Relative survival: Understanding the concepts and using the methods

Slovenia, October 2008

Motivation

- Is breast cancer survival worse among young women?
- Do women have better prognosis than men in colorectal cancer?
- Is it possible to say that a patient with colon cancer is cured, if he survived five years after the diagnosis?
- *The usual evaluation of survival is unable to answer these questions. The relative survival methodology will greatly help.*

I.Understanding the concepts

Net survival & Relative survival

Net survival probability

A patient with cancer is submitted to the mortality hazard due to this specific disease added to that due to other pathologies, as observed in the population to which he belongs. In other words,

At time t after diagnosis, the hazard rate of a patient, diagnosed at age x, is the sum of the cancer hazard rate and of the hazard rate due to other causes

$$\nu_o(t) = \nu_c(t) + \mu_a(x+t)$$

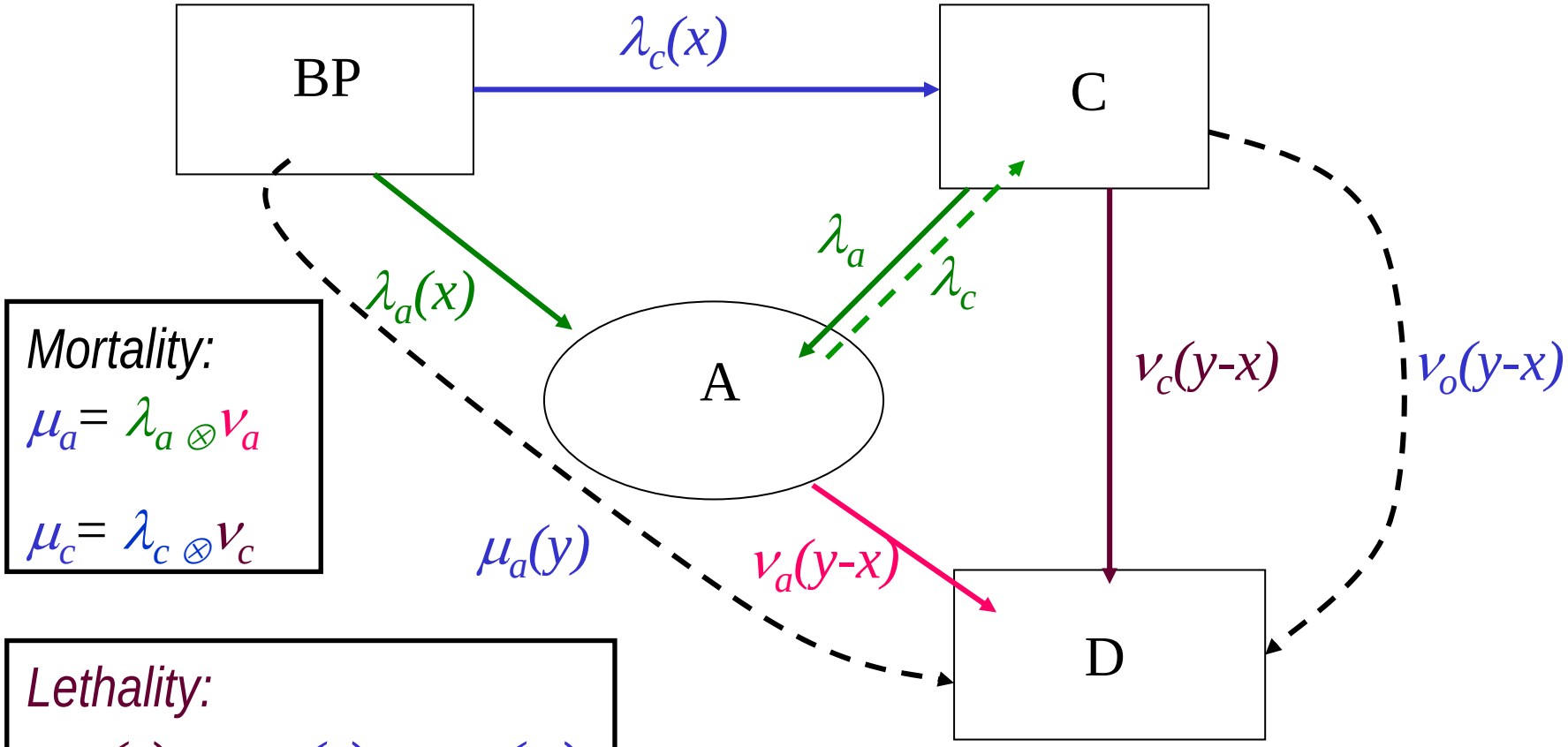
$$S_o(t) = \exp\left(-\int_0^t (\nu_c(u) + \mu_a(x+u))du\right)$$

$$= S_c(t)S_a(t)$$

Therefore the *net cancer survival probability* is obtained from the *crude survival* probability divided by the probability of surviving from other causes *between age x and $x+t$.*

A simple(?) diagram

x:age at diagnosis ***y***:age at death ***t*=*y*-*x***: time since diagnosis



Estimation of the net survival

- If incidence and mortality for a given cancer are known in a given population, the associated survival might be computed (at least in principle). This is the *net survival for this given population*.
- The net survival may be *estimated* from survival data of a given cohort, if the cause of death of each deceased subject is known:
  - An observation is censored at the time of death if the death is not caused by the cancer under study.
- This latter estimation may be biased due to the subjective determination of the cause of death

Relative survival

- If we considered that the mortality from other causes is known and taken equal to the life table mortality, we shall estimate the excess cancer death rate:

$$\nu_r(t) = \nu_o(t) - \mu_{lt}(x+t)$$

- Or equivalently, as explained above, we shall estimate *the relative survival*

$$S_r(t) = S_o(t) / S_{lt}(x, x+t) = S_o(t) / S_e(t)$$

Where $S_e(t)$ is known as the *expected survival*

- The *relative survival* impute to cancer the deaths from causes that are indirectly attributable to the disease, its treatment or to its risk factors.

Relative survival estimation (I)
the “ratio estimate”

- Initially the estimation methods relied on the simple calculation of the expected survival from *ad-hoc* life table.
  - The relative survival was then calculated by dividing the “actuarial” or the “KM” survival estimate by this expected survival
- The methods differed only in the way the expected survival was computed (Ederer *et al*, Hakulinen)

Relative survival estimation (II)

- The life table *expected survival* depends on a set of covariates z , - usually age, sex and year of diagnosis

$$S_r(t) = \frac{S_o(t)}{S_e(t)} = \frac{\sum_i S_e(t, z(i)) S_r(t, z(i))}{\sum_i S_e(t, z(i))}$$

$$= \frac{\sum_z n_z S_e(t, z) S_r(t, z)}{\sum_z n_z S_e(t, z)}$$

Since z is usually a categorical variable

- Therefore, if the relative survival depends effectively on z, the relative survival of the cohort is a weighted average of the relative survival of the various z-categories, the weight being the *number of expected survivors* in the category
- ……

Relative survival estimation (II…)

- ……
- As a consequence the relative survival of the group is *closer to the relative survival of the sub-group with the greater number of expected survivors*
  - Usually the younger, the women, and the most recently diagnosed subjects
  - As time since diagnosis increases the "ratio estimate" may increase since more weight is given to greater relative survival
- The *ratio estimate* will not provide a "*bona fide*" survival curve
- ……

Relative survival estimation (II…)

- ……
- The usual calculation of the *relative survival* of the group

$$S_r(t) = \frac{\sum_z n_z S_r(t,z)}{\sum_z n_z}$$

is different from the "ratio estimate" of the relative survival of the group

- This leads to obvious problems of consistency, in particular for standardization.

- Only when S_r does not depend on z is the *ratio-estimate* clearly defined.

Relative survival estimation (III)
The excess death rate model

- The idea is to estimate the *death rate in excess of the background mortality rate* and to obtain the relative survival from it, instead of estimating the *survival probability* directly:

$$\nu_o(t) = \nu_r(t,\tau) + \mu_{lt}(x+t,z)$$

Where $\nu_r(t,\tau)$ is a parametric function to be estimated from survival data by the maximum likelihood method.

- A natural choice for ν_r is the step function:

$$\nu_r(t,\tau) = \sum_{k=1}^{K} \tau_k I_k(t)$$

where the follow-up time has been split in K intervals.

Illustration I (Estève *et al*)
Colon cancer in Geneva 70-79. Men&women

Causes of death were known and reviewed by a registry physician.
The swiss life table, 1978-83, were used for relative survival computation

		life table				censoring	
		Maximum Likelihood		Ederer		Net survival	
Age	No	5 years	10 years	5 years	10 years	5 years	10 years
<65	322	0.51 (0.03)	0.51 (0.04)	0.54 (0.03)	0.50 (0.04)	0.57 (0.03)	0.55 (0.03)
65-74	292	0.35 (0.03)	0.30 (0.04)	0.37 (0.03)	0.30 (0.05)	0.38 (0.03)	0.31 (0.04)
75+	326	0.24 (0.03)	0.19 (0.03)	0.32 (0.03)	0.30 (0.09)	0.25 (0.03)	0.21 (0.03)
Total	940	0.40 (0.02)	0.36 (0.03)	0.42 (0.02)	0.40 (0.03)	0.40 (0.02)	0.36 (0.02)

Colon cancer in Geneva 70-79. Men&women

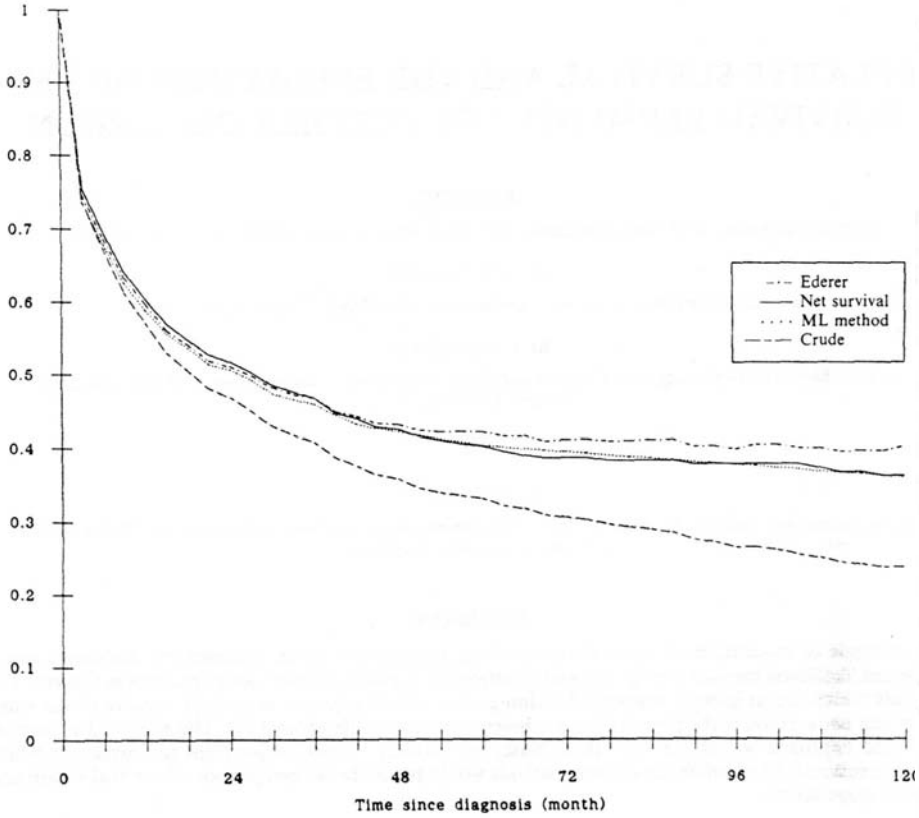


Figure 1. Survival of colon cancer patients, both sexes, Geneva 1970–79

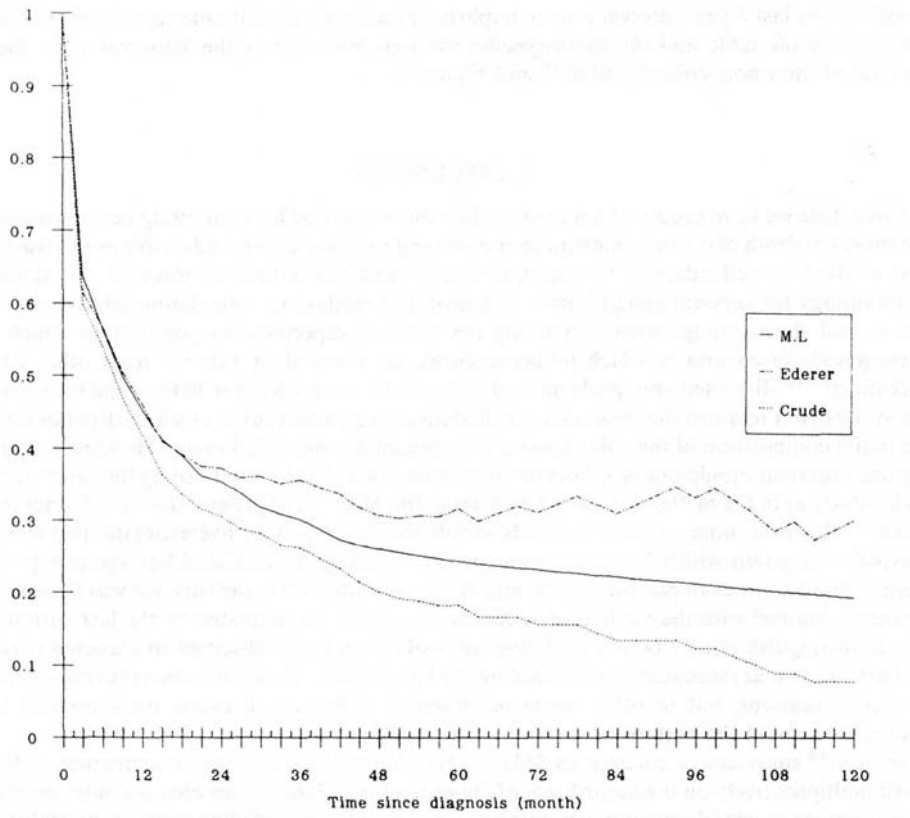


Figure 2. Survival of colon cancer patients, both sexes, age 75 + years, Geneva 1970–79

Illustration II

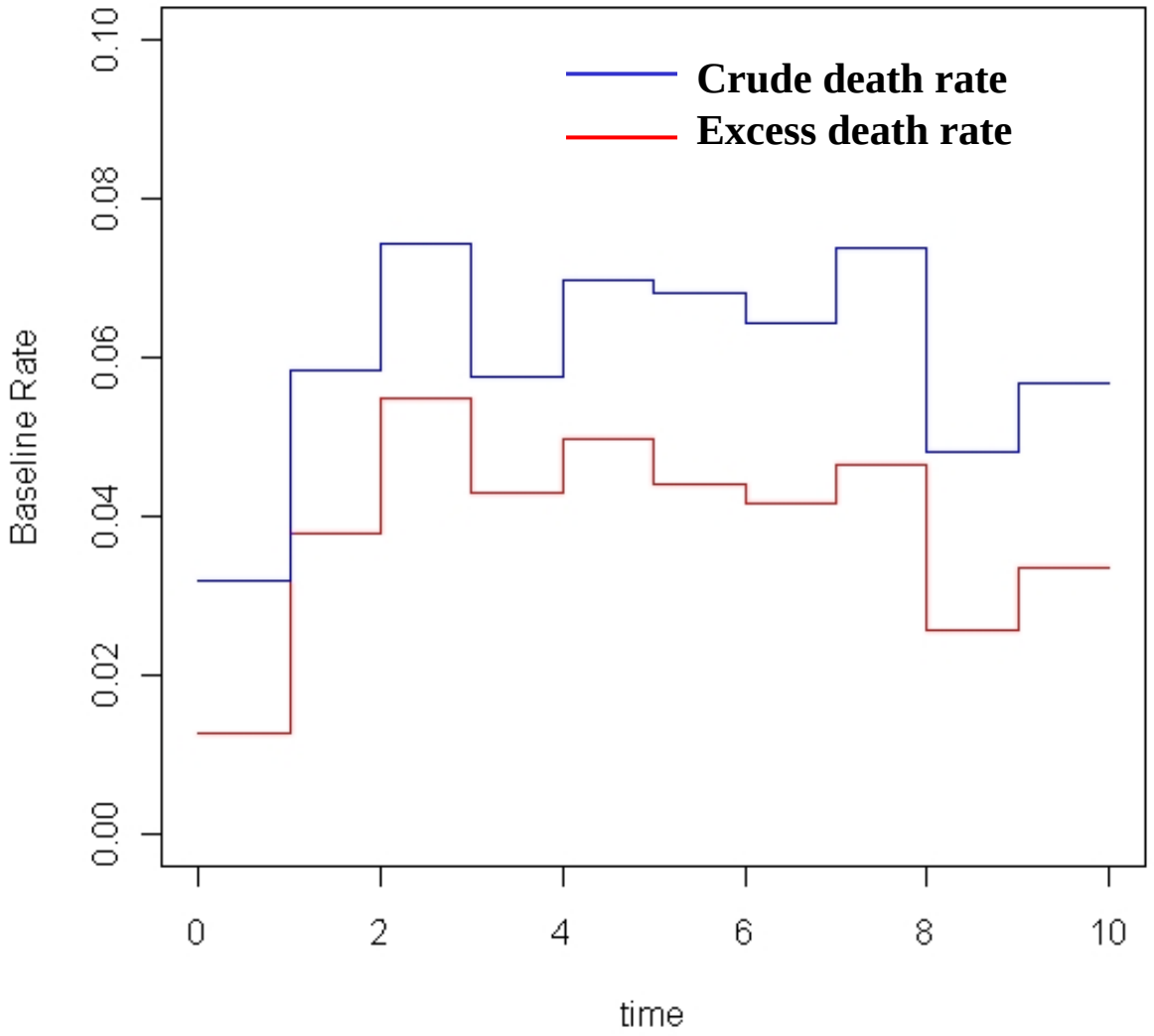
- The Survival data from female breast cancer incident cases in the Eindhoven region of Netherland, between 1984 and 1989

- The available variables are age, year of diagnosis, size of the tumour, number of positive node, number of node examined and also stage, last digit of ICD code, morphology.

- This dataset will serve to illustrate the remainder of the presentation

Breast cancer data: death rate

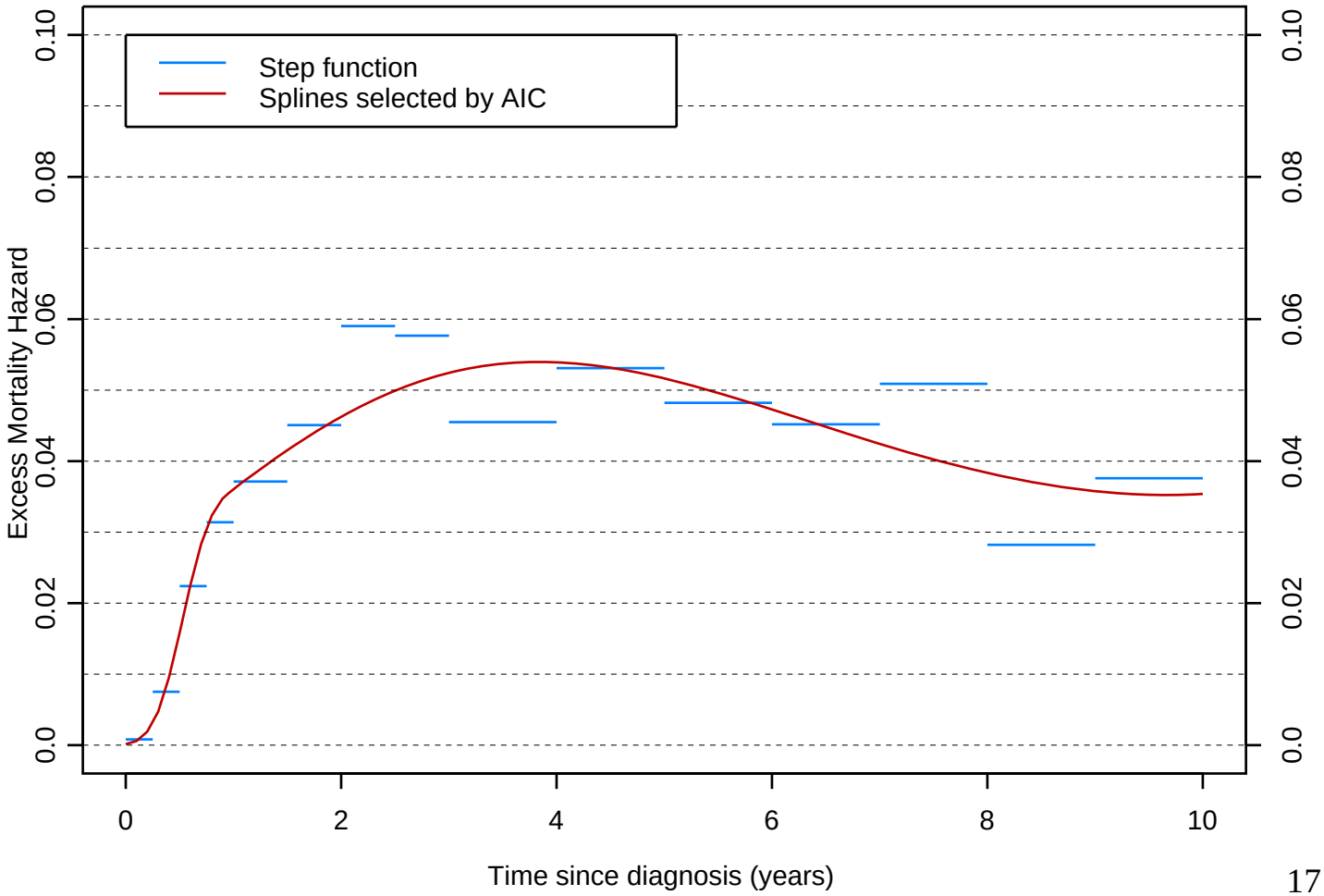
Using a step function, with K=10 and one-year interval, for the excess rate and an exponential regression using the same interval for the crude rate.



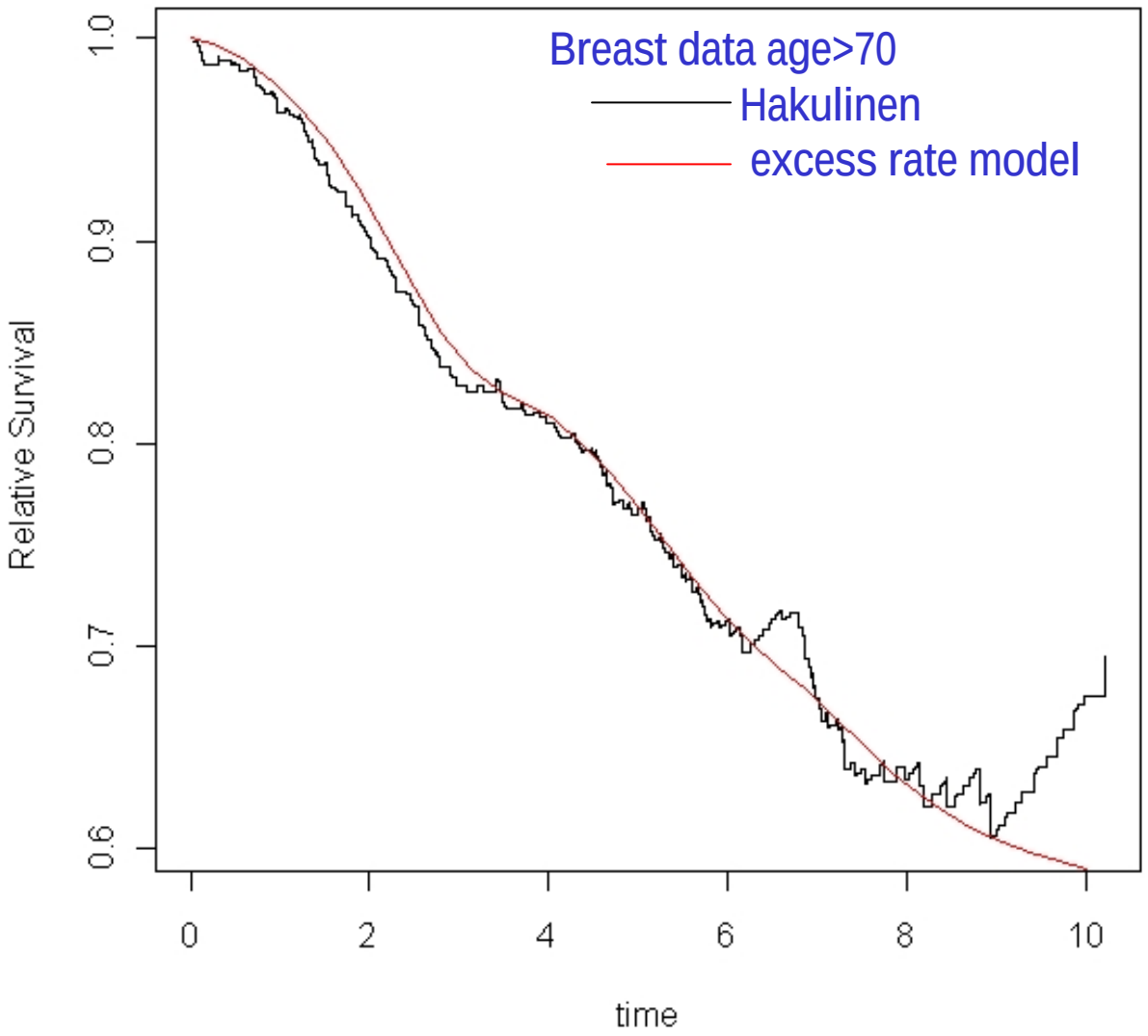
Breast cancer data: an other parametric model

Remontet *et al*

Excess death rate



Relative survival by Hakulinen method



Final notes on the excess death rate estimate

- If the excess death rate is $\nu_r(t,\tau) = \sum_{k=1}^{K} \tau_k I_k(t)$

The ML equations for τ_k are

$$\sum_{t_i \in I_k} \frac{1}{\tau_k + \mu_i} \delta_i - \sum_{i=1}^{n} t_{ki} = 0 \ , \ k = 1 \ldots K \ ,$$

t_{ki} : *time spent by subject i in interval* I_k

If $\tau_k > 0$, these equations are equivalent to:

$$\sum_{t_i \in I_k} \frac{\tau_k}{\tau_k + \mu_i} \delta_i - \tau_k \sum_{i=1}^{n} t_{ki} = 0 \ , \ k = 1 \ldots K$$

- *The two ways of obtaining the # expected deaths, occurring in interval k and caused by the specific disease, must give the same result*

Final notes on the excess death rate estimate

- The previous equation suggest the algorithm

$$\tau_k^{(n+1)} = \left[\sum_{\text{deaths in } I_k} \frac{\tau_k^{(n)}}{\tau_k^{(n)} + \mu_i} \right] \div \left[\sum_{i=1}^{n} t_{ki} \right]$$

- If we increase the number of interval in order to get only one death in each interval the above algorithm becomes:

$$\tau_i^{(n+1)} = \left[\frac{\tau_i^{(n)}}{\tau_i^{(n)} + \mu_i} \right] \div \left[\sum_{j=1}^{n} Y_j(t_i) \right]$$

where the denominator is the number still at risk at time t_i

- This is the *EM algorithm* proposed by *Maja Pohar* when there is no covariate in the additive model

II. Multivariate Regression

For relative survival data

Multivariate Regression for the excess rate

- The extension of the excess death rate model to a multivariate model is straightforward: The function ν_r of slide 12 is now written:

$$\nu_r(\mathbf{t}, \mathbf{z} \mid \tau, \beta) = \nu_b(\mathbf{t} \mid \tau) exp(\beta \mathbf{z})$$

- Where $\nu_b(t, \tau)$ is a parametric function modelling the baseline rate (usually a step function, a regression spline, or a fractional polynomial as in Lambert *et al*)

- The full ML method is used for its estimation, since there is no simplification in using the partial likelihood

A model for grouped data (Hakulinen *et al*)

- Such a model was proposed by Hakulinen et al. in 1987 . Grouping data by follow-up interval permits a GLM to be implemented for parameter estimation

$$Log[-Log[s_{rk}(z)]] = Log\left[-Log\left[\frac{s_{ok}(z)}{s_{ek}(z)}\right]\right] = \gamma_k + \beta z$$

where s_{0k} (s_{ek}) are the crude (expected) survival in the interval *k*

- There is a closed relationship with the model based on individual data

$$-Log[s_{rk}(0)] = -Log\left[\frac{s_{ok}(0)}{s_{ek}(0)}\right] = e^{\gamma_k} = \int_{t_{k-1}}^{t_k} \nu_r(s,0)ds \cong \nu_r(t_{k-1},0) * \Delta t$$

- By definition grouped data *can only deal with categorical explanatory variable*

Computation procedure

- As suggested initially by *Dickman et al.* The estimation of the parameter of the model with *a step function as baseline* is greatly facilitated by the splitting of the individual data at the boundaries of the interval $\boldsymbol{I_k}$
- In the case of the step function, one can take advantage of the proportionality of the “survival likelihood” and of the “Poisson likelihood” to use a GLM algorithm with Poisson error.
- In the case of *other parametric function*, it makes the integration of the baseline rate easier, as shown by *Remontet et al.*
- *Pohar et al.* have implemented several of these approaches in an R-package freely available, where many more is available…

Non-parametric approaches

- Sasieni (1992) developed a non-parametric approach based on the counting process $d\tilde{N}_i(t)=dN_i(t)-Y_i(t)\mu_i(x+t)dt$
  - He provided convergence proofs and examples, but to my knowledge no accessible software exist to illustrate this approach
- Pohar *et al* (2008) developed an approach based on the missing data EM approach. This seems to be based on the counting process $dN^*_i(t)=w_i dN_i(t)$ where $w_i=\nu_{ri}/(\nu_{ri}+\mu_i)$, *the expectation of* $dN_i(t)$ if only death from the specific disease are counted.
  - The method is implemented in the relsurv R-package
- "Full" additive model has been used by Zahl and discussed recently by Cortese *et al*

Which method ?

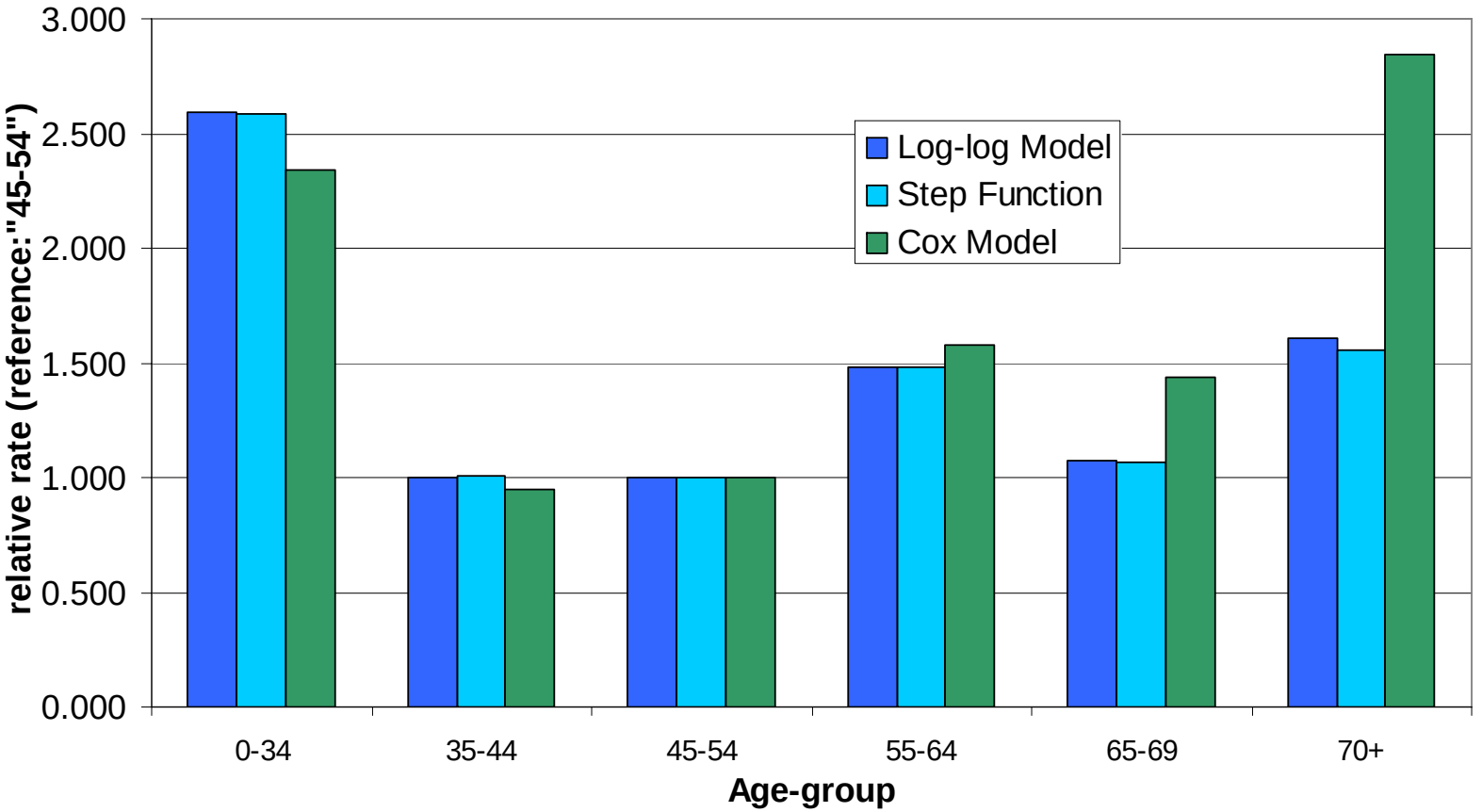
- We have still little experience on the comparison of the various methods but we can anticipate that they will give *close results for the β estimates*.
- The estimate of the *baseline rate* may be crucial in some context and more work is needed in that respect.
  - Is *a posteriori* smoothing better than *a priori* modelling?
- All methods allows *time dependant covariates* and the checking and modelling of *non proportionality*
- The context of the study is certainly important to decide which method to use, in particular for deciding between additive, multiplicative and other models (e.g.: models for individual measure of *Stare et al*)

III-Illustration of the methods

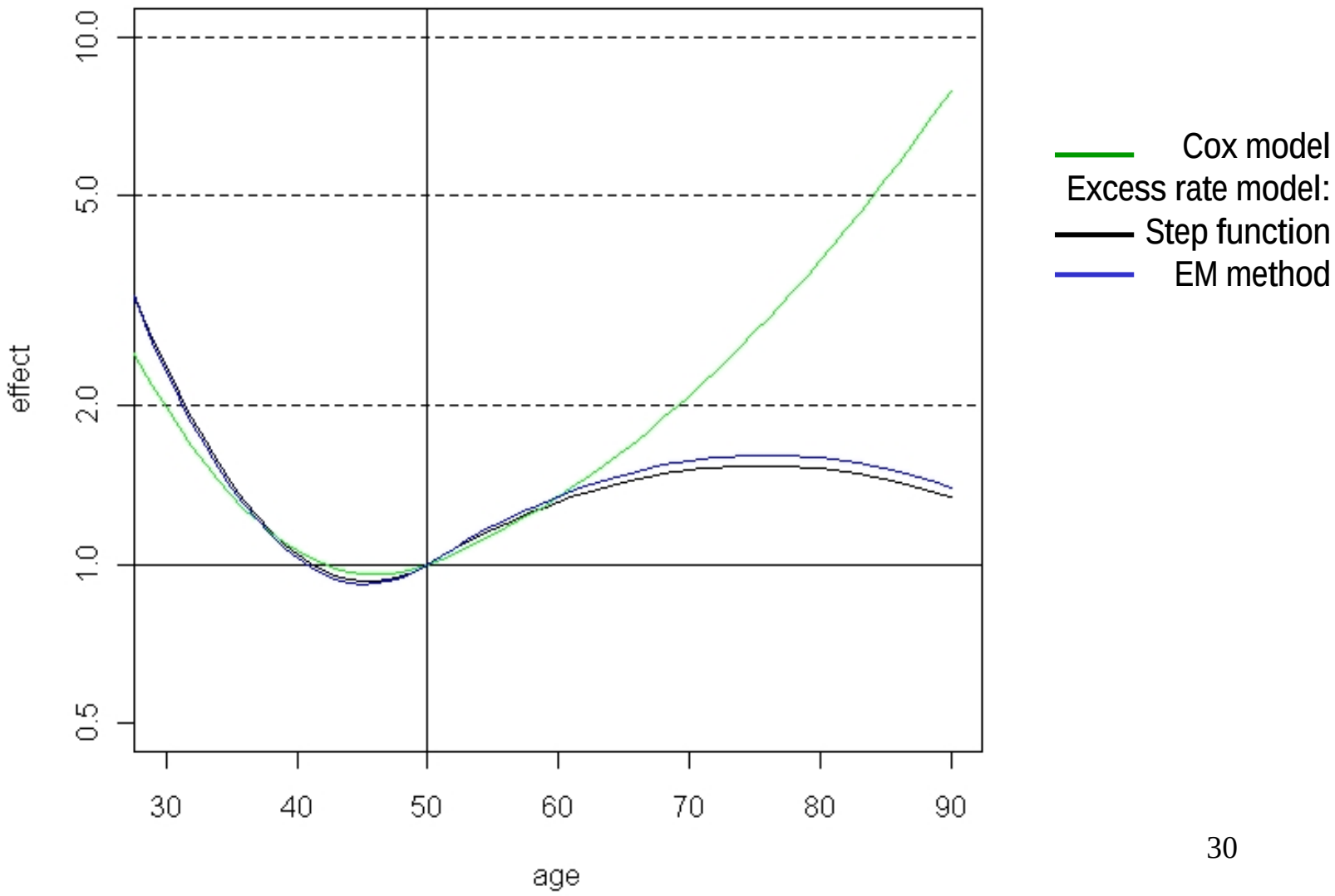
With the breast cancer data

Is breast cancer survival worse among young women?

Age effect on crude and relative survival, age as a categorical variable

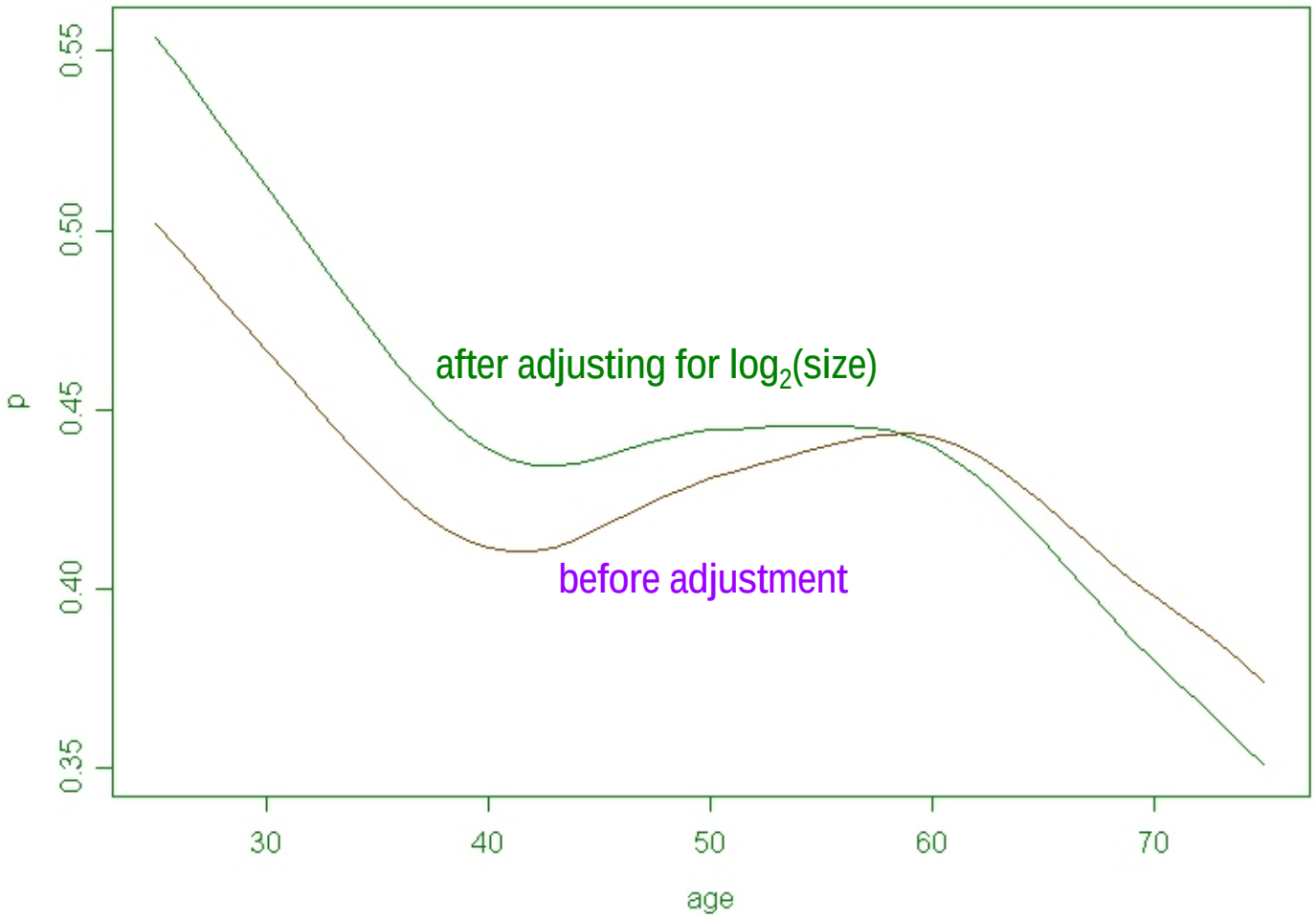


Age effect on crude and relative survival, age as a quadratic spline function (one node at age 50)

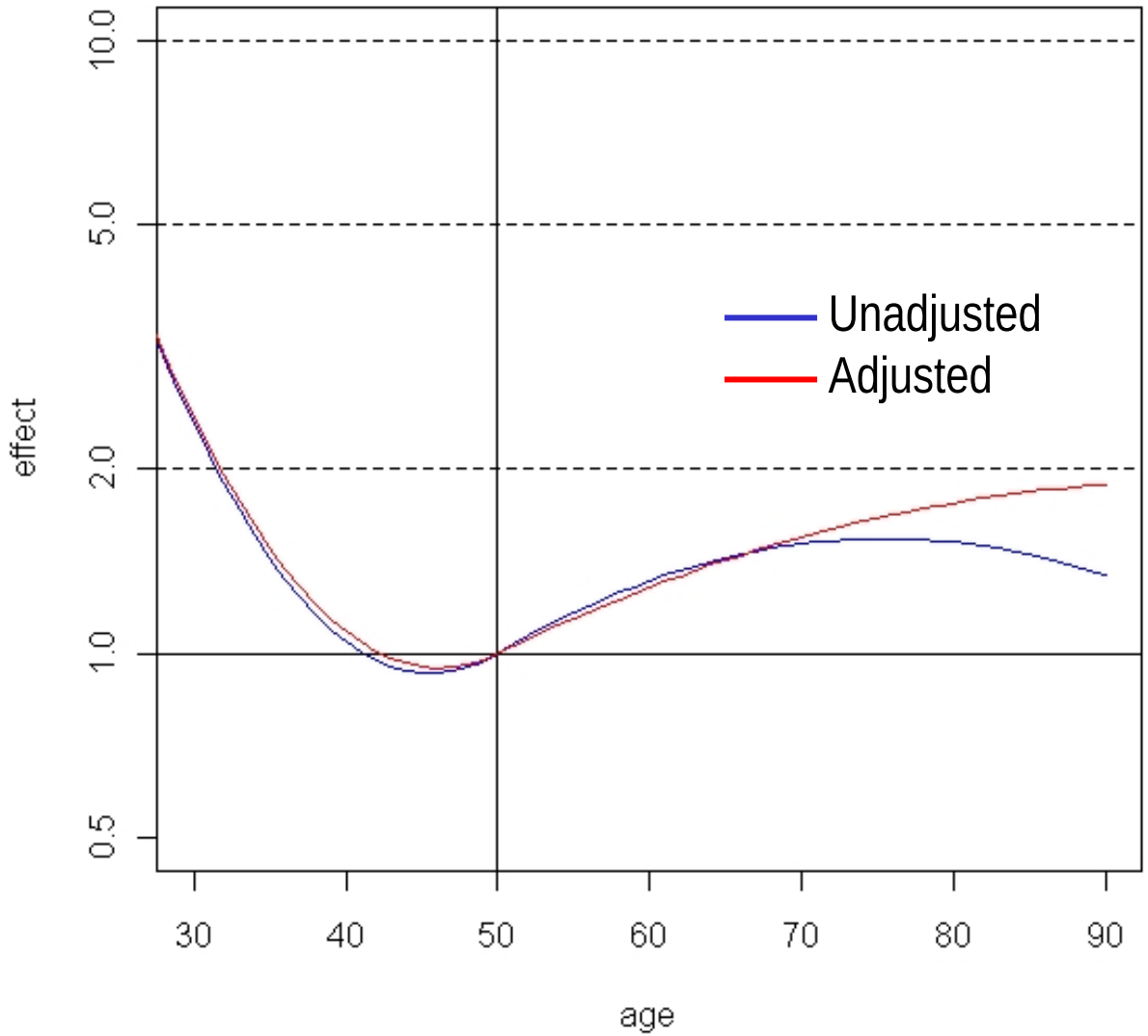


How changes the age effect if we adjust for the size of the tumour and the number of positive node?

Probability of a positive node

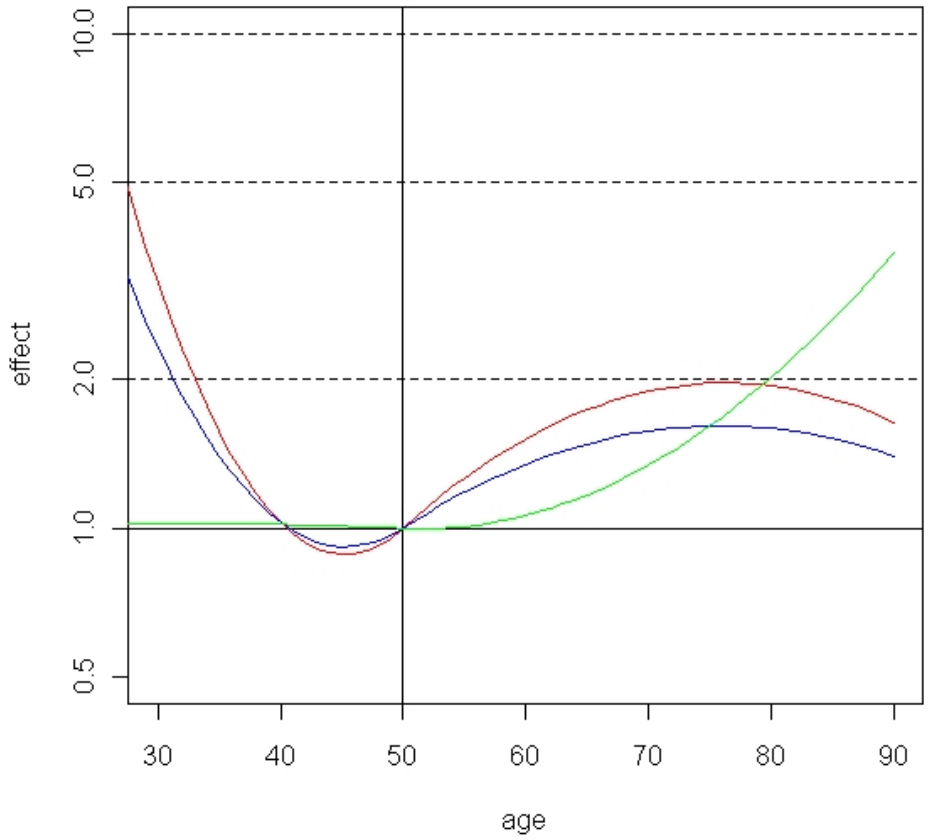


Age effect after adjustment for size and node

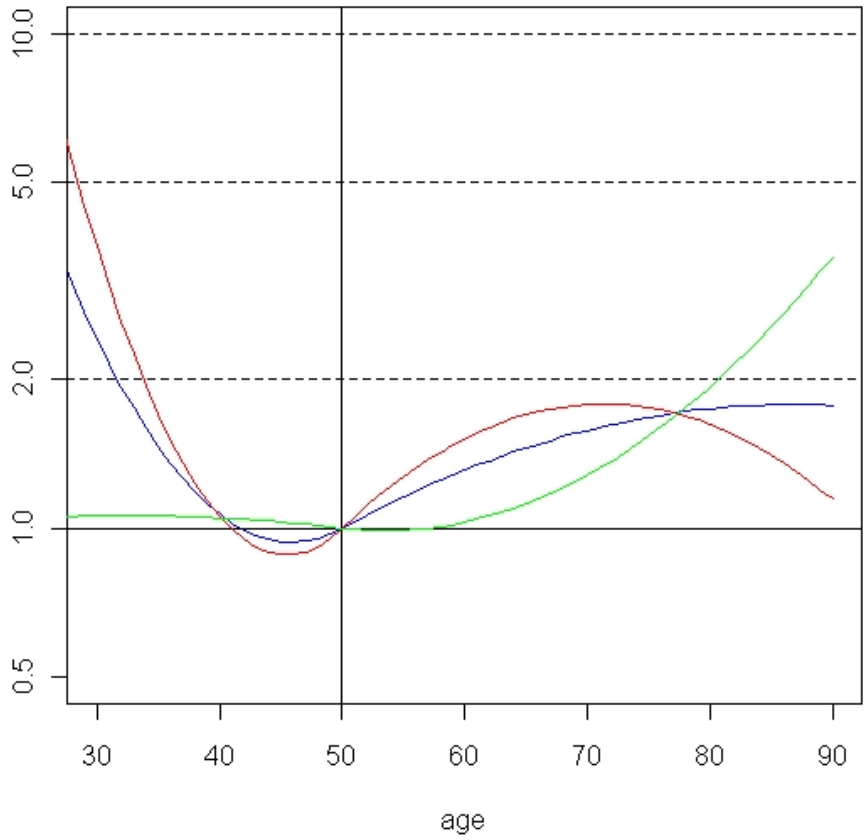


Age effect by node status

Whole cohort Node + Node – (not significant)



No adjustment



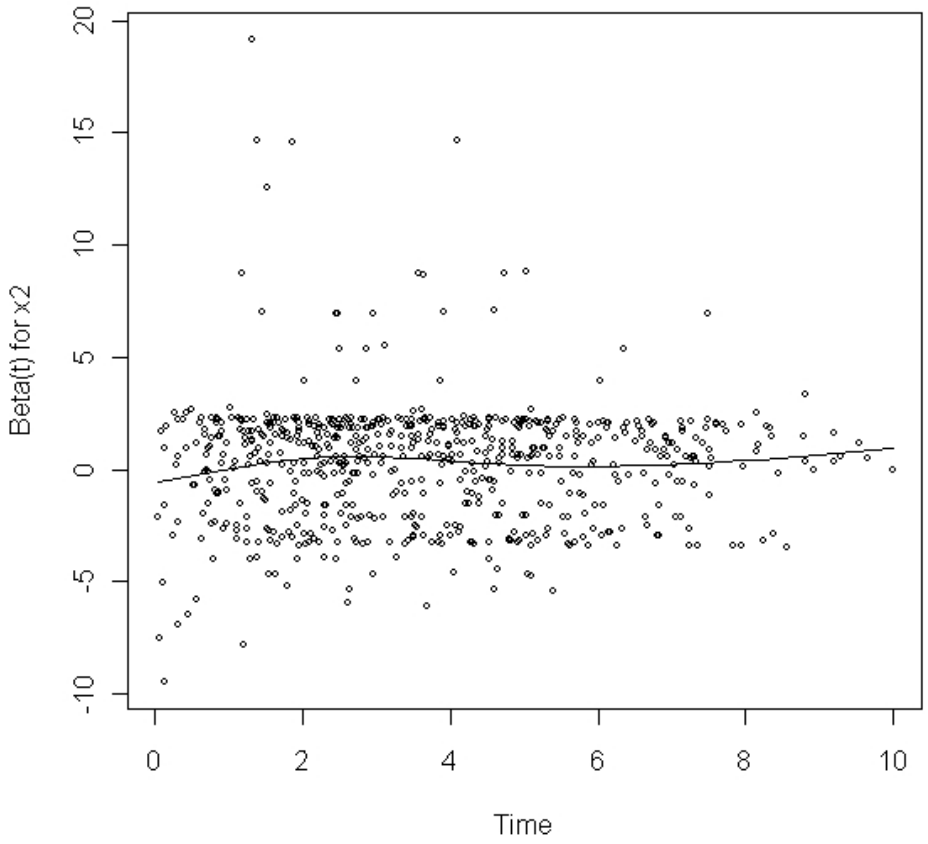
Adjusted for size

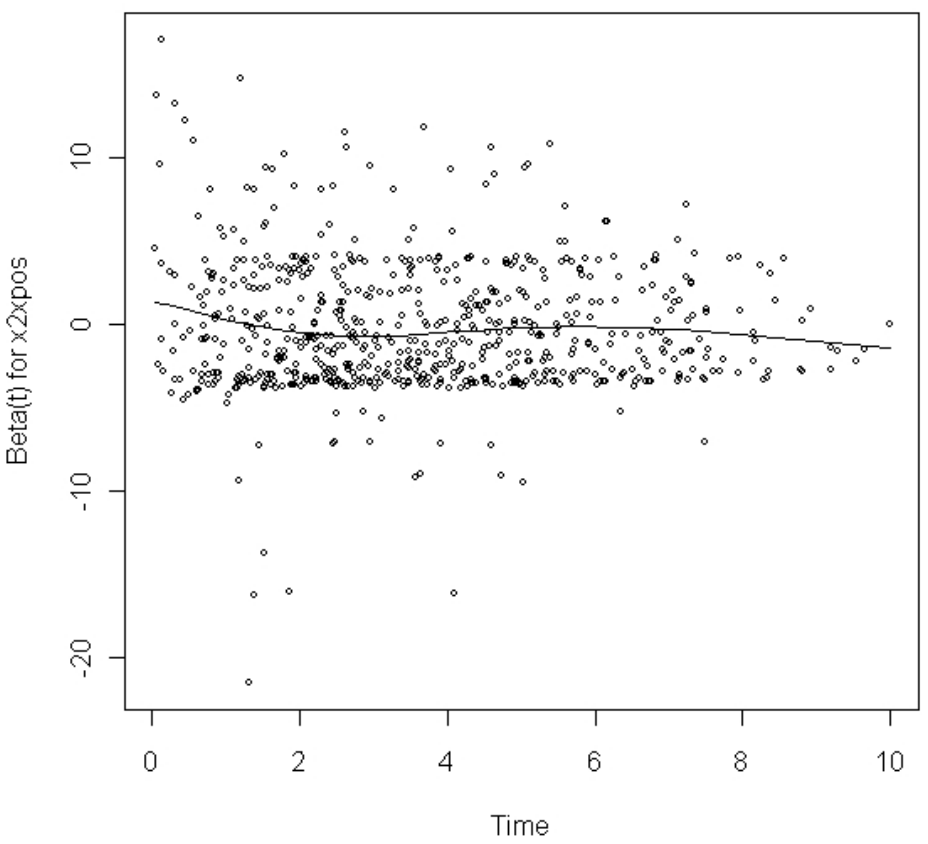
Is the age effect the same over the whole follow-up period?
In other word, is the proportional hazard model acceptable? *is the relative rate β changing with time?*

Is the relative rate β changing with time?

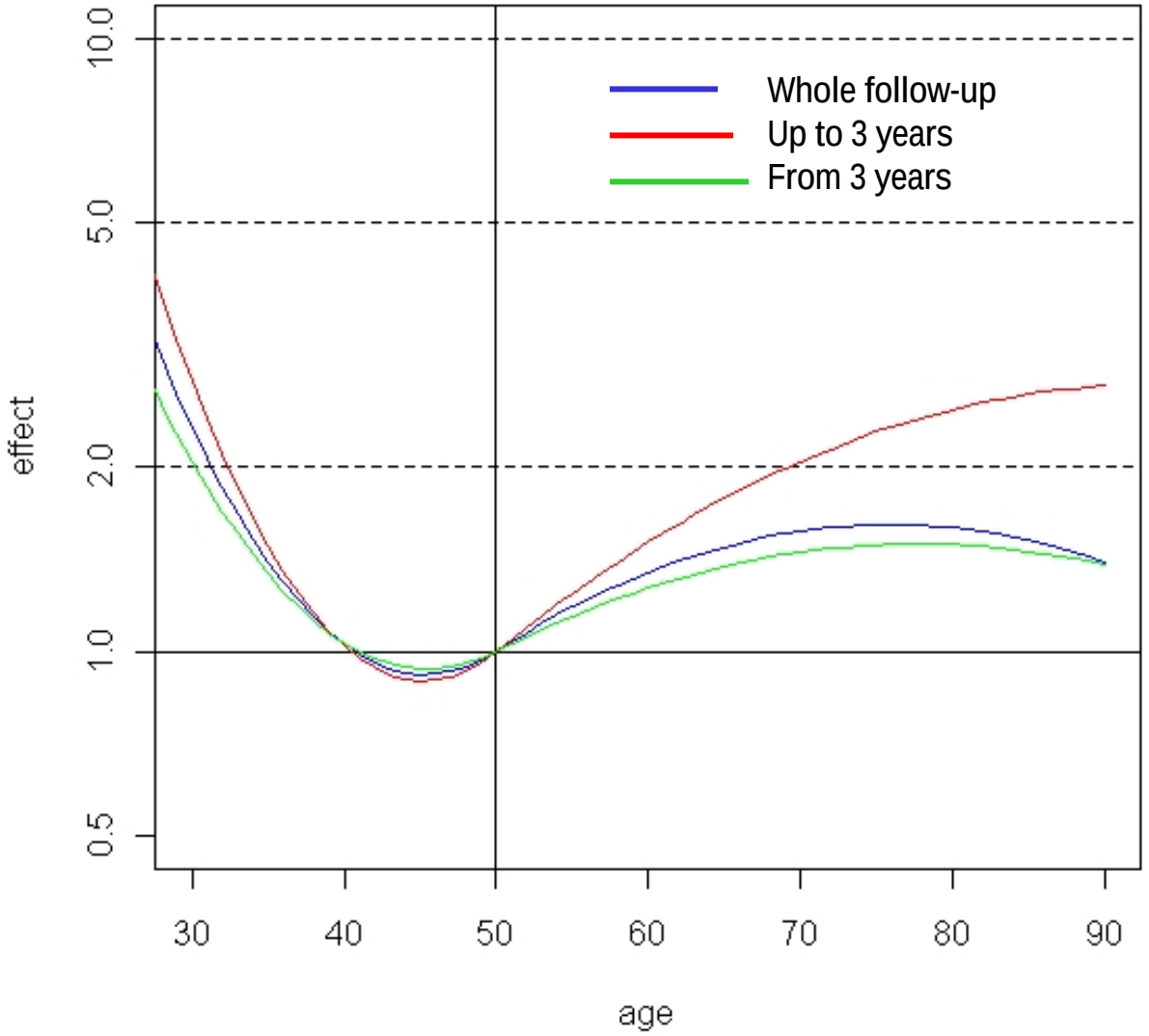
- The approach is the same as that used for crude survival:
  - Compute the Schoenfeld residuals and smooth the associated scatter plot: its graph suggest the functional form of $\beta(t)$ (see Stare *et al*)
  - Fit a model with *time-varying coefficients*
- The latter approach (see *Abramovitz* and *Giorgi*) may be computer intensive if the coefficient is continuous in time
  - The calculation of the cumulative rate in the likelihood implies the evaluation of a complicated integral in a parametric model
  - The interaction term $\Sigma \beta_j g_j(t)z$ must be re-calculated at each failure time in a non-parametric model
- Evaluating the effect for a partition of the follow-up time may be sufficient

Schoenfeld residuals for x^2 and $x^2 x(X>0)$





Changing age effect with follow-up time



Conclusion

- It is clear from our discussion that the "*ratio estimate*" of *relative survival* is less "consistent" than the estimate obtained from the *excess rate model*.
- With little more work *the multivariate regression model* for relative survival will have the same flexibility that the one reached for crude survival (the Cox model tool kit).
- The cancer registries make, at present, little use of this flexibility either because *little promotion of this methodology* has been done or because of the *lack of easy-to-use procedure* in commercially available statistical software.
- The groups engaged in this field of research should make a common effort to correct for this anomaly

Cited Bibliography

M. Abrahamowicz *JASA* 91,1432-1439 (1996)
P. Bolard et al., *J Cancer Epidemiol.Prev.* 7, 113-122 (2002).
G. Cortese and T. H. Scheike, *Stat.Med.* 27, 3563-3584 (2008).
P. W. Dickman, A. Sloggett, M. Hills, T. Hakulinen, *Stat.Med.* 23, 51-64 (2004).
F. Ederer et al., *Natl Cancer Inst.Monogr.* 6:101-21., 101-121 (1961).
J. Esteve, E. Benhamou, M. Croasdale, L. Raymond, *Stat Med* 9, 529-538 (1990).
R. Giorgi et al., *Stat.Med.* 22, 2767-2784 (2003).
T. Hakulinen, *Biometrics.* 38, 933-942 (1982).
T. Hakulinen *et al. Applied Statistics,* 36, 309-17 (1987)
P. Lambert et al, *Stat.Med.* 24, 3871-3885 (2005).
M. P. Perme, R. Henderson, J. Stare, *Biostatistics.* (2008).
M. Pohar and J. Stare, *Comput.Methods Programs Biomed.* 81, 272-278 (2006).
M. Pohar and J. Stare, *Comput.Biol.Med.* 37, 1741-1749 (2007).
L. Remontet, N. Bossard, A. Belot, J. Esteve, *Stat.Med.* 26, 2214-2228 (2007).
P. D. Sasieni, *Biometrika* 83, 127-141 (1996).
J. Stare, M. Pohar, R. Henderson, *Stat.Med.* 24, 3911-3925 (2005).
P. H. Zahl and O. O. Aalen, *Lifetime.Data Anal.* 4, 149-168 (1998).

Book chapters
J.Estève *et al.* Stastitical methods in descriptive epidemiology Ch 4
T.Therneau *et al.* Modeling survival data, Ch 10