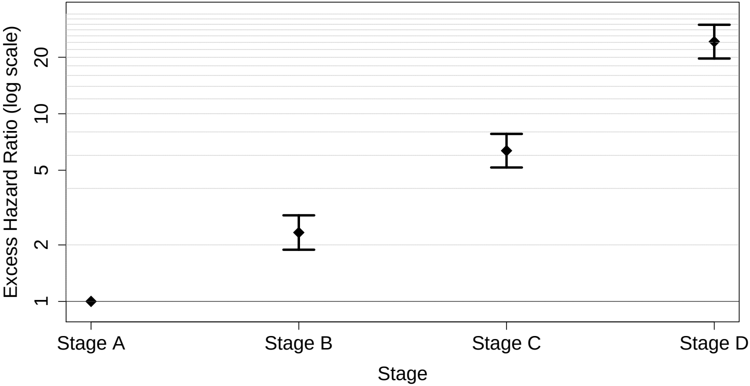
Results using the first model (Age groups)

Effect of deprivation



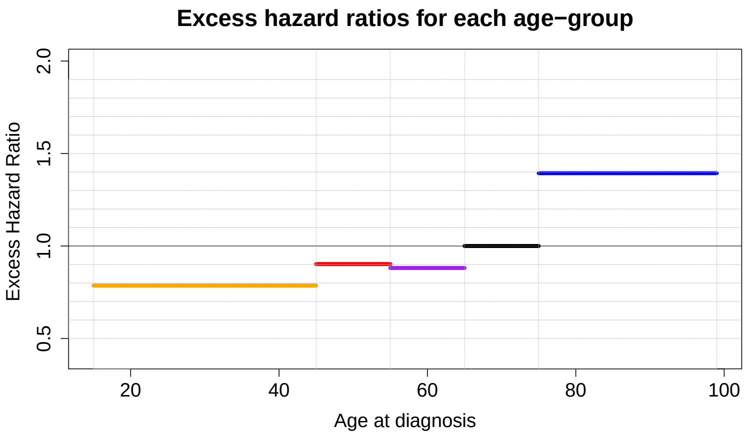
Results using the first model (Age groups)

Effect of stage



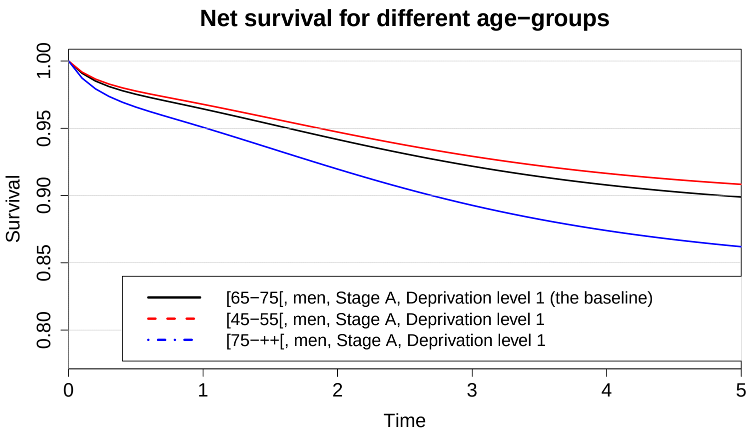
Results using the first model (Age groups)

Effect of age groups



Results using the first model (Age groups)

Net survival by age-group



Refinements

Second model: linear effect of age

$$\lambda_E(t;x) = \lambda_0(t)\exp\Big(\sum_{i=2}^{4}\alpha_i stage_i + \sum_{i=2}^{5}\beta_i dep_i + \theta\, age\Big)$$

where $\lambda_0(t)$ is the (exponential of a) B-spline of degree 3, with 1 knot located at 1 year

R syntax using the `mexhaz` package

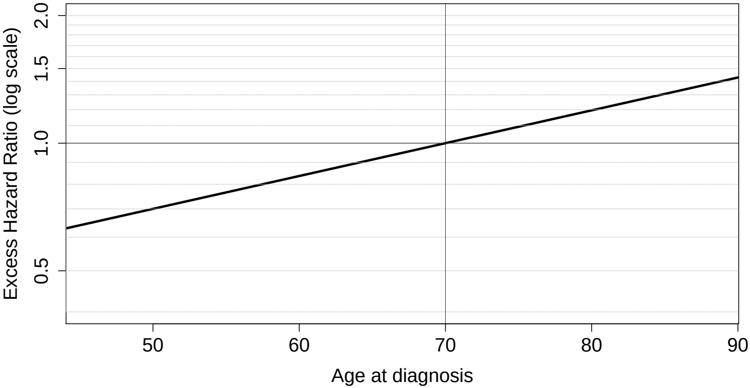
R-code for the second model:

```
mexhaz(Surv(timey, dead) ~ Idep2+Idep3+Idep4+Idep5
                          + IstageB+IstageC+IstageD
                          + agediagc,
       data=temp, base= "exp.bs", degree=3, knots=c(1),
       verbose = 500, expected="rate")
```

The variable `agediagc` was created before, and correspond to `agediag` centered: `agediagc = agediag-70`

Results using the second model (linear effect of age)

For 1-year increase of age, θ-increase of the linear predictor



More refinements

Third model: Non-linear effect of age

$$\lambda_E(t;x) = \lambda_0(t)\exp\Big(\sum_{i=2}^{4}\alpha_i stage_i + \sum_{i=2}^{5}\beta_i dep_i + \beta_a age + f(age)\Big)$$

where $\lambda_0(t)$ is the (exponential of a) B-spline of degree 3, with 1 knot located at 1 year and $f()$ is a flexible function (B-spline, degree 2, 1 knot at age 70 (age centred))

R syntax using the `mexhaz` package

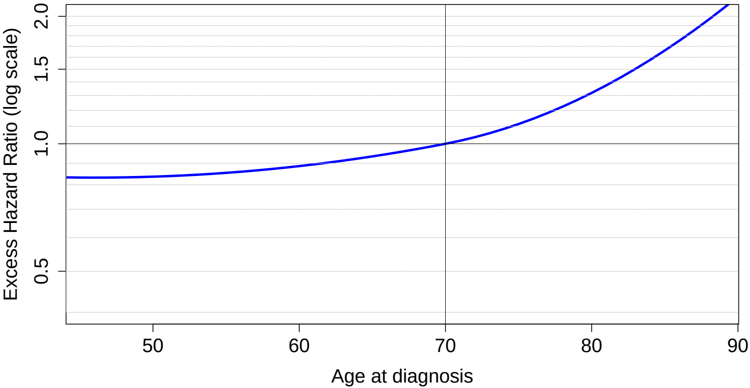
R-code for the third model:

```
mexhaz(Surv(timey, dead) ~ Idep2+Idep3+Idep4+Idep5
                          + IstageB+IstageC+IstageD
                          + agediagc + agediagc2
                          + agediagc2plus ,
       data=temp, base= "exp.bs", degree=3, knots=c(1),
       verbose = 500, expected="rate")
```

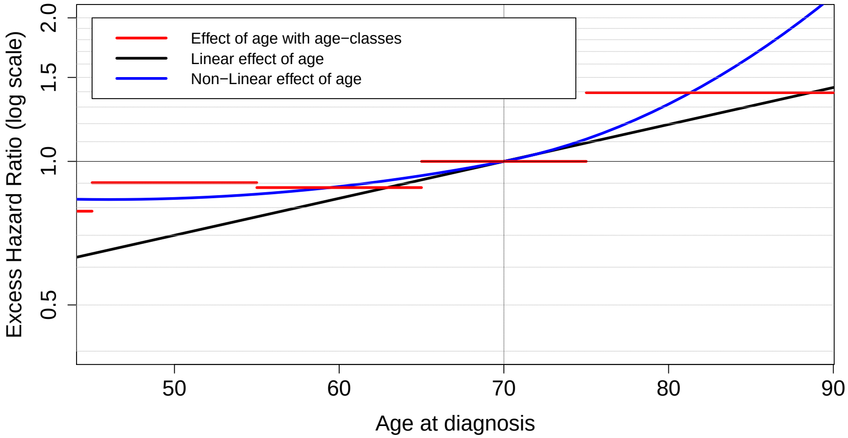
The variables `agediagc2` and `agediagc2plus` were created before, and correspond to $agediagc^2$, and $agediagc_+^2$

Results using the third model (Non-linear effect of age)

For 1-year increase of age, the increase of the linear predictor is different when comparing 45 with 44 years old, than when comparing 84 with 83 years old



Comparison of the 3 first models



More refinements:Time-dependent effect

Fourth model: Non-linear and time-dependent effect of age

$$\lambda_E(t;x) = \lambda_0(t)\exp\Big(\sum_{i=2}^{4}\alpha_i stage_i + \sum_{i=2}^{5}\beta_i dep_i + \beta_a(t)age + f(age)\Big)$$

where $\lambda_0(t)$ is the (exponential of a) B-spline of degree 3, with 1 knot located at 1 year. $f()$ and $\beta_a()$ are flexible functions (B-spline)

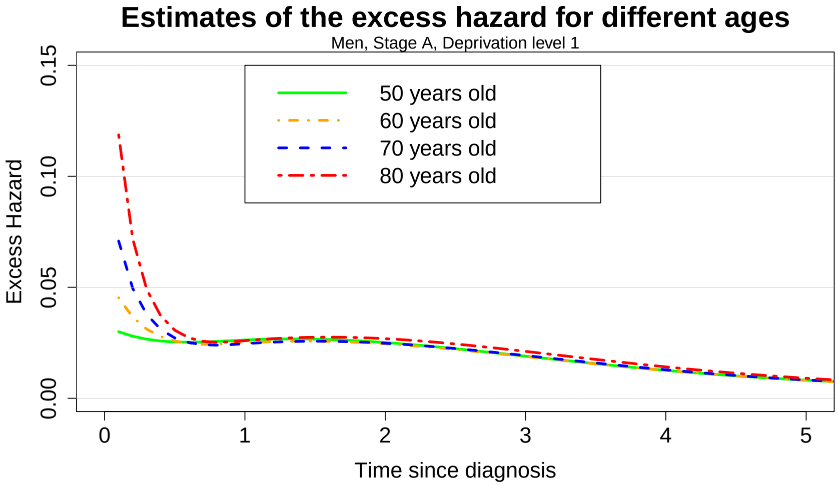
R syntax using the `mexhaz` package

R-code for the fourth model:

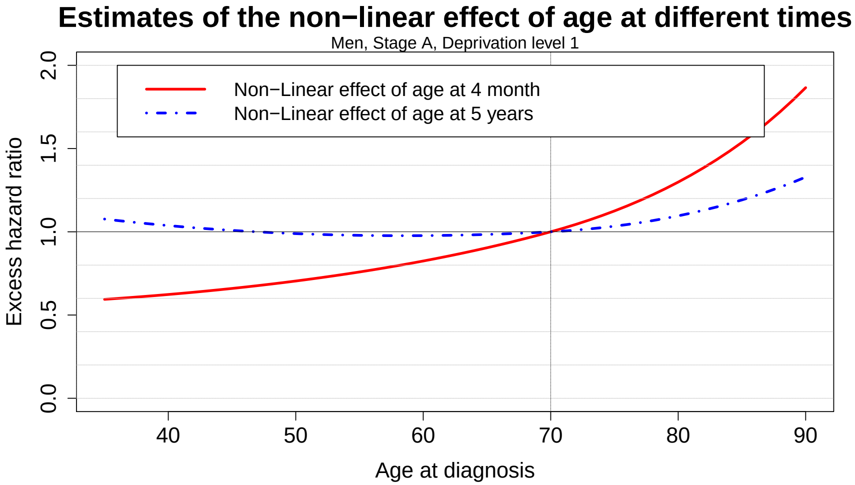
```
mexhaz(Surv(timey, dead) ~ Idep2+Idep3+Idep4+Idep5
                         + IstageB+IstageC+IstageD
                         + agediagc + agediagc2
                         + agediagc2plus + nph(agediagc),
        data=temp, base= "exp.bs", degree=3, knots=c(1),
        verbose = 500, expected="rate")
```

The variables `agediagc2` and `agediagc2plus` were created before, and correspond to $agediagc^2$, and $agediagc_+^2$

Results fourth model



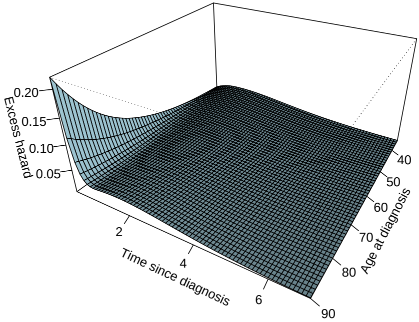
Results fourth model



Results fourth model

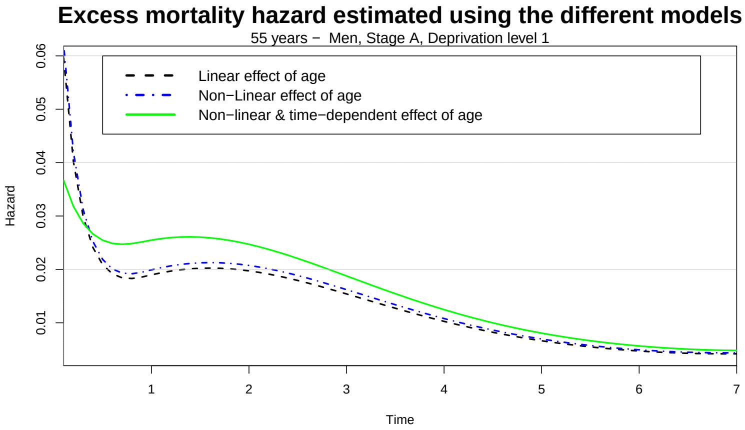
The model: Non-linear and time-dependent effect of age

$$\lambda_E(t;x) = \lambda_0(t)\exp\Big(\sum_{i=2}^{4}\alpha_i stage_i + \sum_{i=2}^{5}\beta_i dep_i + \nu(t)age + f(age)\Big)$$



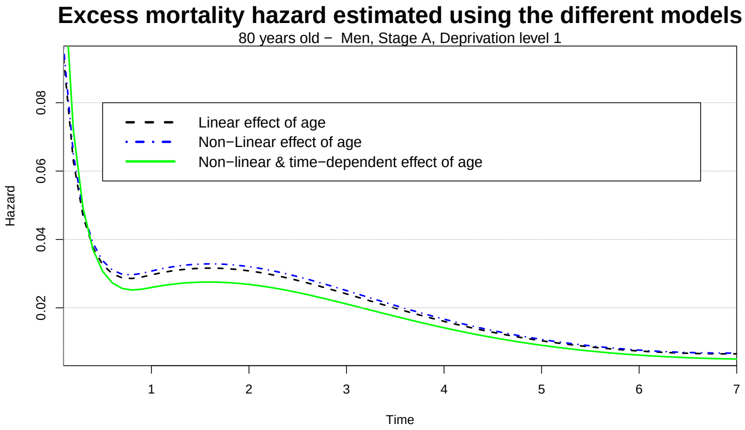
Comparison of the 3 models keeping age continuous

Excess hazard for 55 years old



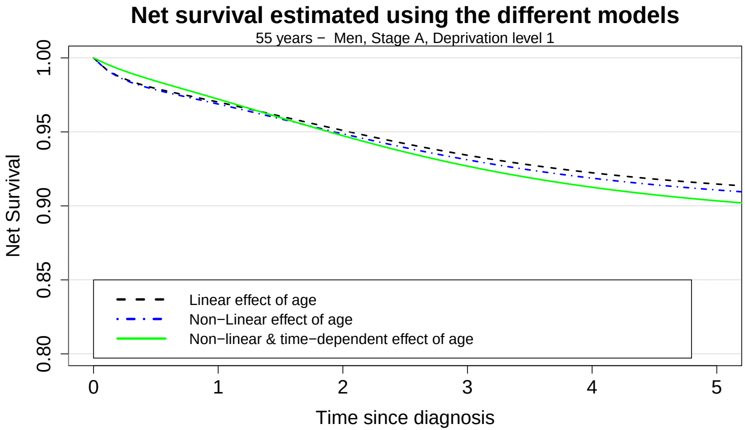
Comparison of the 3 models keeping age continuous

Excess hazard for 80 years old



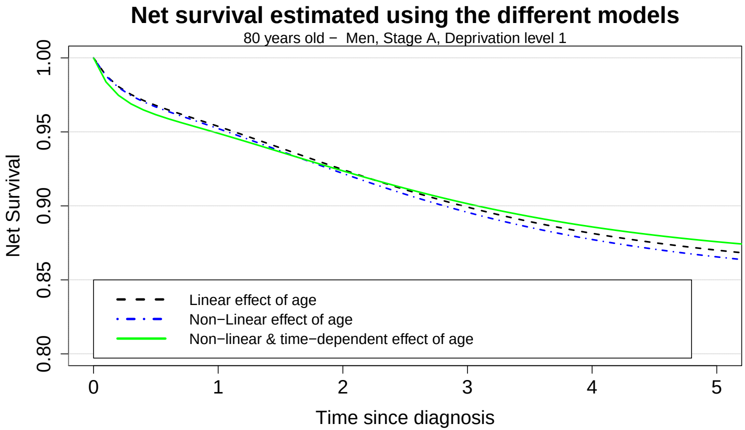
Comparison of the 3 models keeping age continuous

Net survival for 55 years old



Comparison of the 3 models keeping age continuous

Net survival for 80 years old



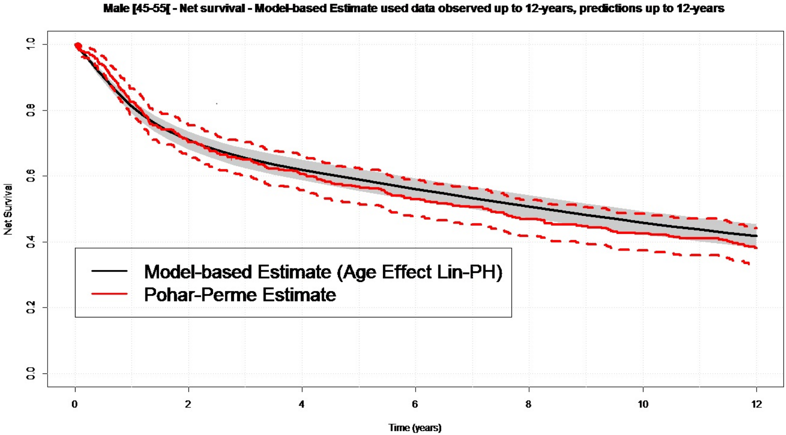
How to choose the “best” model?

In term of best fitting model, we can look at the AKAIKE INFORMATION CRITERION (lower is better)

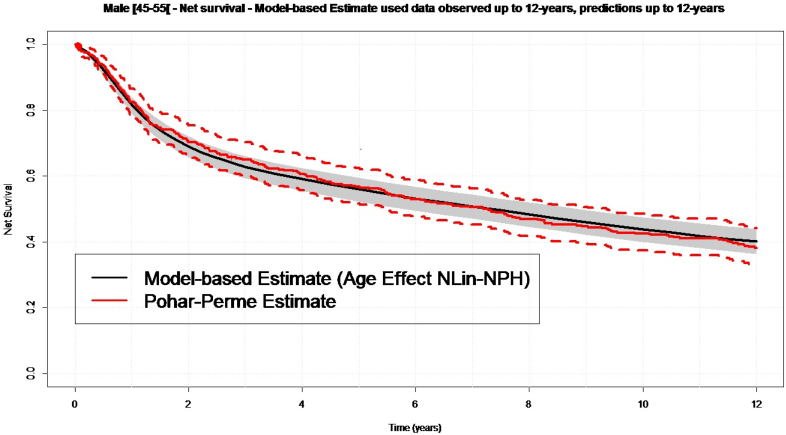
$$AIC = -2 \times LL + 2 \times Nparameters$$

Model with
age with linear effect: 43787.46
age with Non-linear effect: 43710.12
age with Non-linear and Time-dependent effect: 43416.64

The importance of Time-dependent effect



The importance of Time-dependent effect



A full-week short course: Corsican Summer School on Modern Methods in Biostatistics and Epidemiology 2017

- Statistical methods and recent advances in statistical methods for excess risk analysis
- Monday 3rd July to Friday 7th July 2017, in Corte (Corsica, France)

"http://sesstim.univ-amu.fr/hearstat-2017"

References

- Hakulinen T, Tenkanen L. Regression analysis of survival rates. Applied in Statistics 1987; 36: 309-317.
- Esteve J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. Statistics in Medicine 1990; 9(5):529-538.
- Bolard P, Quantin C, Esteve J, Faivre J, Abrahamowicz M. Time dependent hazard ratio in relative survival with application in colon cancer. Journal of Clinical Epidemiology 2001; 54: 986-96.
- Bolard P, Quantin C, Abrahamowicz M, Estve J, Giorgi R, Chadha-Boreham H, Binquet C, Faivre J. Assessing time-by-covariate interactions in relative survival models using restrictive cubic spline functions. Journal of Cancer Epidemiology and Prevention 2002;7(3): 113-22.
- Giorgi R, Abrahamowicz M, Quantin C et al. A relative survival regression model using B-spline functions to model non-proportional hazards. Statistics in Medicine 2003; 22:2767-2784.
- Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. Stat Med 2004;23:51-64.
- Lambert PC, Smith LK, Jones DR, Botha JL. Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. Statistics in Medicine 2005;24:3871-3885.

References

- Giorgi R, Payan J, Gouvernet J. RSurv: a function to perform relative survival analysis with S-PLUS or R. Computer Methods and Programs in Biomedicine 2005; 78: 175-178
- Pohar M, Stare J. Relative survival analysis in R. Computer Methods and Programs in Biomedicine 2006;81(3):272-8
- Pohar M, Stare J. Making relative survival analysis relatively easy. Computers in biology and medicine 2007; 37 : 17411749
- Nelson CP, Lambert PC, Squire IB, Jones DR. Flexible parametric models for relative survival, with application in coronary heart disease. Statistics in Medicine 2007;26(30):5486-98
- Remontet L, Bossard N, Belot A, Estve J, and the French network of cancer registries FRANCIM. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. Statistics in Medicine 2007; 26:2214-28
- Bossard N, Velten M, Remontet L, Belot A, Maarouf N, Bouvier AM, et al. Survival of cancer patients in France: a population-based study from The Association of the French Cancer Registries (FRANCIM). European Journal of Cancer 2007;43(1):149-60

References

- Pohar Perme M, Henderson R, Stare J. An approach to estimation in relative survival regression. Biostatistics 2009; 10:136146
- Mahboubi A, Abrahamowicz M, Giorgi R, et al. Flexible modeling of the effects of continuous prognostic factors in relative survival. Statistics in Medicine 2011;30(12):1351-65.
- 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 Epidemiology. 2013;37(4):449-56
- Mounier M, Bossard N, Belot A, Remontet L, Iwaz J, Dandoit M, Girard-Boulanger S, Herry A, Woronoff AS, Casasnovas RO, Maynadi M, Giorgi R; FRANCIM Network and MESURE Working Survival Group. Trends in excess mortality in follicular lymphoma at a population level. European Journal of Haematology. 2015; 94(2):120-9. doi: 10.1111/ejh.12403
- Remontet L, Bossard N, Iwaz J, Estve J, Framework and optimisation procedure for flexible parametric survival models. Statistics in medicine. 2015; 34(25):3376-7. doi: 10.1002/sim.6489
- 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. Statistics in medicine. 2016;35: 3066-3084. doi: 10.1002/sim.6881

References

- Stare J, Pohar M, Henderson R. Goodness of fit of relative survival models. Statistics in Medicine 2005; 24(24):3911-3925
- Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Statistics in Medicine 2007; 26:55125528
- Cortese G, Scheike TH. Dynamic regression hazards models for relative survival. Statistics in Medicine 2008; 27(18):3563-3584
- Royston P, Sauerbrei W. Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Wiley series in probability and statistics. John Wiley and Sons, 2008
- Binder H, Sauerbrei W, Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. Statistics in Medicine 2013; 32:22622277
- Wynant W, Abrahamowicz M. Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. Stat Med. 2014 Aug 30; 33(19):3318-37