Modelling excess mortality using fractional polynomials and spline functions

Bernard Rachet

Modelling time-dependent excess hazard

Excess hazard function $\lambda_c(t)$

$$\lambda_c(t) = \lambda_0(t)\exp(\beta(t)X(t))$$

with

- $\lambda_0(t)$ time-dependent baseline excess hazard
- $\boldsymbol{\beta}(t)$ possibly time-dependent vector of coefficients
- $\boldsymbol{X}(t)$ possibly time-dependent vector of co-variables

Fractional polynomials

A fractional polynomial of degree *m* consists of *m* integer/fractional powers

$$fp(x) = a_0 + \sum_{j=1}^{m} a_j x^{(p_j)}$$

where the power (p_j) is generally chosen from a restricted set including

{-2; -1; -0.5; 0; 0.5; 1; 2; 3}

with $$x^{(0)} = \log x$$

Fractional polynomials

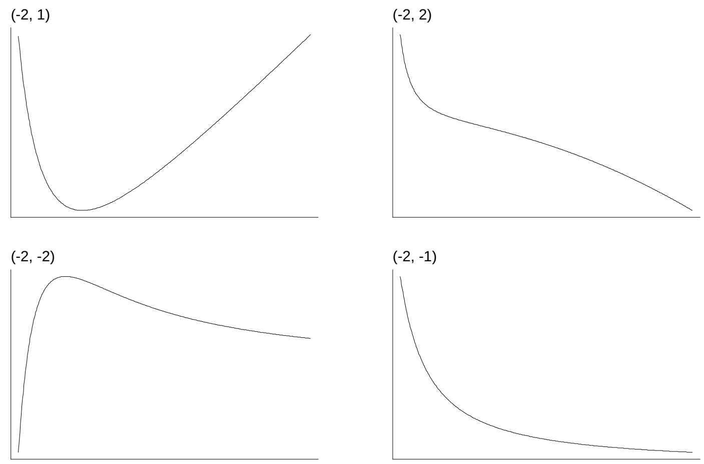
A fractional polynomial of first degree (*m* = 1) can include repeated powers of *p*

$$fp(x) = a_0 + a_1 x^{(p1)} + \sum_{j=2}^{m} a_j x^{(p_j)} (\log x)^{m-1}$$

Example of *m* = 2 with repeated powers of 0.5

$$a_0 + a_1 x^{0.5} + a_2 x^{0.5} \log x$$

Possible curve shapes with second-degree fractional polynomials



From Royston *et al.*, 1999

Algorithm to choose the most suitable model

Sauerbrei W, Royston P. Journal of the Royal Statistical Society, Series A 1999; 162:71–94.

Implemented in Stata® command *mfp.ado* and adapted to *mvrs.ado*

Principles of the algorithm

1. Determine fitting order of co-variables from linear model
2. For each co-variable *X*:
   - Fit model with each combination of powers
   - Choose model with lowest deviance
   - Compare FP*m* with FP(*m*-1)
   - Functions for other *X* are fixed, but not their β's
3. Iterate step 2 until powers of FP functions do not change

Fractional polynomials

Excess hazard function $\lambda_c(t)$

$$\lambda_c(t) = fp_0(t)\exp(\beta(t)X(t))$$

with

$\boldsymbol{\beta}(t)$ a vector of fractional polynomials of time

Example: patients diagnosed with colon cancer

Fit a multiplicative regression model for relative survival

grouped data

small intervals

fractional polynomial for baseline excess hazard

```
. strs using popmort, br(0(.1)10) mergeby(_year sex _age) ///
    by(sex year8594 agegrp) notables ...

. use colon_grouped, clear
. gen midtime=(start+end)/2

. xi: mfp glm d midtime (i.sex i.year8594 i.agegrp), ///
    adjust(no) df(6) alpha(-1) ///
    xorder(n) family(poisson) link(rs d_star) lnoffset(y)
```

Stata/SE 9.2 - [Results]

File Edit Prefs Data Graphics Statistics User Window Help

```
Deviance for model with all terms untransformed =  4568.354, 1600 observations

Variable     Model (vs.)    Deviance  Dev diff.   P      Powers   (vs.)

midtime      lin.           4568.354                     1
             FP1            4502.797                     0
             FP2            4464.240                     -.5 3
             FP3            4438.248                     -2 -2 3
             Final          4438.248                     -2 -2 3

[_Isex_2 _Iyear8594_1 _Iagegrp_1 _Iagegrp_2 _Iagegrp_3 included with 1 df in model]


Fractional polynomial fitting algorithm converged after 1 cycle.

Transformations of covariates:

-> gen double Imidt__1 = midtime^-2 if e(sample)
-> gen double Imidt__2 = midtime^-2*ln(midtime) if e(sample)
-> gen double Imidt__3 = midtime^3 if e(sample)

Final multivariable fractional polynomial model for d

    Variable  |    -----Initial-----              -----Final-----
              |    df     Select    Alpha    Status    df     Powers
--------------+------------------------------------------------------
      midtime |     6     1.0000    A.I.C.     in       6     -2 -2 3
   _Isex_2 ...|     1     1.0000    A.I.C.     in       1     1
---------------------------------------------------------------------

Generalized linear models                        No. of obs      =      1600
```

Command

D:\data

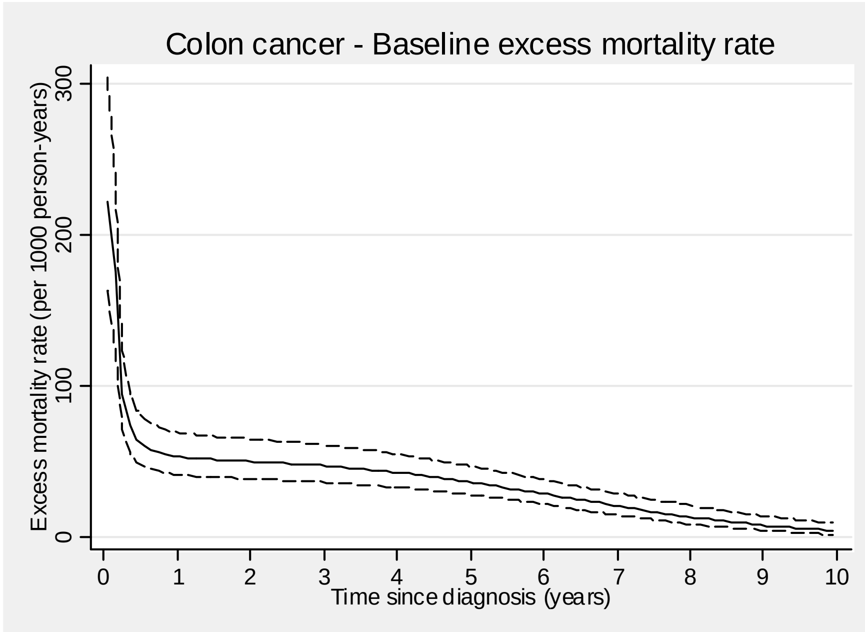
Stata/SE 9.2 - [Results]

File Edit Prefs Data Graphics Statistics User Window Help

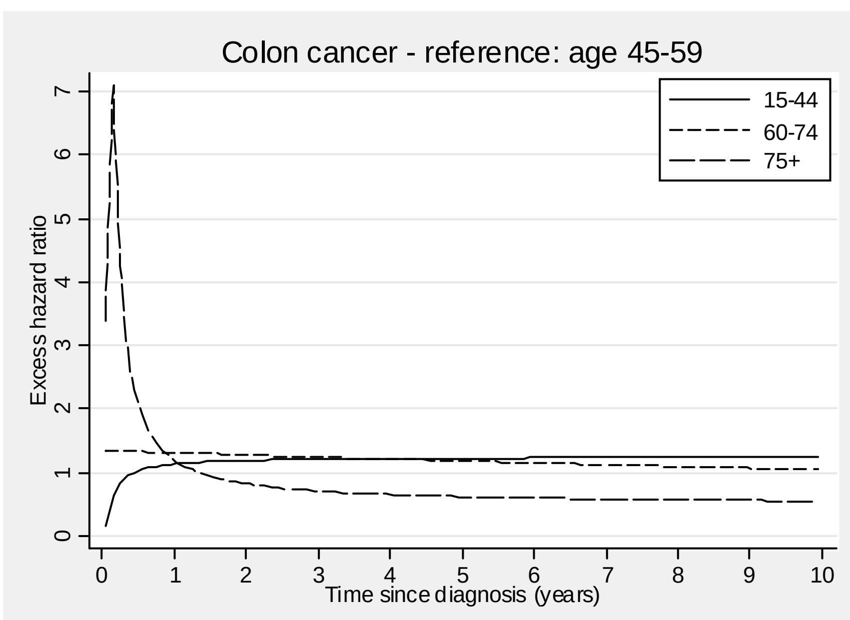
```
Generalized linear models                        No. of obs      =      1600
Optimization     : ML                            Residual df     =      1591
                                                 Scale parameter =         1
Deviance         =  1932.888375                  (1/df) Deviance =  1.214889
Pearson          =  2241.143026                  (1/df) Pearson  =  1.408638

Variance function: V(u) = u                      [Poisson]
Link function    : g(u) = log(u-d*)              [Relative survival]

                                                 AIC             =  2.785155
Log likelihood   = -2219.124179                  BIC             = -9805.126

------------------------------------------------------------------------------
             |                 OIM
           d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Imidt__1 |   .0707225   .0074299     9.52   0.000     .0561602    .0852848
    Imidt__2 |   .0223596   .0024654     9.07   0.000     .0175276    .0271917
    Imidt__3 |  -.0026192   .0004411    -5.94   0.000    -.0034838   -.0017545
     _Isex_2 |  -.0707154   .0707442    -1.00   0.318    -.2093715    .0679407
_Iyear8594_1 |  -.2630362   .0691814    -3.80   0.000    -.3986292   -.1274431
  _Iagegrp_1 |  -.1077499   .1456763    -0.74   0.460    -.3932703    .1777705
  _Iagegrp_2 |   .0980065   .1344159     0.73   0.466    -.1654438    .3614568
  _Iagegrp_3 |   .4203599   .1390874     3.02   0.003     .1477536    .6929661
       _cons |  -3.003356   .1348498   -22.27   0.000    -3.267657   -2.739056
           y | (exposure)
------------------------------------------------------------------------------
Deviance: 4438.248.
```

Command

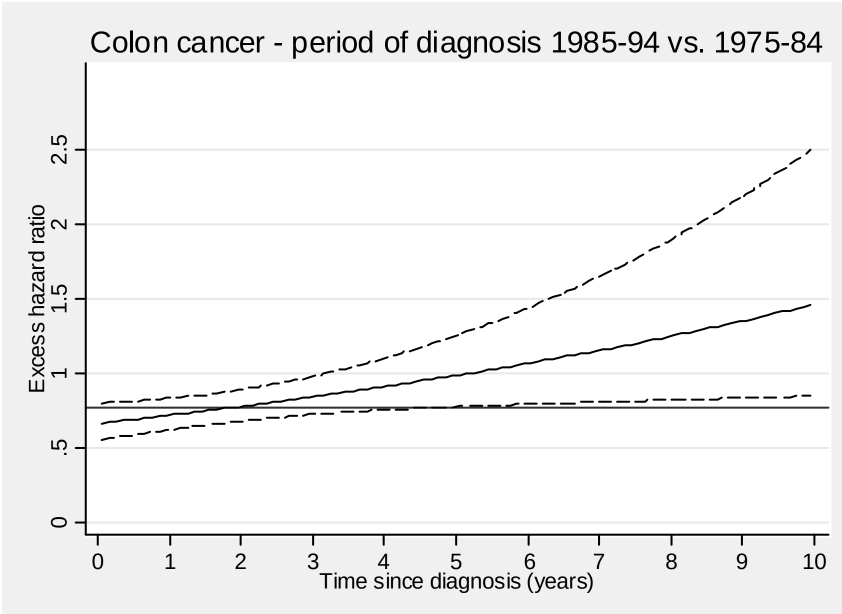
D:\data



Colon cancer - Baseline excess mortality rate
Excess mortality rate (per 1000 person-years)
0
100
200
300
0
1
2
3
4
5
6
7
8
9
10
Time since diagnosis (years)



Colon cancer - reference: age 45-59
Excess hazard ratio
0
1
2
3
4
5
6
7
15-44
60-74
75+
0
1
2
3
4
5
6
7
8
9
10
Time since diagnosis (years)



Splines

Excess hazard function $\lambda_c(t)$

$$\lambda_c(t) = \exp\big(f(t) + \beta(t)X(t)\big)$$

with

$f(t)$ spline function and $\exp(f(t)) = \lambda_0(t)$

$\boldsymbol{\beta}(t)$ a vector of spline functions

Main broad classes

- regression or natural splines (incl. restricted, B-splines, …)
- smoothing splines
- *penalised splines*

Regression or natural splines

Linear combination of piecewise polynomial functions

+ linear combination of truncated functions

A k-degree polynomial function consists of $a_0 + a_1 t^1 + \ldots + a_k t^k$

Interval-specific polynomials joined smoothly at distinct knots (continuity of the function and the first two derivatives at the knots)

$$f(t) = a_0 + a_1 t^1 + \ldots + a_k t^k + \sum_{j=1}^{m} \theta_j \left(t - t_j\right)_+^k$$

where $u_+ = \max(0, u)$ and $\{t_j\}_m$ are m knots

Regression splines

Flexibility

Degree of the polynomial

cubic, quadratic …

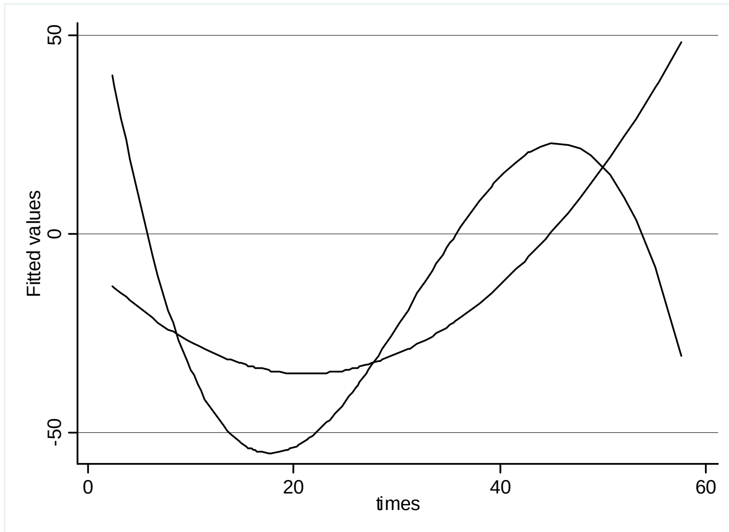
Knots

sensitivity to their location and their number

Degrees of freedom

number of internal knots + 1

Regression splines



How flexible?

Baseline excess hazard

cubic splines with 2 knots or 3-degree fractional polynomial

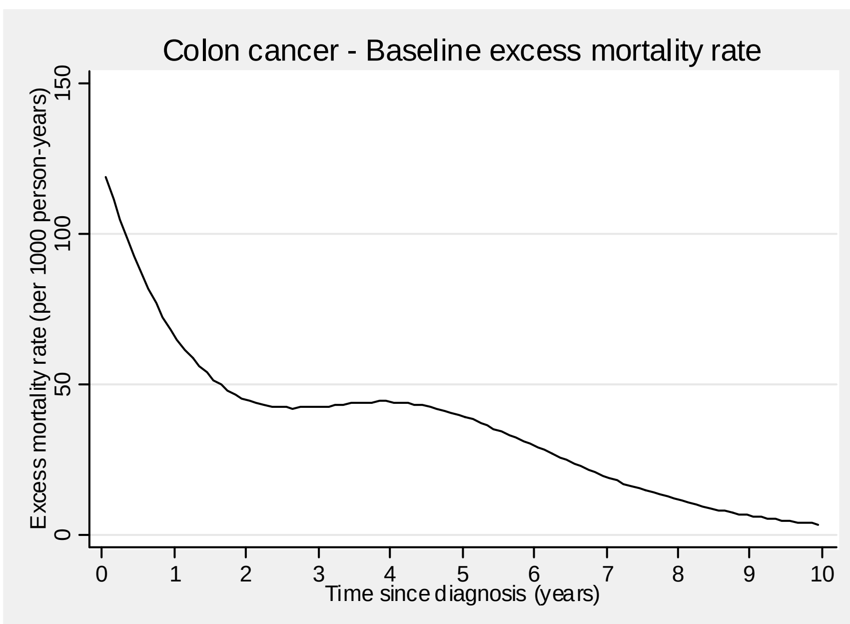
Time-dependent effects – categorical co-variable

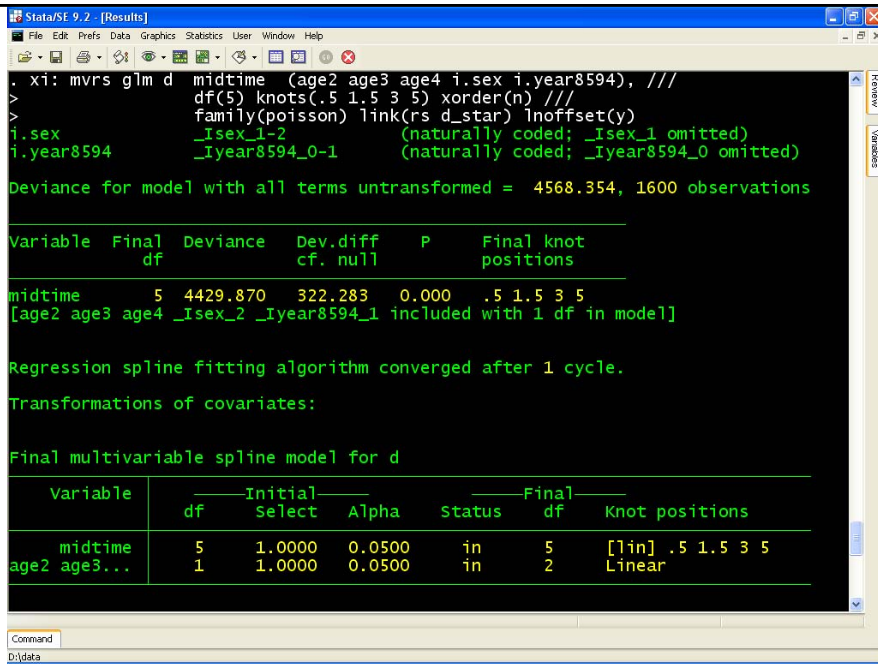
cubic splines with 1 knot or 2-degree fractional polynomial for each interaction term “dummy co-variable by time”

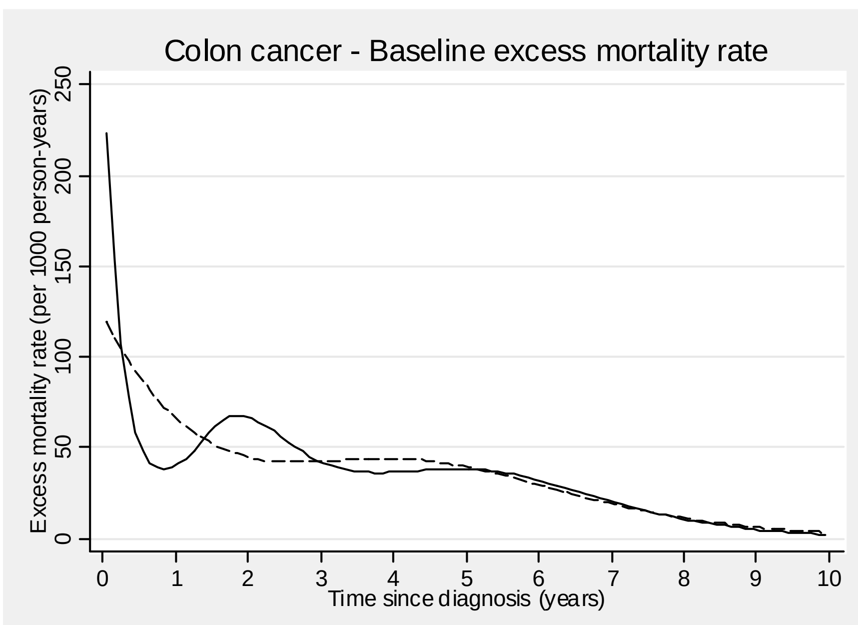
Time-dependent effects – continuous co-variable

cubic splines with 1 knot or 2-degree fractional polynomial for both co-variable (knot at the mean) and interaction “co-variable by time”

```
. xi: mvrs glm d  midtime  (age2 age3 age4 i.sex i.year8594), ///
>                  df(5) xorder(n) family(poisson) link(rs d_star) lnoffset(y)
i.sex            _Isex_1-2           (naturally coded; _Isex_1 omitted)
i.year8594       _Iyear8594_0-1      (naturally coded; _Iyear8594_0 omitted)

Deviance for model with all terms untransformed =  4568.354, 1600 observations

------------------------------------------------------------
Variable  Final  Deviance   Dev.diff    P      Final knot
            df              cf. null           positions
------------------------------------------------------------
midtime       3  4540.946   211.206   0.000   1.95 3.95
[age2 age3 age4 _Isex_2 _Iyear8594_1 included with 1 df in model]


Regression spline fitting algorithm converged after 1 cycle.

Transformations of covariates:


Final multivariable spline model for d
------------------------------------------------------------------------------
    Variable |    -----Initial-----          -----Final-----
             |   df     Select   Alpha     Status    df    Knot positions
-------------+----------------------------------------------------------------
     midtime |    5     1.0000   0.0500      in       3    [lin] 1.95 3.95
age2 age3... |    1     1.0000   0.0500      in       2    Linear
------------------------------------------------------------------------------
```


Stata/SE 9.2 - [Results]
File Edit Prefs Data Graphics Statistics User Window Help
. xi: mvrs glm d midtime (age2 age3 age4 i.sex i.year8594), ///
> df(5) knots(.5 1.5 3 5) xorder(n) ///
> family(poisson) link(rs d_star) lnoffset(y)
i.sex _Isex_1-2 (naturally coded; _Isex_1 omitted)
i.year8594 _Iyear8594_0-1 (naturally coded; _Iyear8594_0 omitted)
Deviance for model with all terms untransformed = 4568.354, 1600 observations
Variable Final df Deviance Dev.diff cf. null P Final knot positions
midtime 5 4429.870 322.283 0.000 .5 1.5 3 5
[age2 age3 age4 _Isex_2 _Iyear8594_1 included with 1 df in model]
Regression spline fitting algorithm converged after 1 cycle.
Transformations of covariates:
Final multivariable spline model for d
Variable Initial df Select Alpha Final Status df Knot positions
midtime 5 1.0000 0.0500 in 5 [lin] .5 1.5 3 5
age2 age3... 1 1.0000 0.0500 in 2 Linear
Command
D:\data


Colon cancer - Baseline excess mortality rate
Excess mortality rate (per 1000 person-years)
0
50
100
150
200
250
0
1
2
3
4
5
6
7
8
9
10
Time since diagnosis (years)

```
Survival time in months (vital status 1 or 2)
-------------------------------------------------------------
      Percentiles      Smallest
 1%           .5             .5
 5%           .5             .5
10%          2.5             .5       Obs                3291
25%         13.5             .5       Sum of Wgt.        3291

50%         34.5                      Mean           48.95943
                        Largest       Std. Dev.      46.81865
75%         71.5          239.5
90%        116.5          239.5       Variance       2191.986
95%        147.5          241.5       Skewness       1.361728
99%        202.5          241.5       Kurtosis       4.669213
```



```
. xi: mvrs glm d  midtime  (age2 age3 age4 i.sex i.year8594), ///
     df(5) xorder(n) family(poisson) link(rs d_star) lnoffset(y)


Deviance for model with all terms untransformed = 38521.869,
 312192 observations

------------------------------------------------------------
Variable  Final  Deviance   Dev.diff    P     Final knot
            df              cf. null          positions
------------------------------------------------------------
convergence not achieved
```

To conclude…

- Continuous functions provide clinically relevant patterns of excess hazard
- No inflation of the number of parameters
- Sparse data when problematic partitioning
- Flexibility *vs.* smoothness

Fractional polynomials

- Stata *mfp.ado* (P Royston) using generalised linear model for relative survival (P Dickman)

Splines

- Stata *mvrs.ado* (P Royston) using generalised linear model for relative survival (P Dickman)
- R/S Rsurv (R Giorgi)
  Rsurv-v2 (L Remontet)

References

- Royston P and Sauerbrei W. Multivariable modeling with cubic regression splines: A principled approach. *The Stata Journal* 2007; **7**:45–70.

- Royston P and Gambler G. sg81: Multivariable fractional polynomials. *Stata Technical Bulletin Reprints* 1998; **8**:123-132.

- 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**:3871-3885.

- Remontet L, Bossard N, Belot A, Estève J and the French network of cancer registries FRANCIM2. 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**:2214–2228.

- Giorgi R, Abrahamowicz M, Quantin C, Bolard P, Estève J, Gouvernet J, Faivre J. A relative survival regression model using B-spline functions to model non-proportional hazards. Stat Med 2003; 22:2767-2784.

- Bollard P, Quantin C, Estève J, Abrahamowicz M. Modelling time-dependent hazard ratios in relative survival: application to colon cancer. J Clin Epidemiol 2001; 54:986-996.

- Bollard P, Quantin C, Abrahamowicz M, Estève 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 Can Epidemiol Prev* 2002; **7**:113-122.

- Harrell FEJ. *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer-Verlag: New York, 2002; 19–23.