Flexible parametric survival models in cancer epidemiology

Paul W Dickman

University of Lund
LUPOP seminar series
14 Oct 2020
http://pauldickman.com/talk/

Today's talk

- Overview of some of my research, focussing on those projects that extend and/or apply flexible parametric survival models.
- A non-technical introduction to flexible parametric survival models and why I like them.
- I have many more slides than I can show in an hour; will tailor the talk based on questions and audience interest.

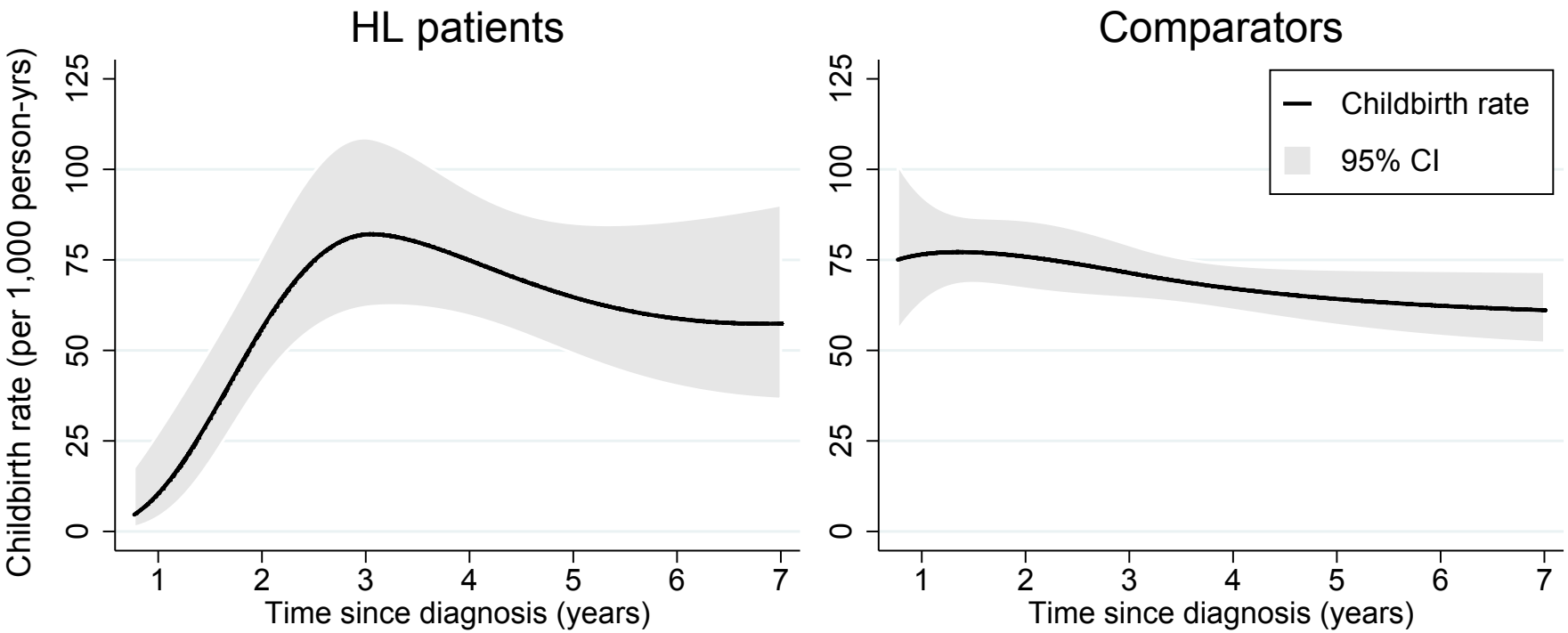
My research interests

- Primary research interests are in the development and application of methods for population-based cancer survival analysis, particularly the estimation and modeling of net survival.
- General interest in statistical aspects of the design, analysis, and reporting of epidemiological studies.
- Epidemiology, with particular focus on cancer epidemiology and perinatal/reproductive epidemiology.
- Co-PI for a node of the VR funded environment grant within register-based research. Focus of our node is methods for register-based research.
- Working to establish a masters program in biostatistics and develop doctoral education in biostatistics.

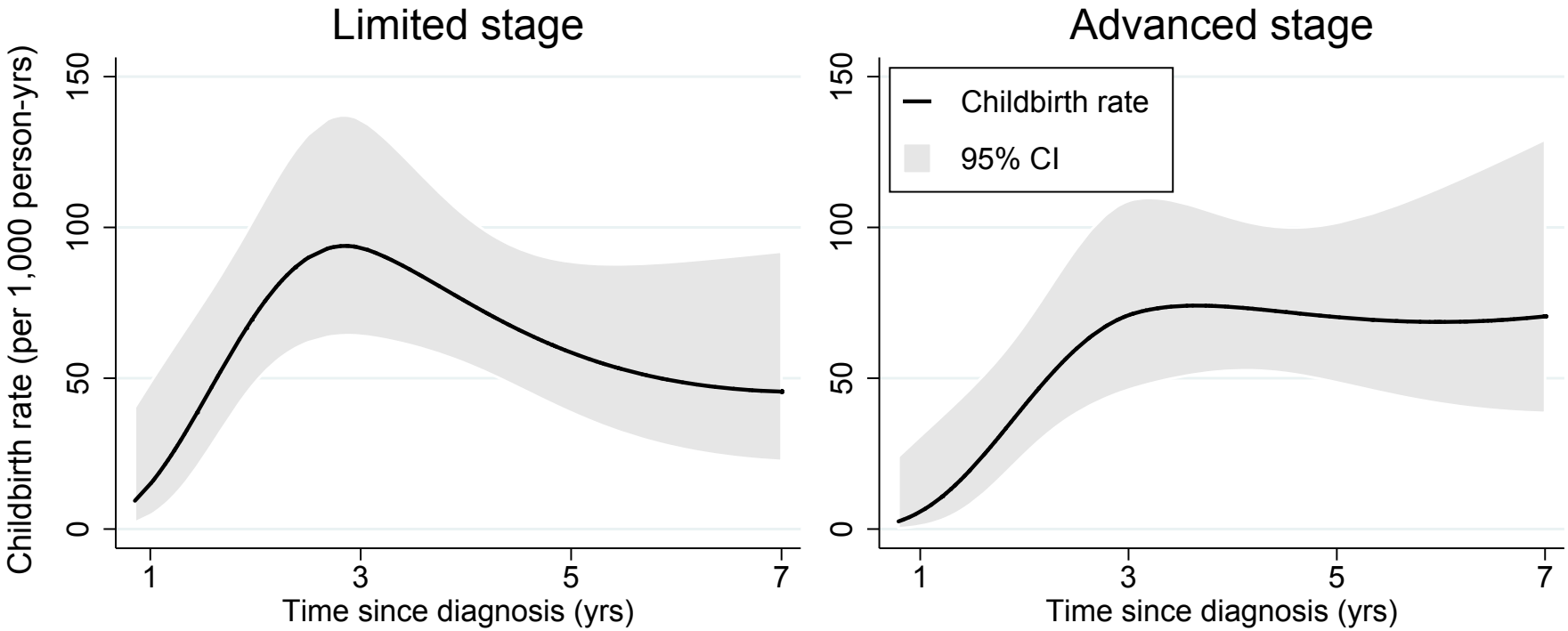
Why I use parametric survival models

- I analyse large population-based datasets where
  - The proportional hazards assumption is rarely appropriate.
  - The hazard function is of interest.
  - A hazard ratio does not tell the whole story.
- I model excess mortality/net survival among cancer patients.
  - Not possible to fit the Cox model.
  - Proportional excess hazards rarely true.
  - Quantities other than the excess hazard ratio are of interest.
- Quantification and presentation of absolute risks and rates.
  - Should be done more than it is.
  - Much easier if you estimate the baseline.
- Many useful extensions are much easier in a parametric setting.

Childbirth rates among Hodgkin lymphoma survivors in Sweden (Weibull *et al.* 2018 [1])



Childbirth rates among Hodgkin lymphoma survivors in Sweden (Weibull *et al.* 2018 [1])



Sex differences in bladder cancer survival [3]

European Journal of Cancer 95 (2018) 52–58

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Bladder cancer survival: Women better off in the long run

Bettina Kulle Andreassen [a,*], Tom Kristian Grimsrud [a], Erik Skaaheim Haug [b,c]

- See Radkiewicz *et al.* (2017) [2] for a similar Swedish study.

Time-varying excess hazard ratio [3]

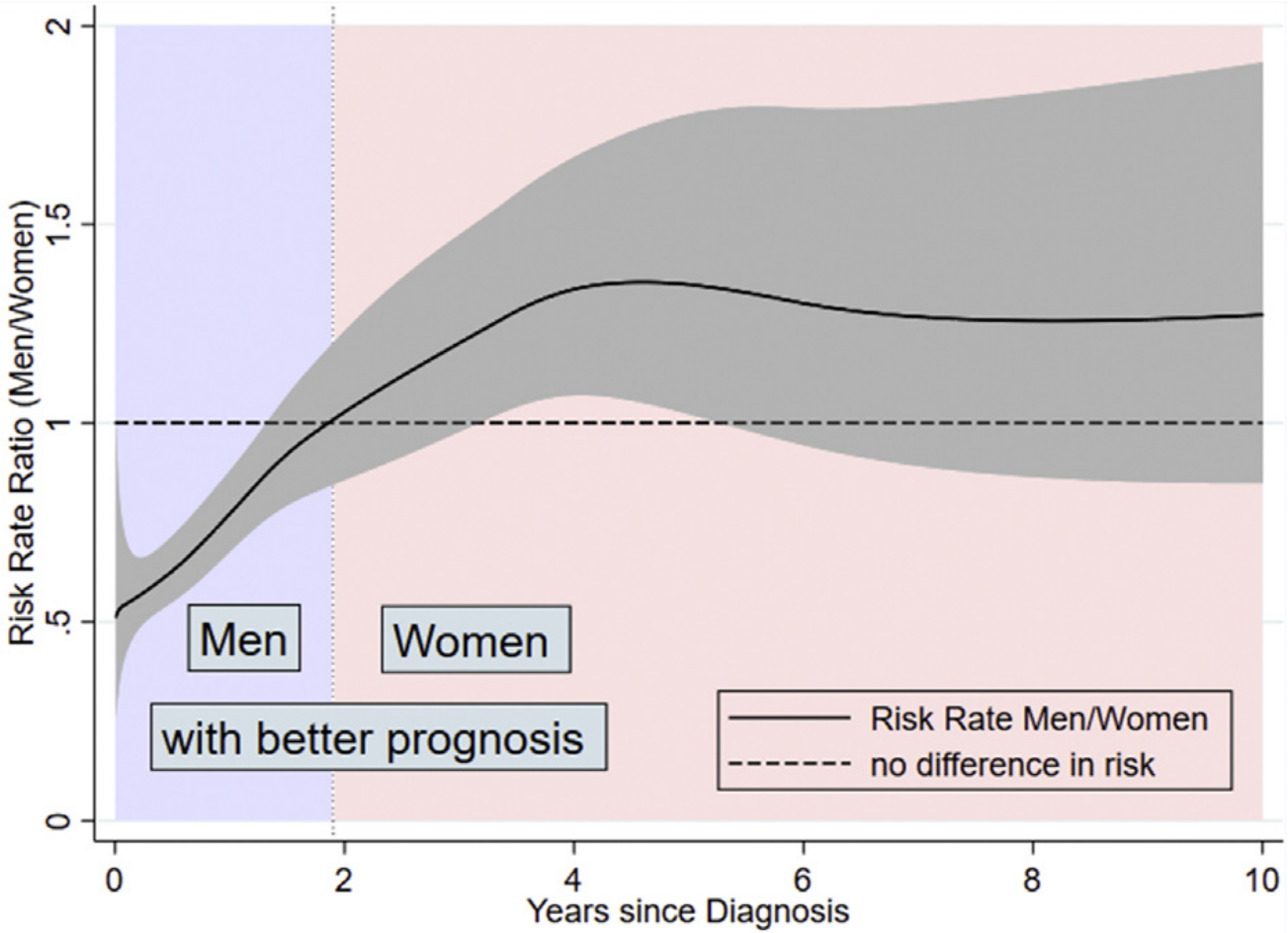


Fig. 2. Risk ratio (excess mortality rate ratio) including confidence intervals for men versus women with bladder cancer diagnosis. The

'The hazard ratio has a built-in selection bias' [4]

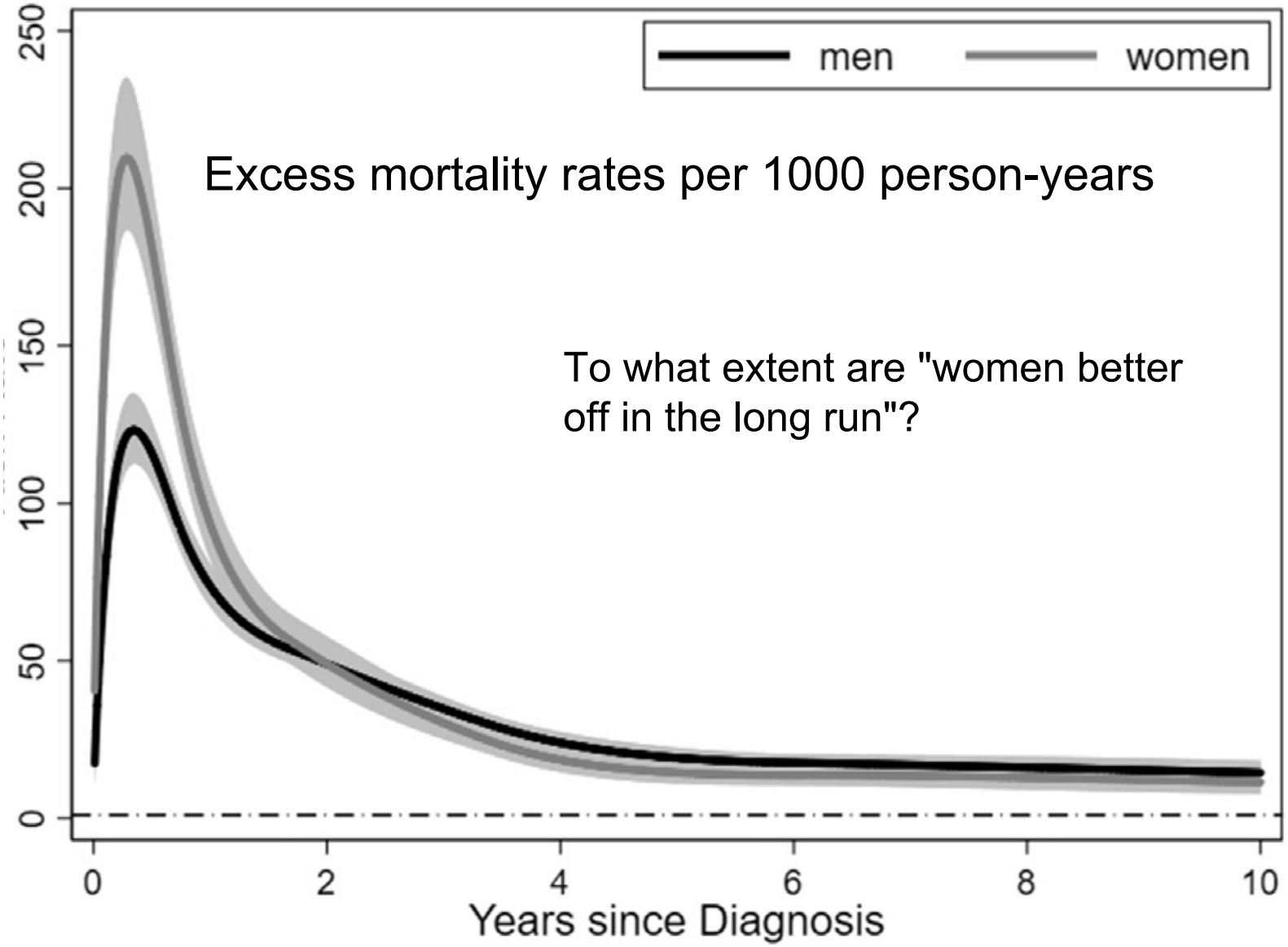
The Hazards of Hazard Ratios

Miguel A. Hernán

The hazard ratio (HR) is the main, and often the only, effect measure reported in many epidemiologic studies. For dichotomous, non–time-varying exposures, the HR is defined as the hazard in the exposed groups divided by the hazard in the unexposed groups. For all practical purposes, hazards can be thought of as incidence rates and thus the HR can be roughly interpreted as the incidence rate ratio. The HR is commonly and conveniently estimated via a Cox proportional hazards model, which can include potential confounders as covariates.

Unfortunately, the use of the HR for causal inference is not straightforward even in the absence of unmeasured confounding, measurement error, and model misspecification. Endowing a HR with a causal interpretation is risky for 2 key reasons: the HR may change over time, and the HR has a built-in selection bias. Here I review these 2 problems and

Baseline excess mortality rates [3]



Marginal and standardised survival [3, 5]

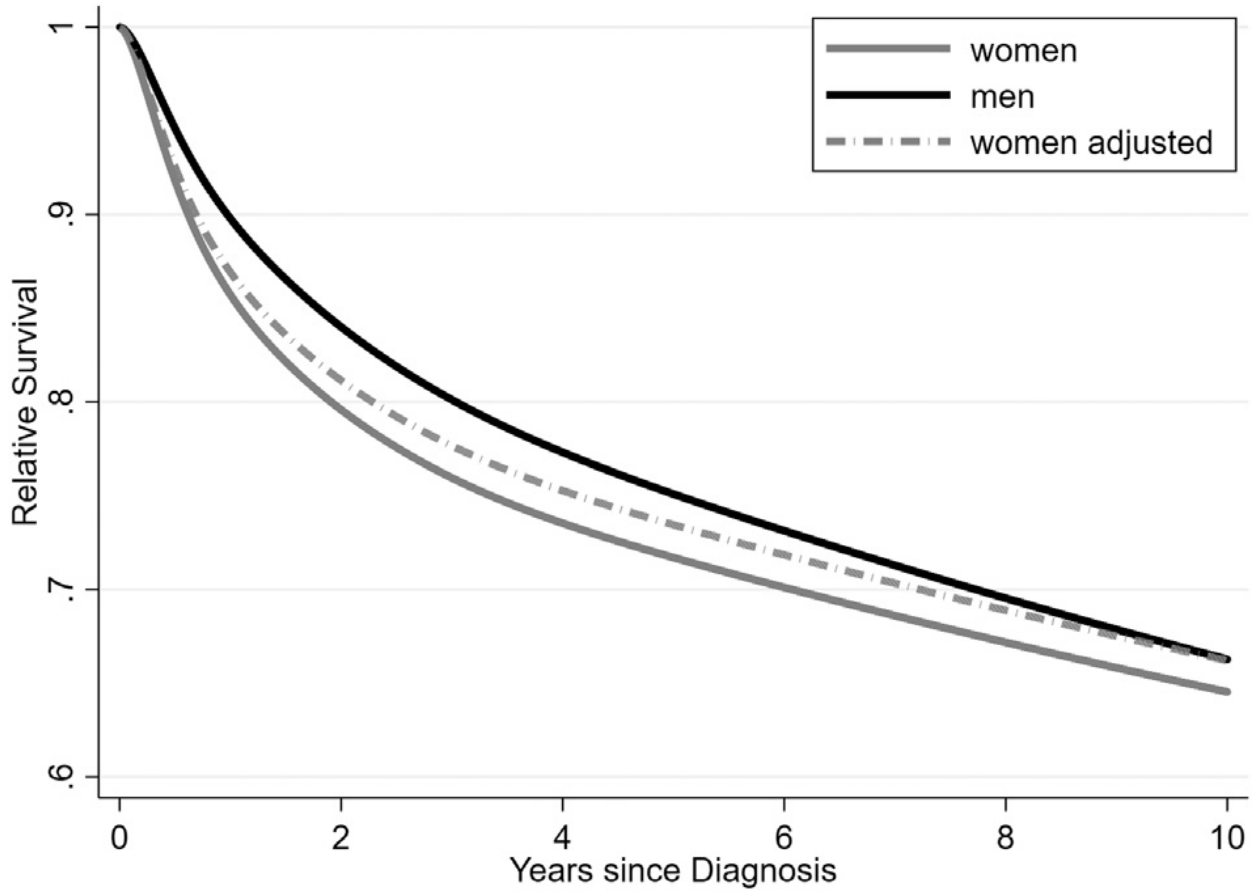
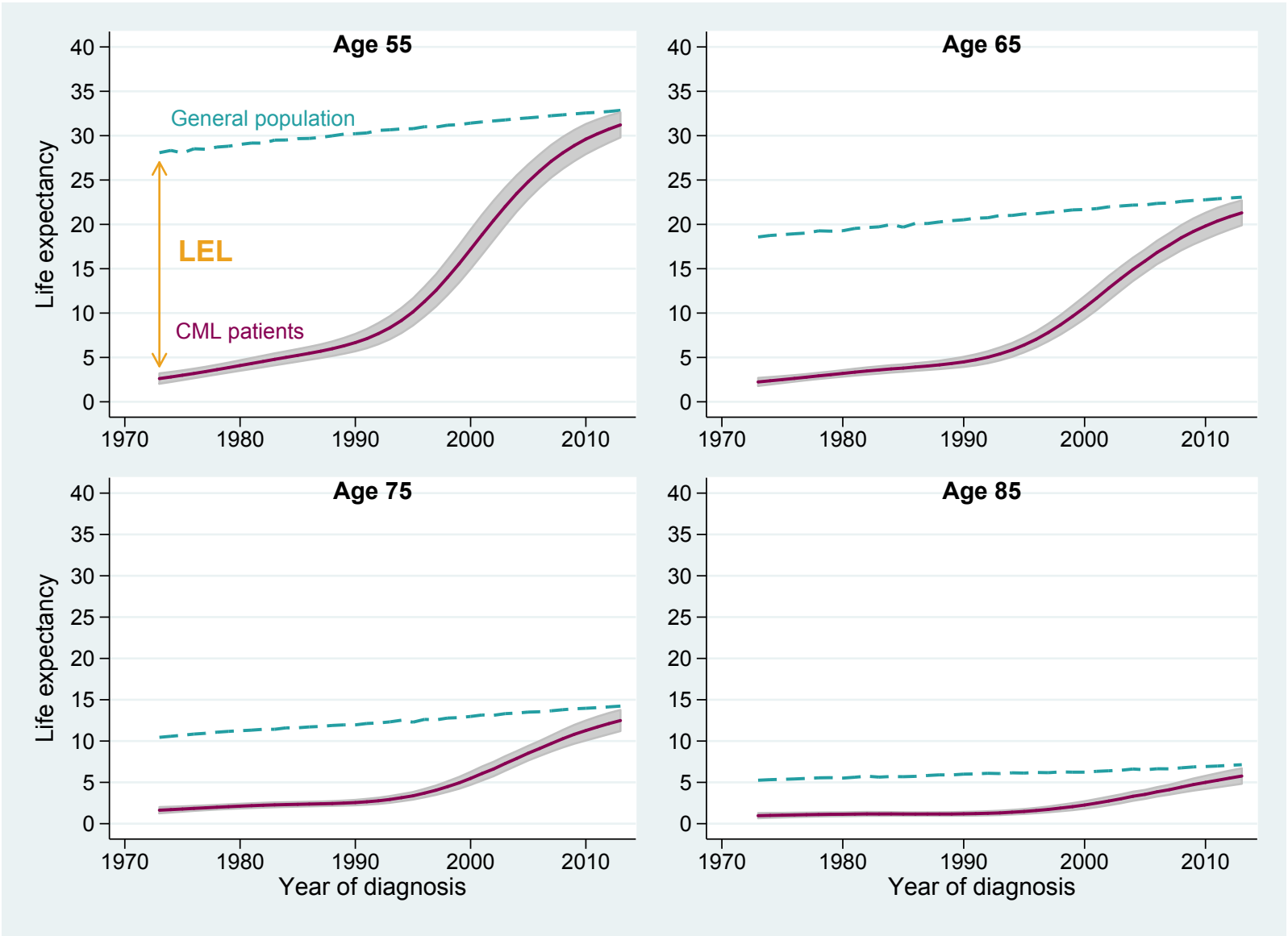


Fig. 3. Relative survival for men, women and women assuming the same T-stage distribution as men. Black (grey) lines: mean survival curve for men (women); Dashed grey line: survival curve for women when assuming men's covariate pattern.

Loss in expectation of life: CML (Sweden) [6]



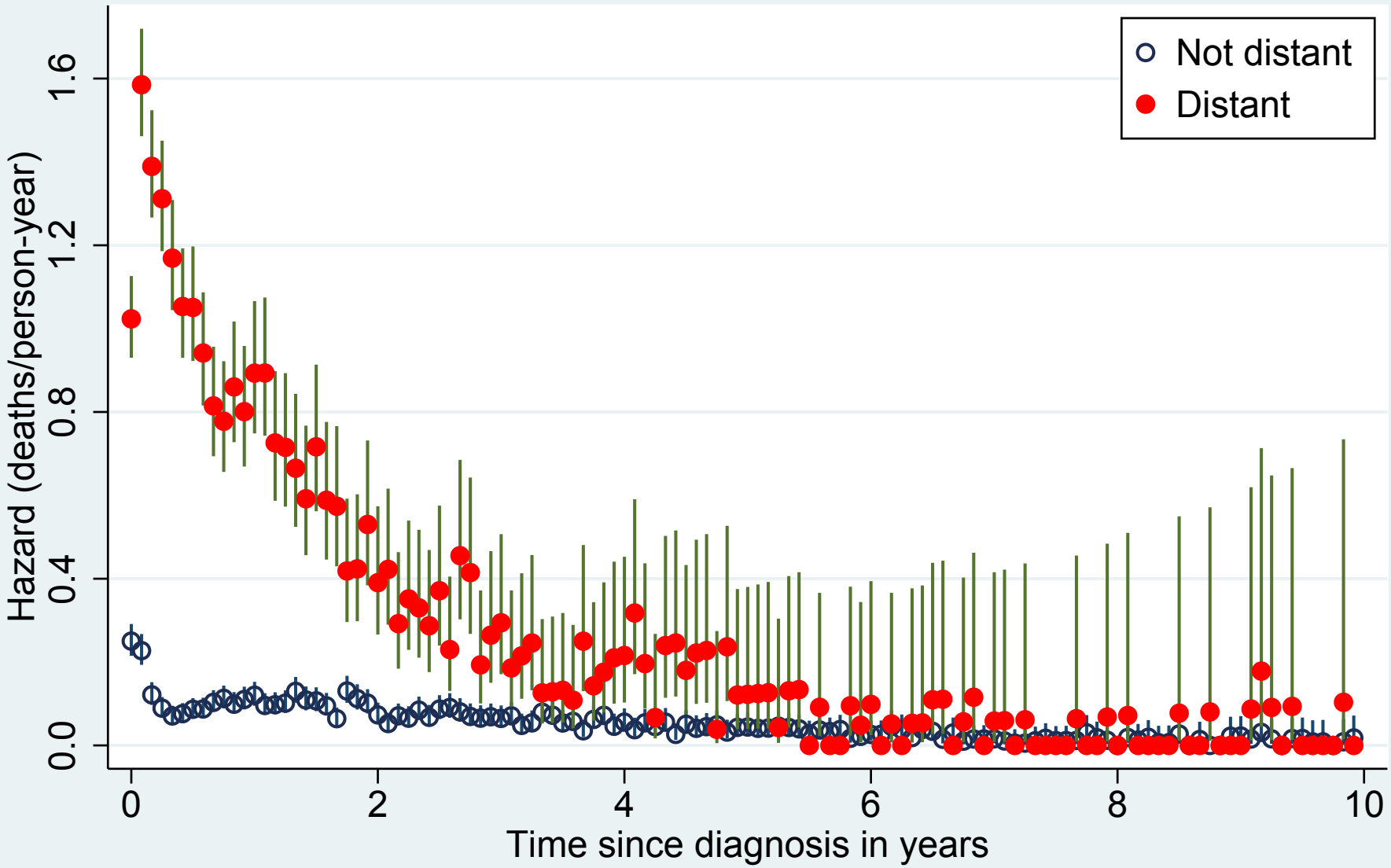
A sneak peek at my conclusions

- I use and advocate flexible parametric survival models. However,
- There is nothing wrong with using a Cox model.
- If you only want to estimate a hazard ratio and you 'know' you have proportional hazards then a Cox model is ideal.
- Can relax the PH assumption in the Cox model, and can estimate quantities other than HR.
- However, a parametric approach makes it easier to estimate quantities that provide more insight and may be more relevant to your research question.
- You will get the same hazard ratio, but a whole lot more.

Example: survival of patients diagnosed with colon carcinoma

- Patients diagnosed with colon carcinoma 1984–95. Potential follow-up to end of 1995; censored after 10 years.
- Outcome is death due to colon carcinoma.
- Interest is in the effect of clinical stage at diagnosis (distant metastases vs no distant metastases).
- How might we specify a statistical model for these data?

Empirical hazards by stage



The Cox model (Cox 1972 [7])

- Over 33,000 citations (Web of Science, November 2018).
- 24th on Nature's 2014 list of most-cited paper of all time for all fields.

$$h_i(t|\mathbf{x}_i) = h_0(t)\exp(\mathbf{x}_i\beta)$$

- Estimates (log) hazard ratios.
- Advantage: The baseline hazard, $h_0(t)$ is not estimated.
- Disadvantage: The baseline hazard, $h_0(t)$ is not estimated.

An interview with Sir David Cox (Reid 1994 [8])

Reid "What do you think of the cottage industry that's grown up around [the Cox model]?"

Cox "In the light of further results one knows since, I think I would normally want to tackle the problem parametrically. . . . I'm not keen on non-parametric formulations normally."

Reid "So if you had a set of censored survival data today, you might rather fit a parametric model, even though there was a feeling among the medical statisticians that that wasn't quite right."

Cox "That's right, but since then various people have shown that the answers are very insensitive to the parametric formulation of the underlying distribution. And if you want to do things like predict the outcome for a particular patient, it's much more convenient to do that parametrically."

Some common survival models in epidemiology

- Commonly used models have the same basic formulation.

$$h_i(t) = h_0(t)\exp(\mathbf{x_i}\beta)$$

$$\ln(h_i(t)) = \ln(h_0(t)) + \mathbf{x_i}\beta$$

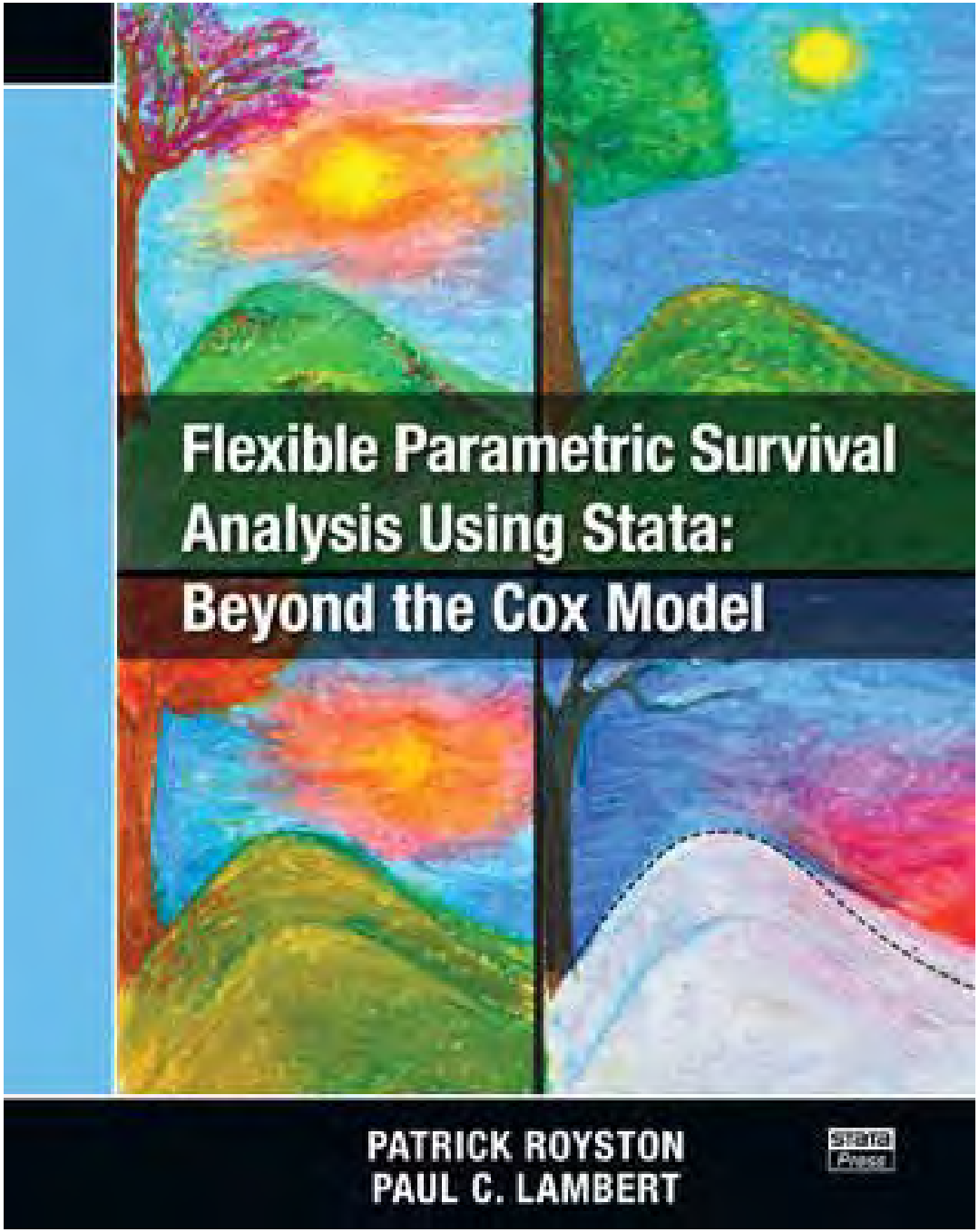
- Proportional hazards assumed by default (but can be relaxed).
- Primary difference is in specification of the baseline hazard:
  - Cox model: $h_0(t)$ an arbitrary function of time; not estimated.
  - Poisson regression model: $h_0(t)$ is a step function.
  - Weibull model: $h_0(t) = \lambda\gamma t^{\gamma-1}$
  - Flexible parametric model: $h_0(t)$ modelled using splines.

Flexible Parametric Survival Models [9, 10, 11]

- First introduced by Royston and Parmar (2002) [9].
- Applicable for 'standard' and relative survival models.
- Usually easier to model on the log cumulative hazard scale
- $\ln(H_0(t))$ modelled using restricted cubic splines.

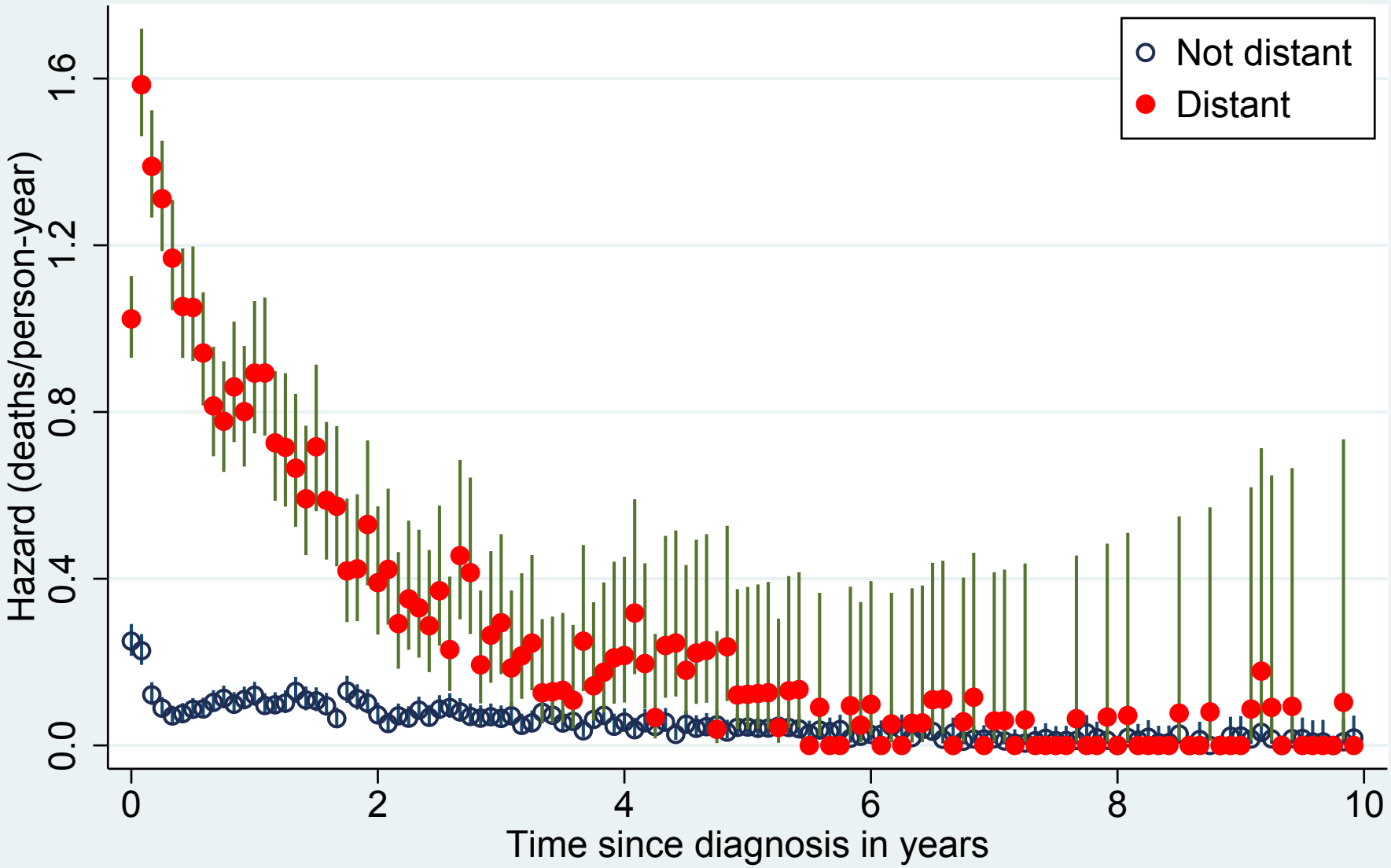
$$H_i(t) = H_0(t)\exp(\mathbf{x_i}\beta)$$

$$\ln(H_i(t)) = \ln(H_0(t)) + \mathbf{x_i}\beta$$


Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model
PATRICK ROYSTON
PAUL C. LAMBERT
Stata Press

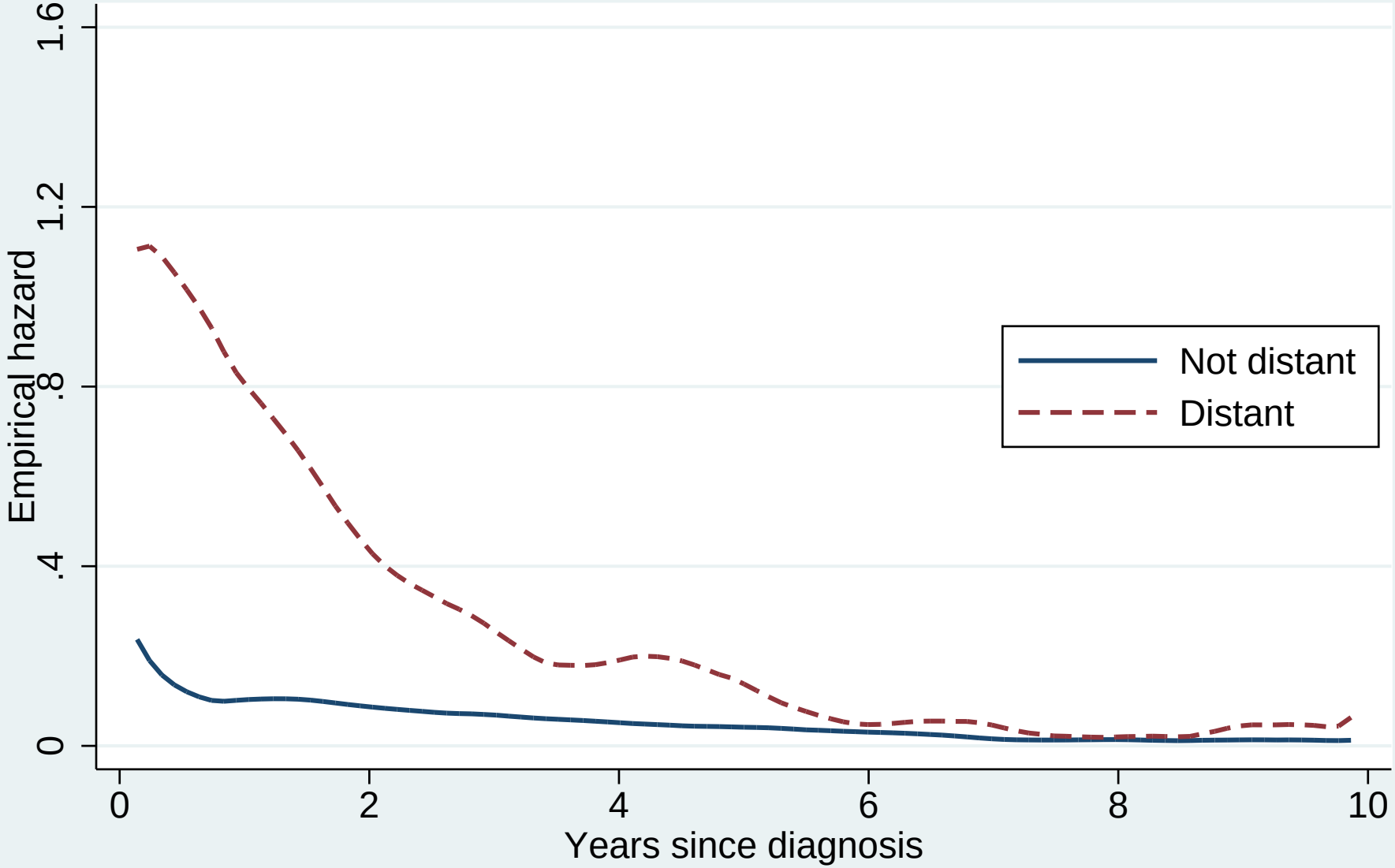

www.cansurv.net

Empirical hazards by stage



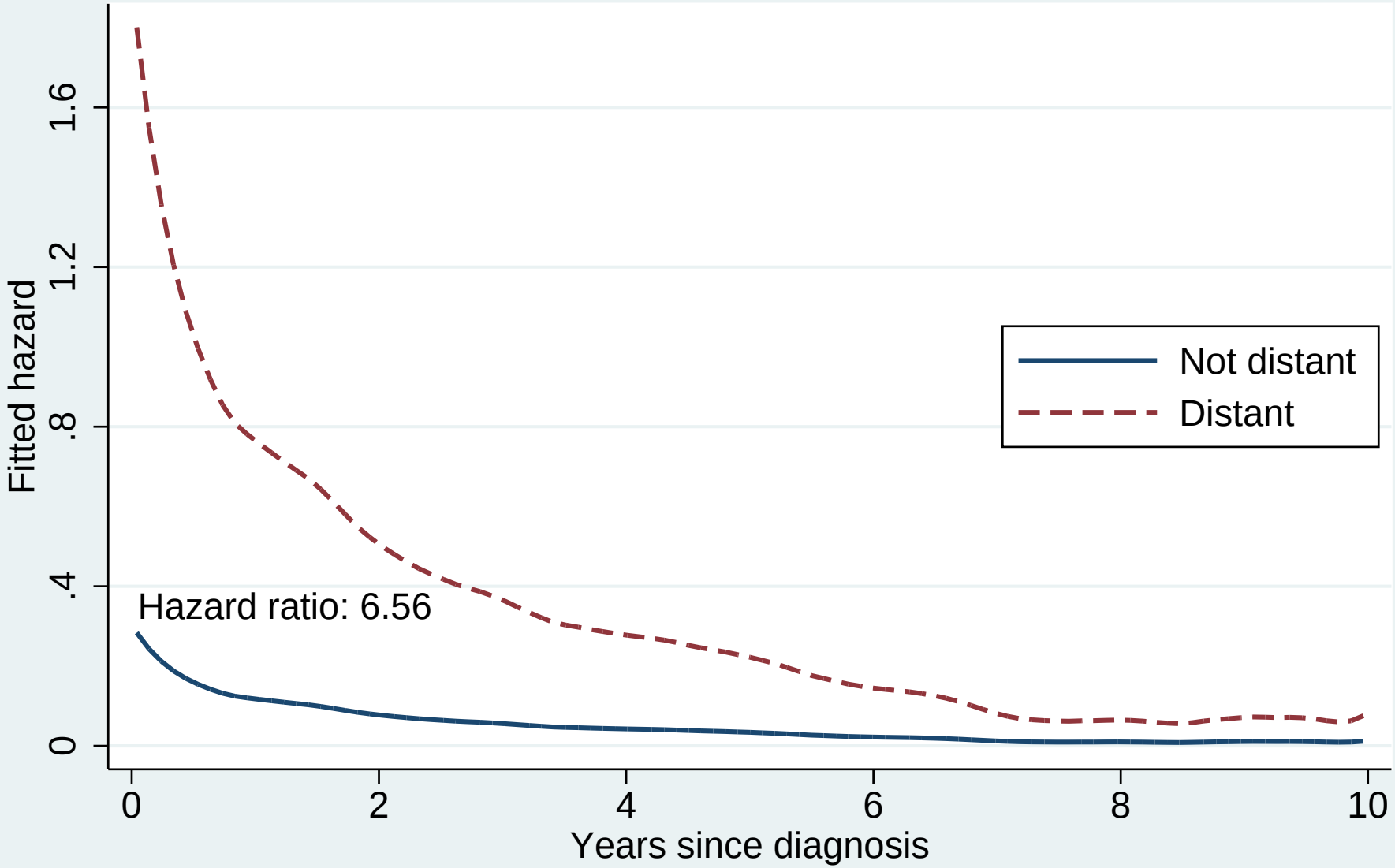
Smoothed empirical hazards (cancer-specific mortality rates)

sts graph, by(distant) hazard kernel(epan2)

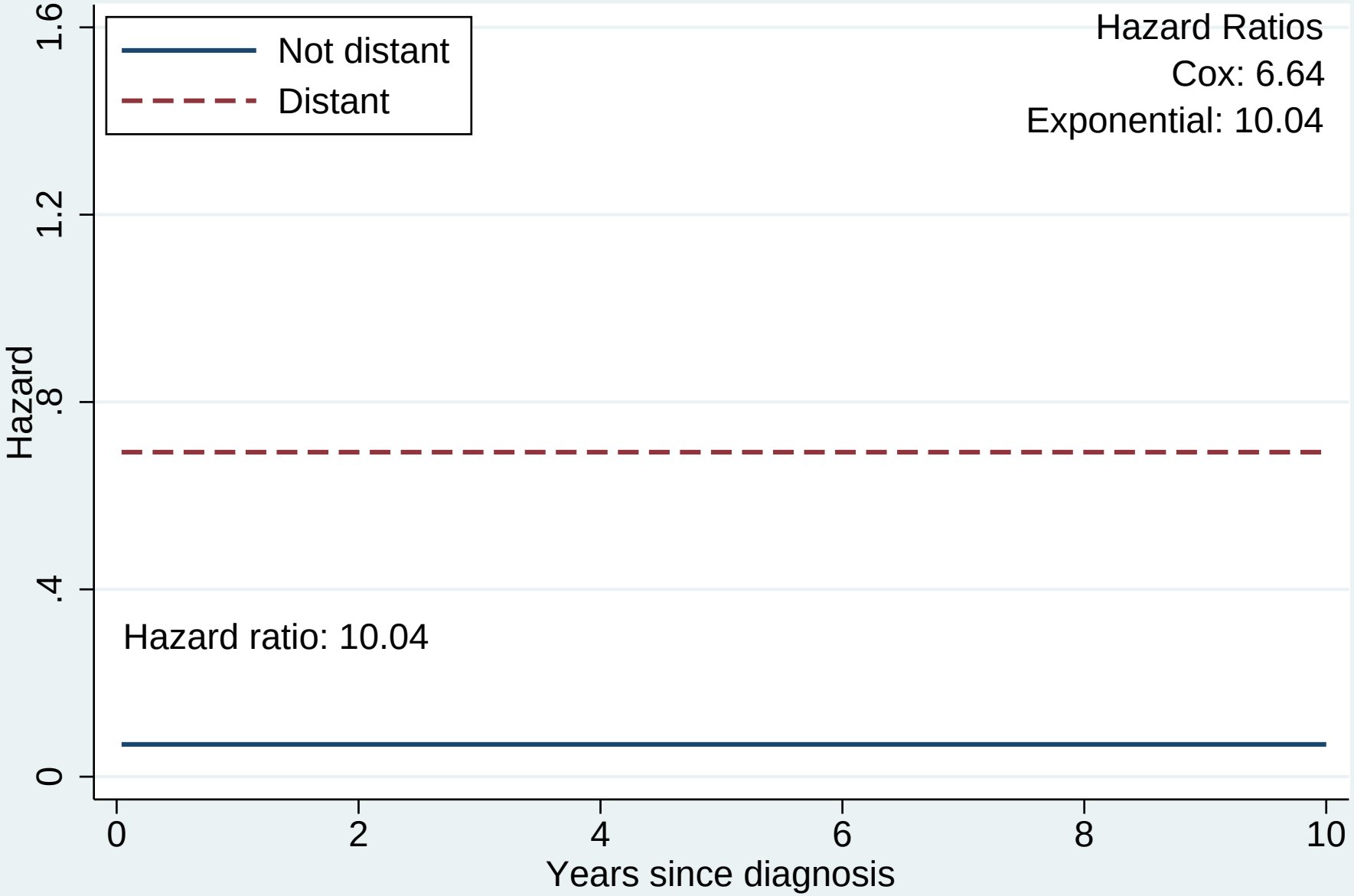


Fitted hazards from Cox model

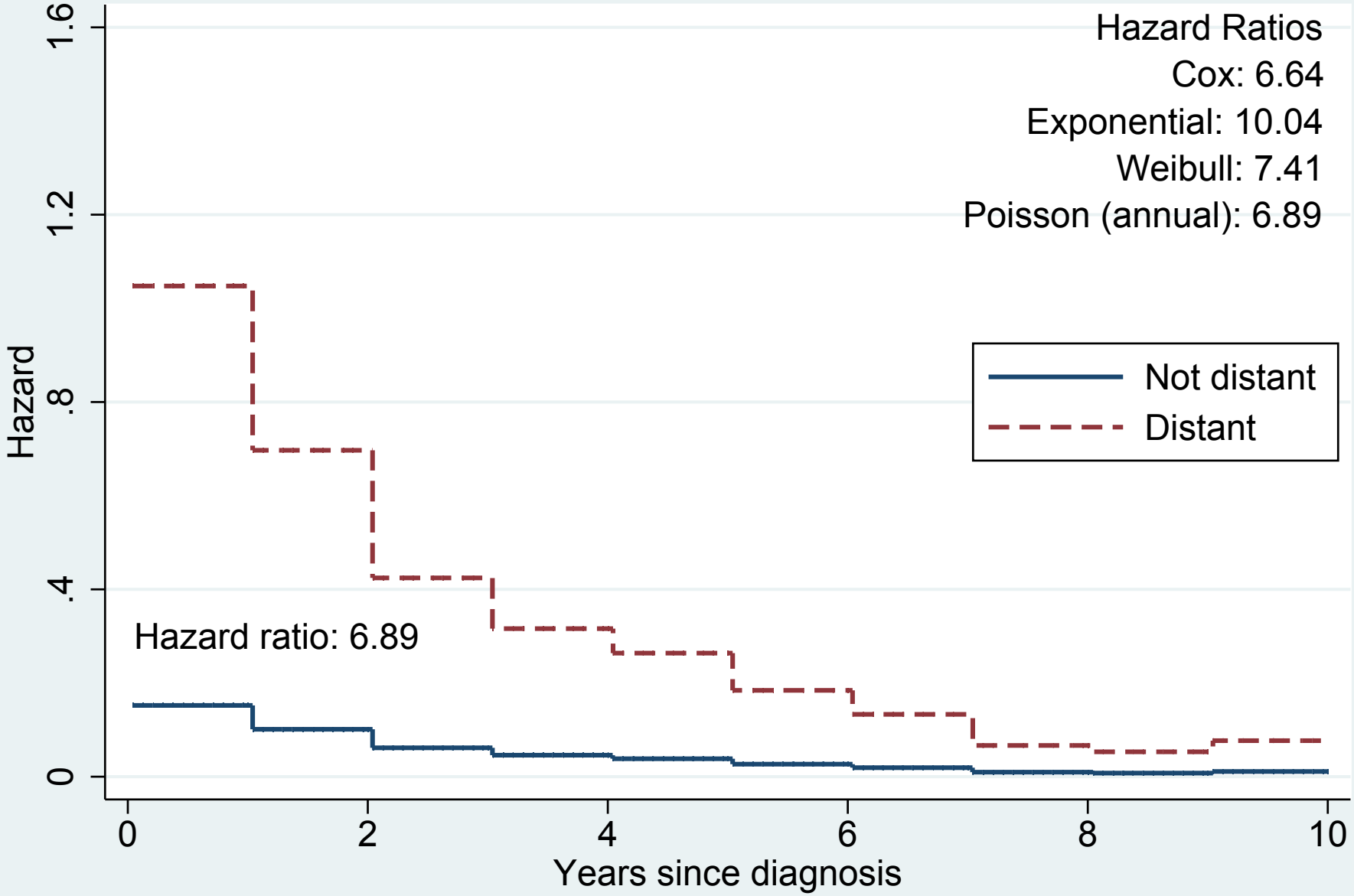
stcurve, hazard at1(distant=0) at2(distant=1) kernel(epan2)



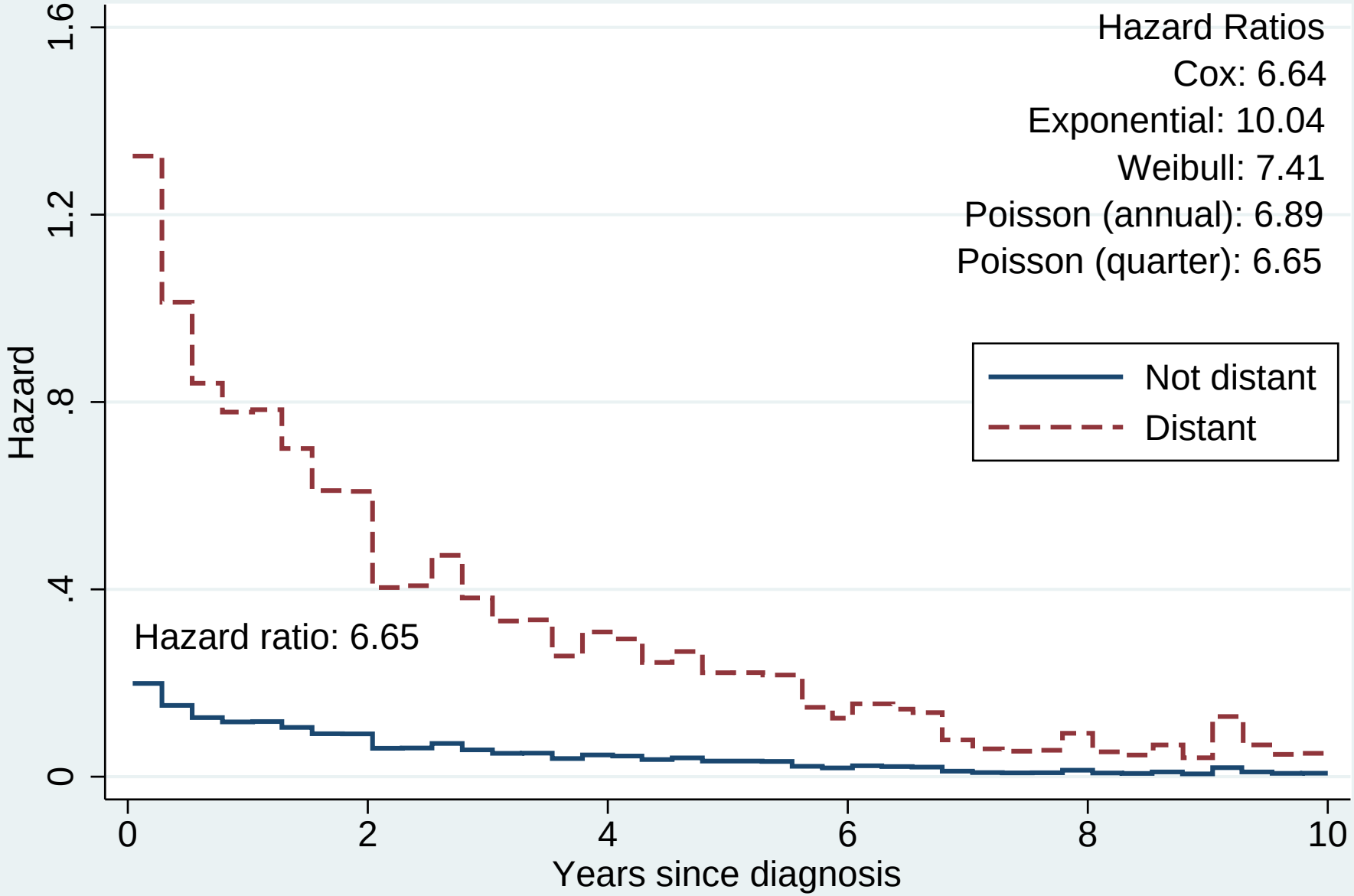
Fitted hazards from parametric survival model (exponential)



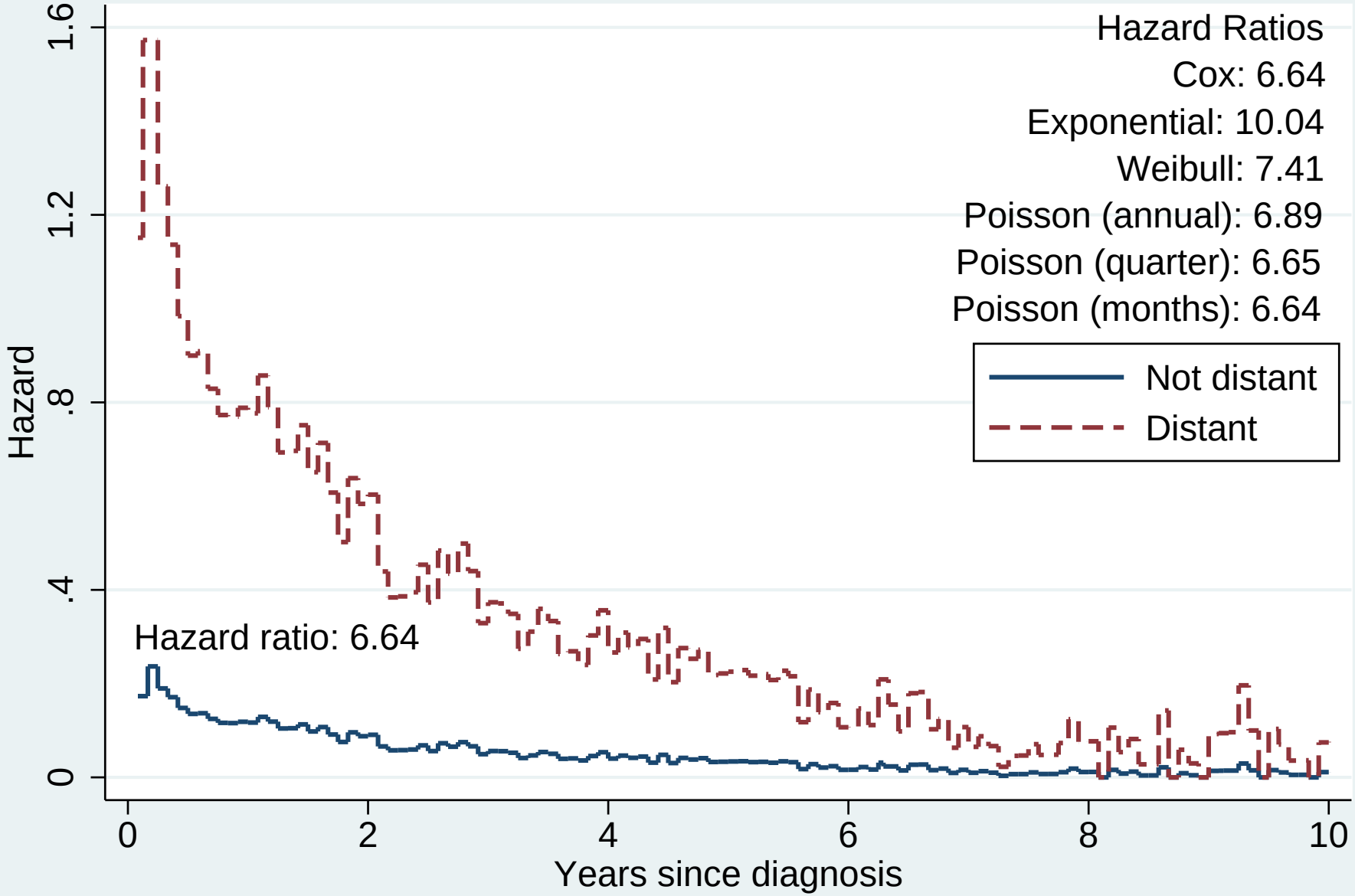
Fitted hazards from Poisson model (yearly intervals)



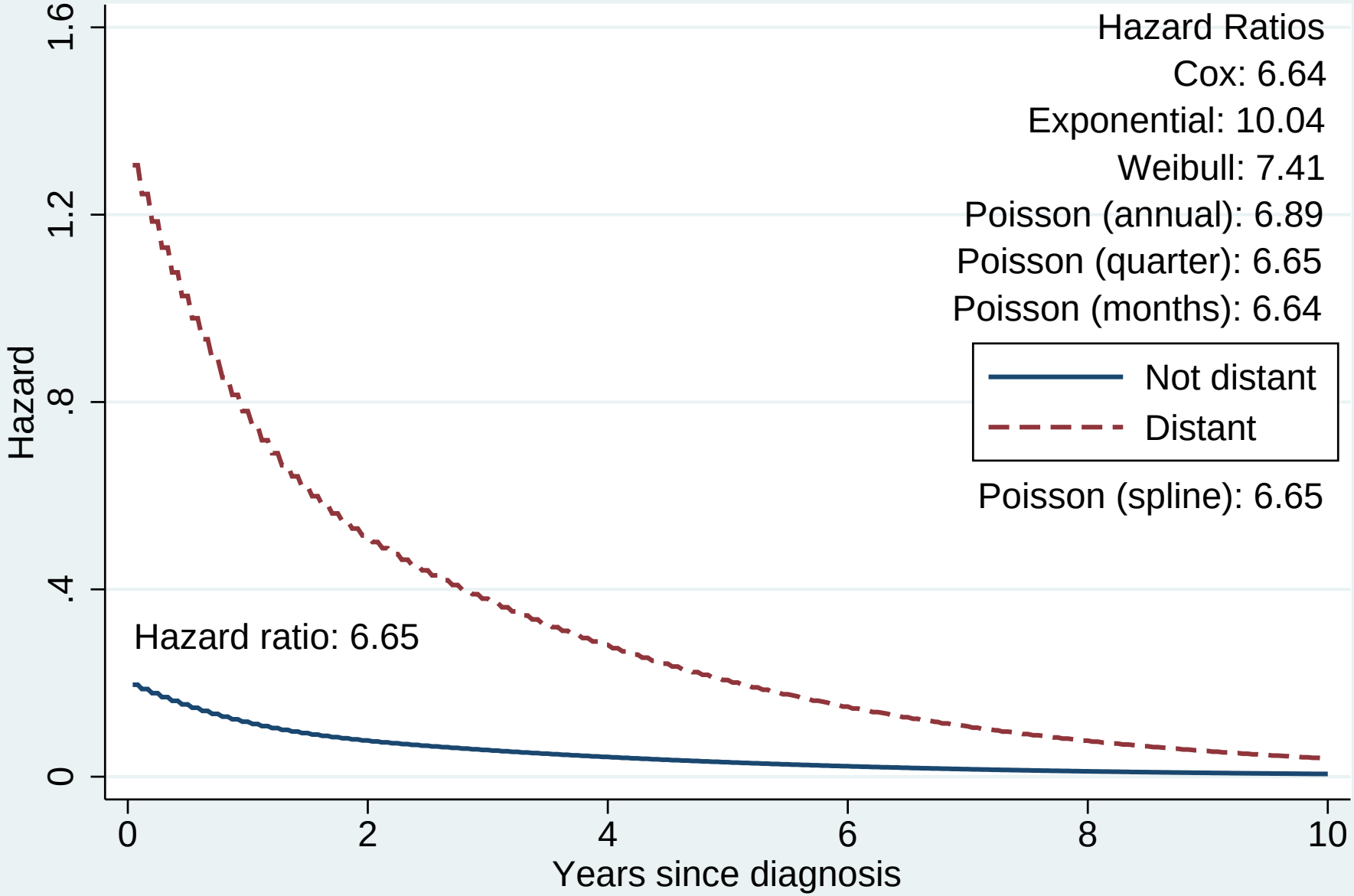
Fitted hazards from Poisson model (3-months)



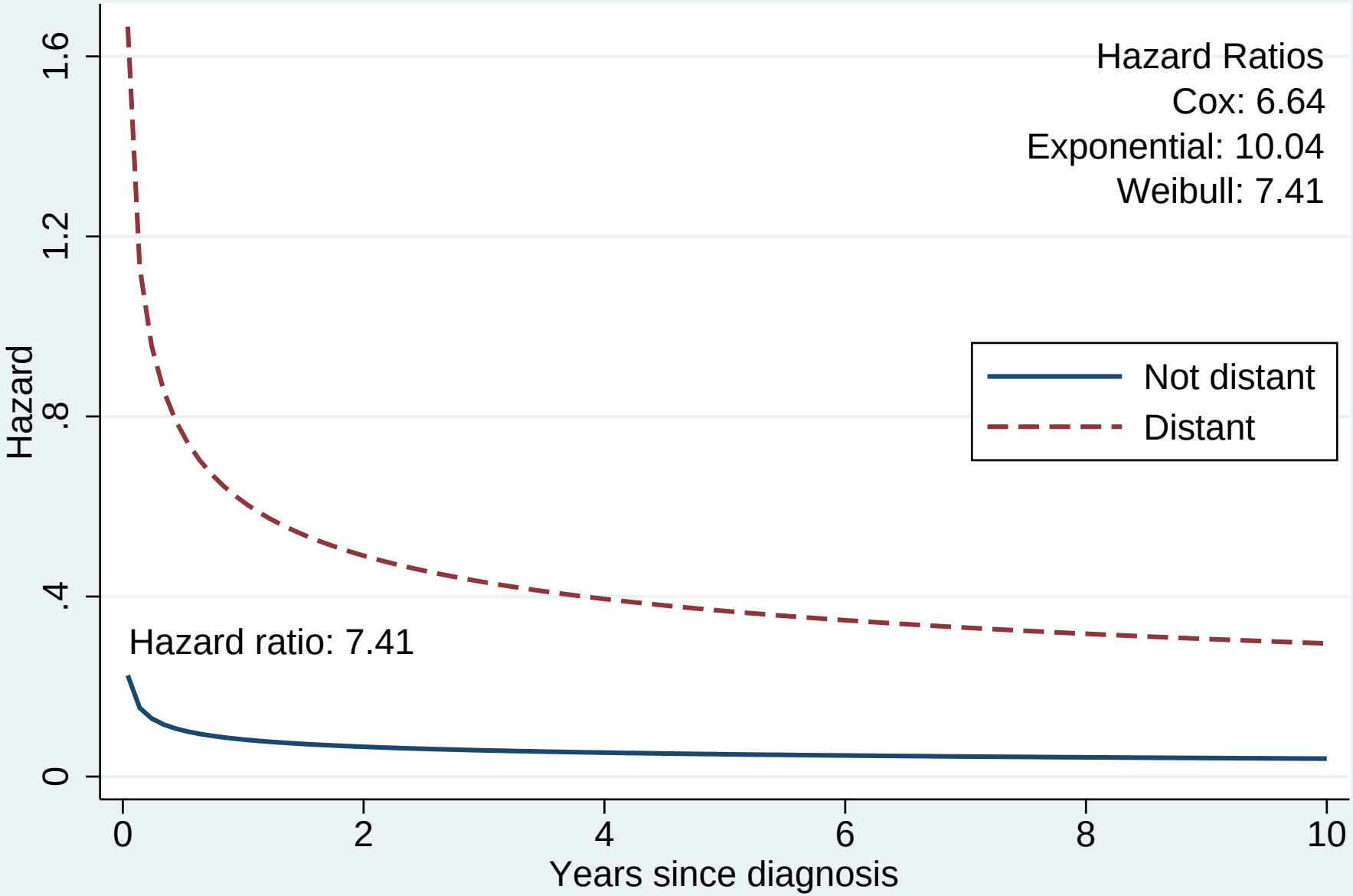
Fitted hazards from Poisson model (months)



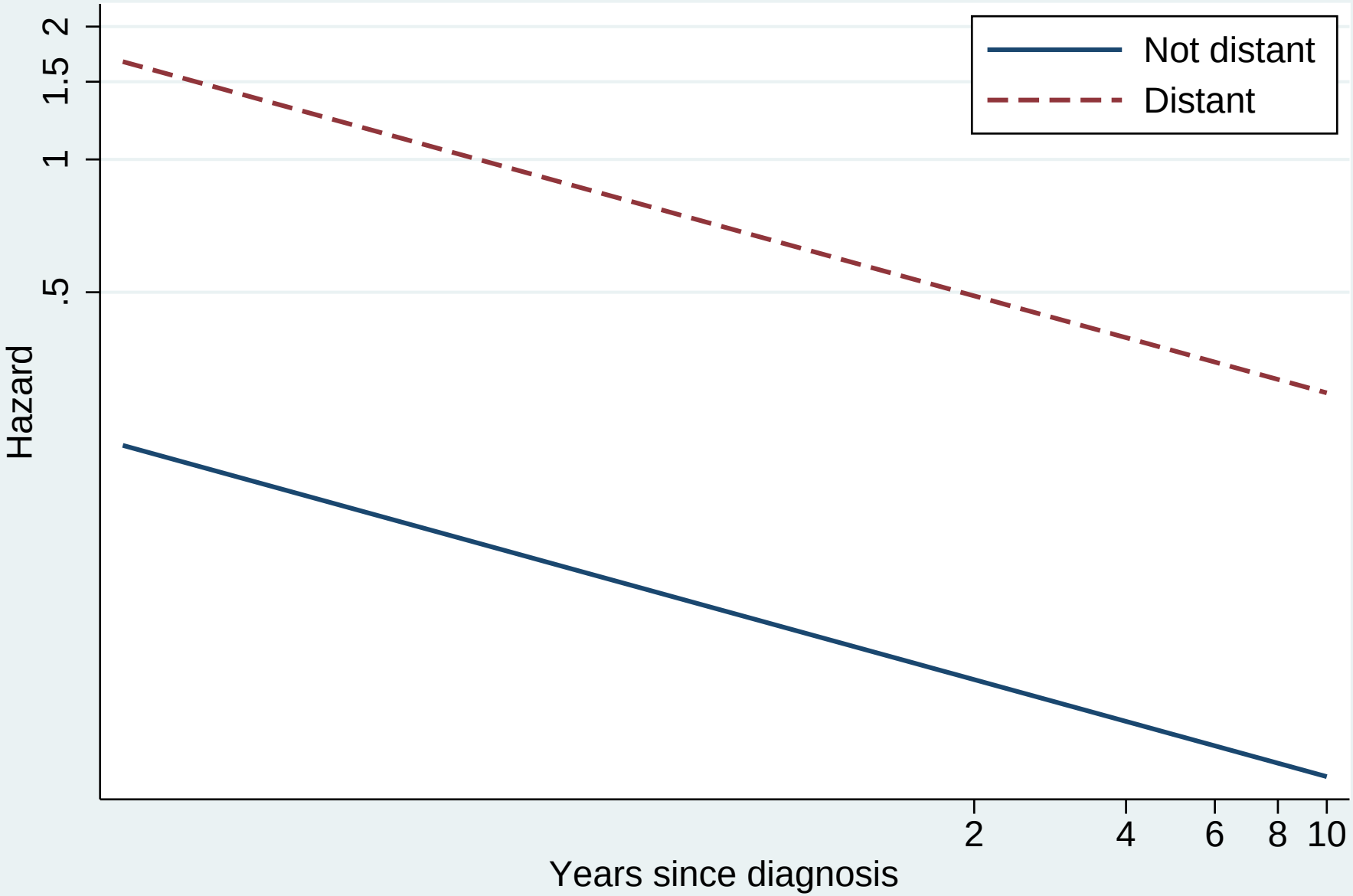
Fitted hazards from Poisson model (rcs 5df)



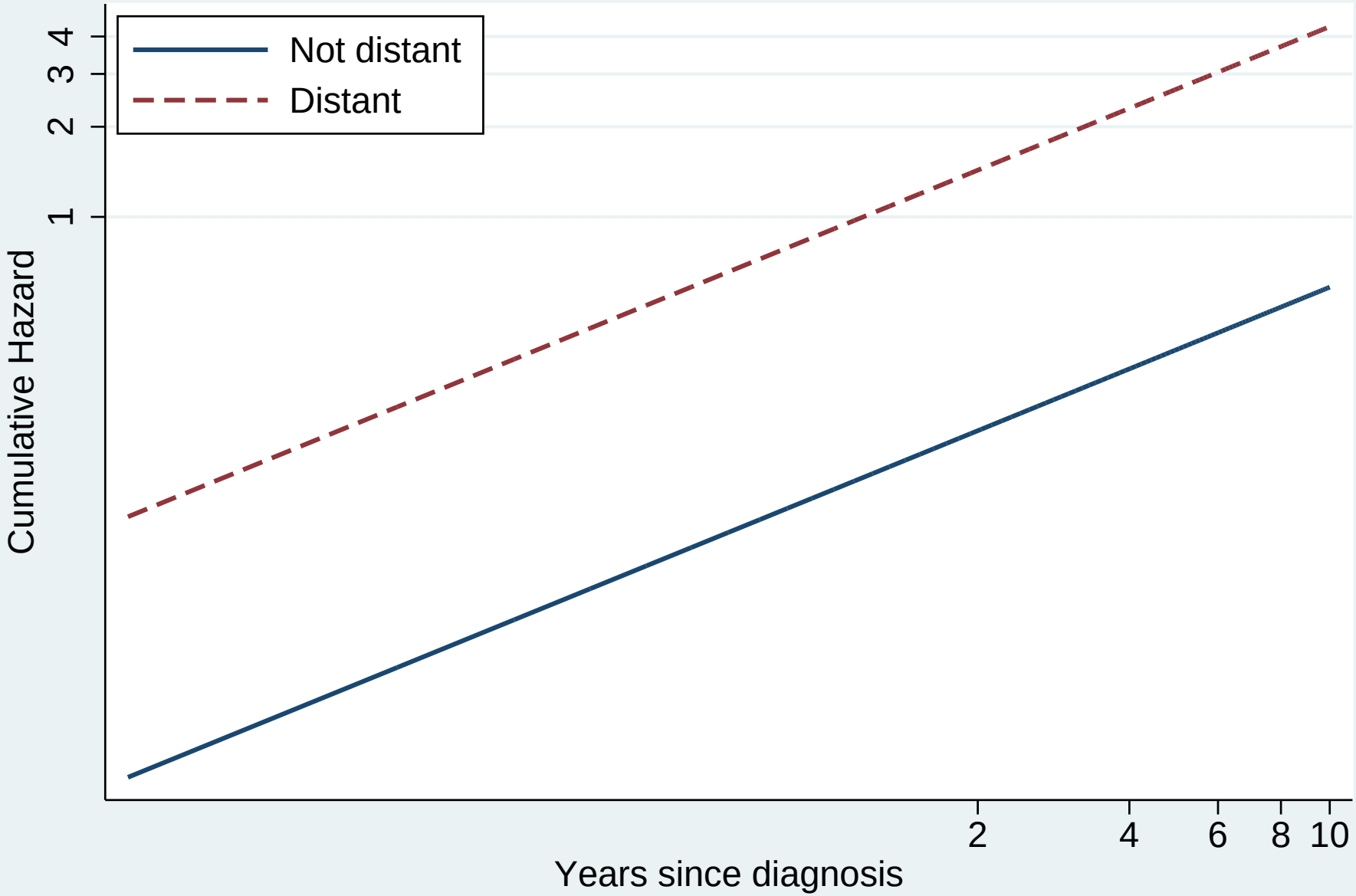
Fitted hazards from parametric survival model (Weibull)

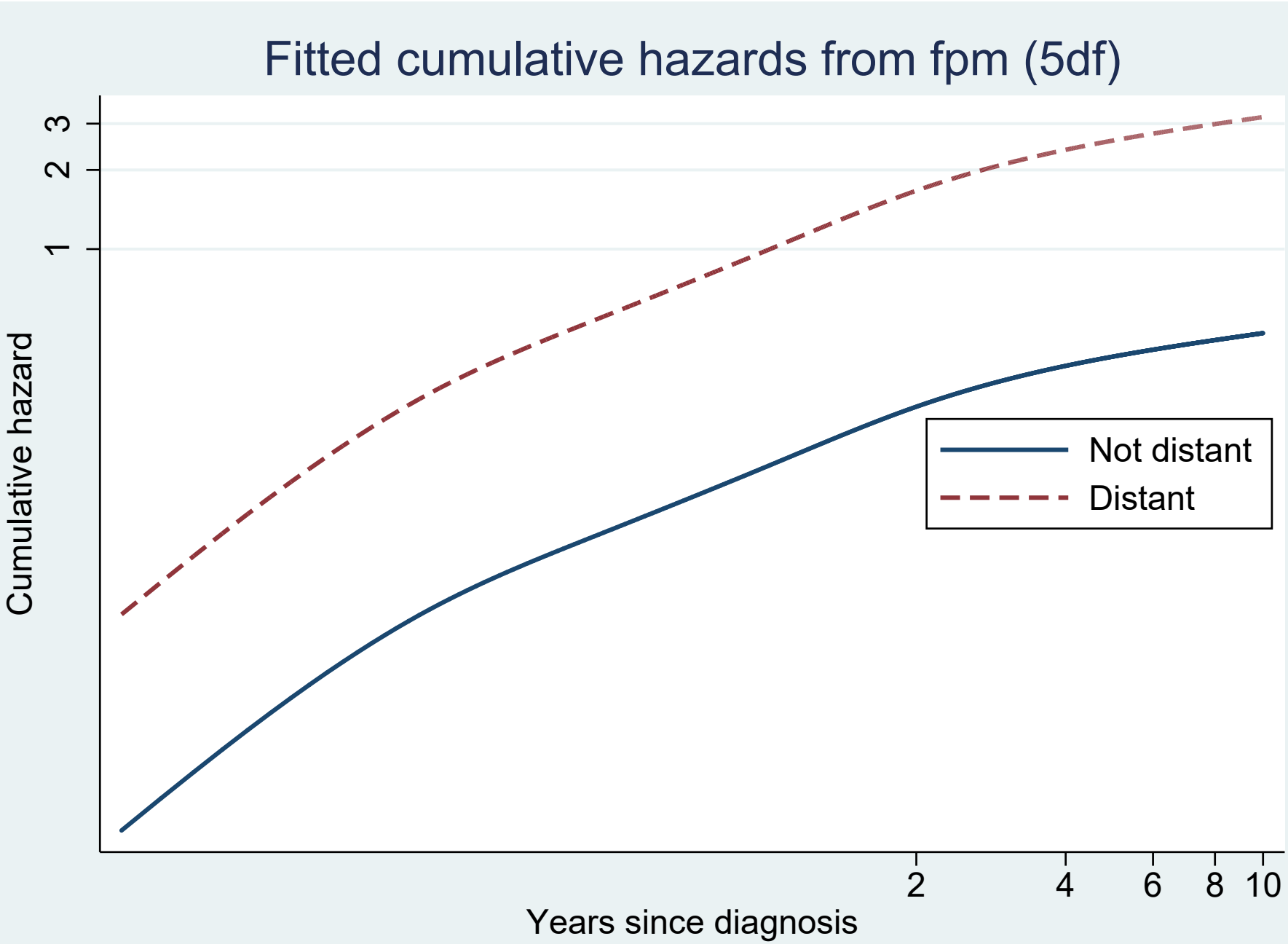


Fitted hazards from parametric survival model (Weibull)

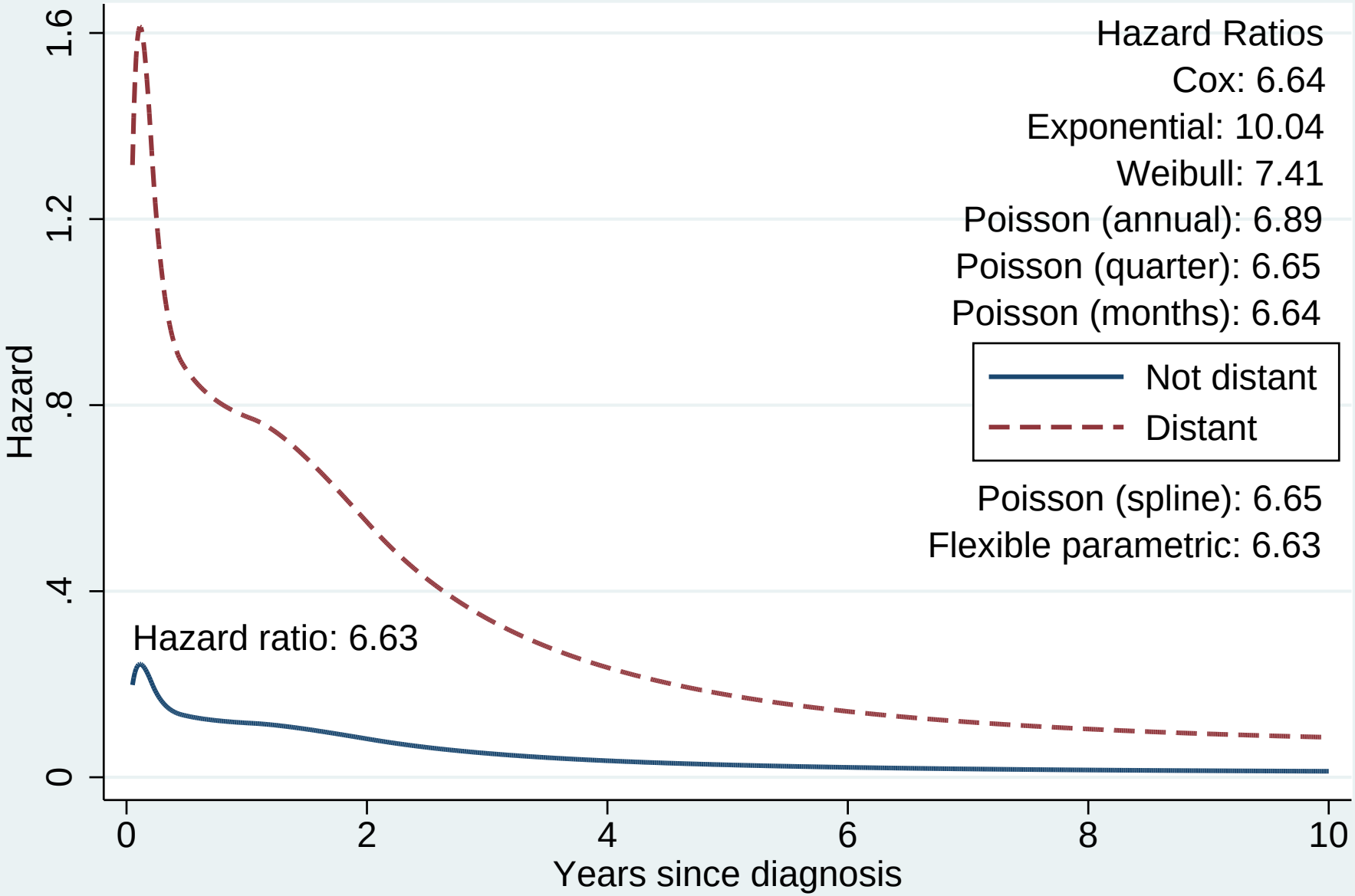


Fitted cumulative hazards from Weibull model




Fitted cumulative hazards from fpm (5df)
Cumulative hazard
1
2
3
Not distant
Distant
2
4
6
8
10
Years since diagnosis

Fitted hazards from flexible parametric model (5df)



Do splines capture the underlying shape?

- With any statistical method we need to assess its performance.
- We have performed a number of simulation studies.
- In summary, the models can capture many complex shapes of the underlying hazard and survival functions for both proportional hazards [12] and, importantly, when relaxing the proportional hazards assumption [13].
- Results are, in general, not sensitive to choices of number and location of knots.

Example using attained age as the time-scale

- Study from Sweden[14] comparing incidence of hip fracture of,
  - 17,731 men diagnosed with prostate cancer treated with bilateral orchiectomy (surgical removal of testicles).
  - 43,230 men diagnosed with prostate cancer not treated with bilateral orchiectomy.
  - 362,354 men randomly selected from the general population.
- Study entry is 6 months post diagnosis.
- Outcome is femoral neck fracture.
- Attained age is used as the primary time-scale.
- Provides estimates of age-specific incidence rates.
- Actually two timescales of interest, but we will initially ignore time since exposure.

Estimates from a proportional hazards model

- Estimated IRRs compared to population comparators.

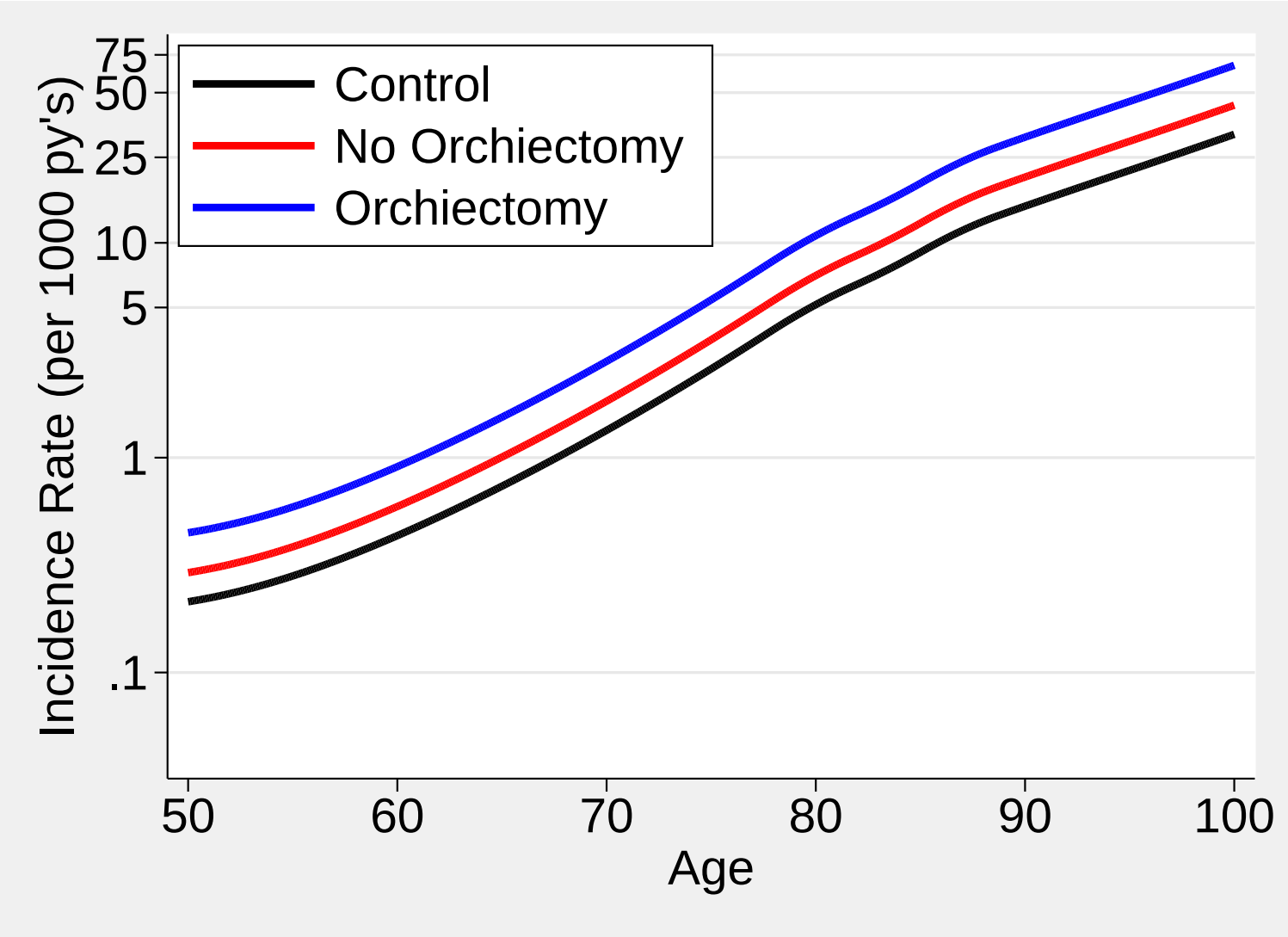
Cox Model

Incidence rate ratio (no orchiectomy)	=	1.37 (1.28 to 1.46)
Incidence rate ratio (orchiectomy)	=	2.09 (1.93 to 2.27)

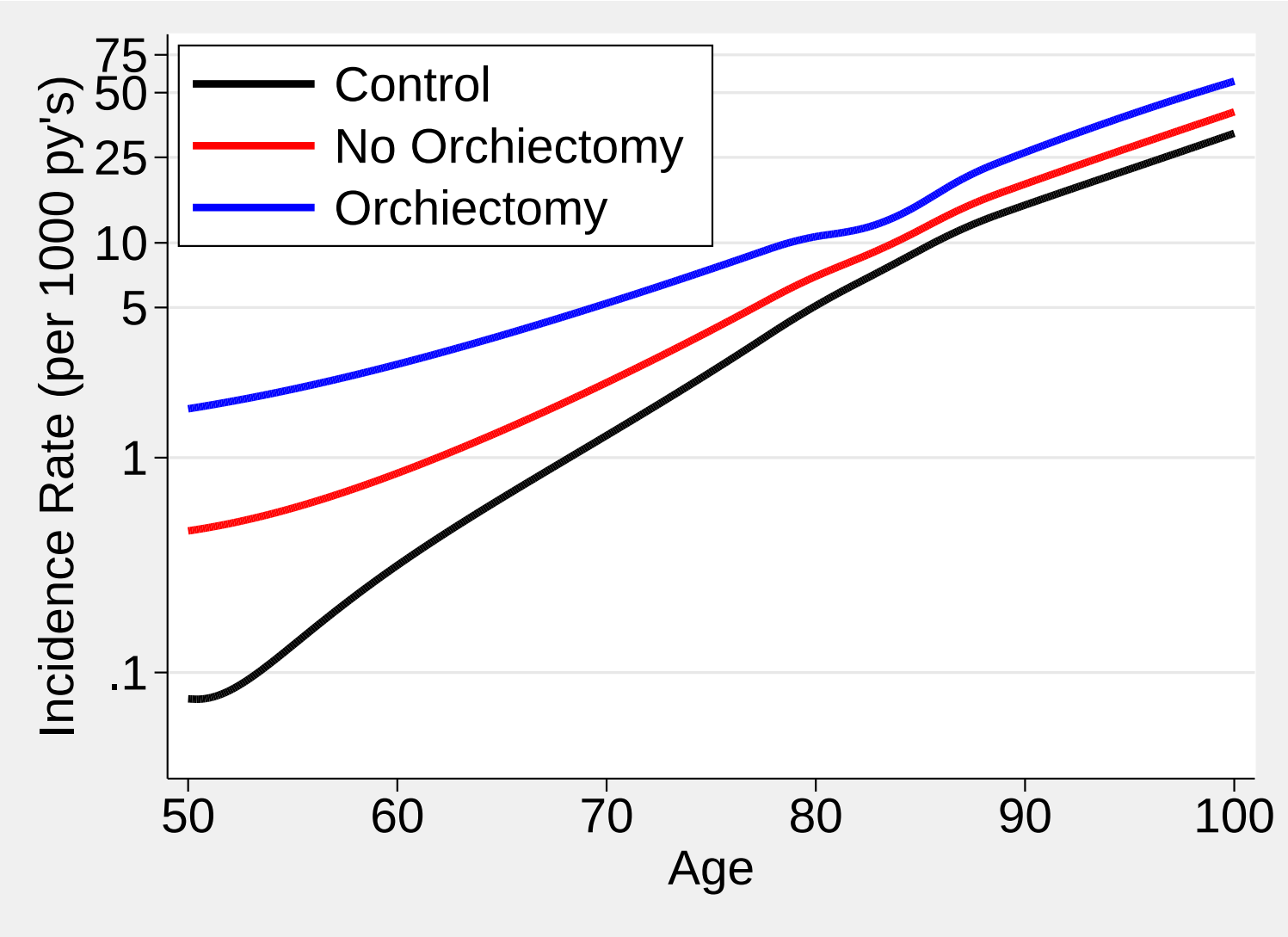
Flexible Parametric Model

Incidence rate ratio (no orchiectomy)	=	1.37 (1.28 to 1.46)
Incidence rate ratio (orchiectomy)	=	2.09 (1.93 to 2.27)

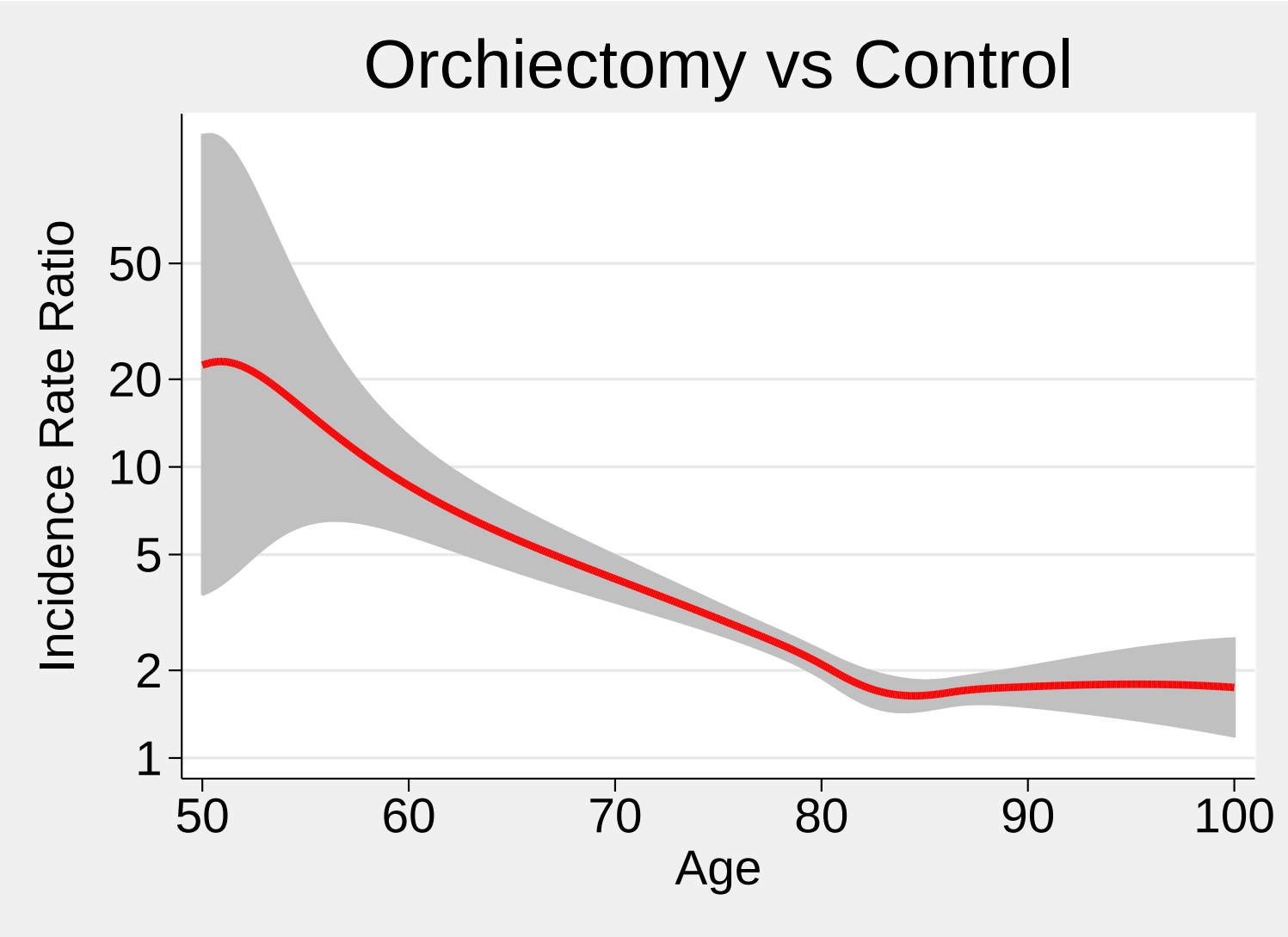
Proportional Hazards



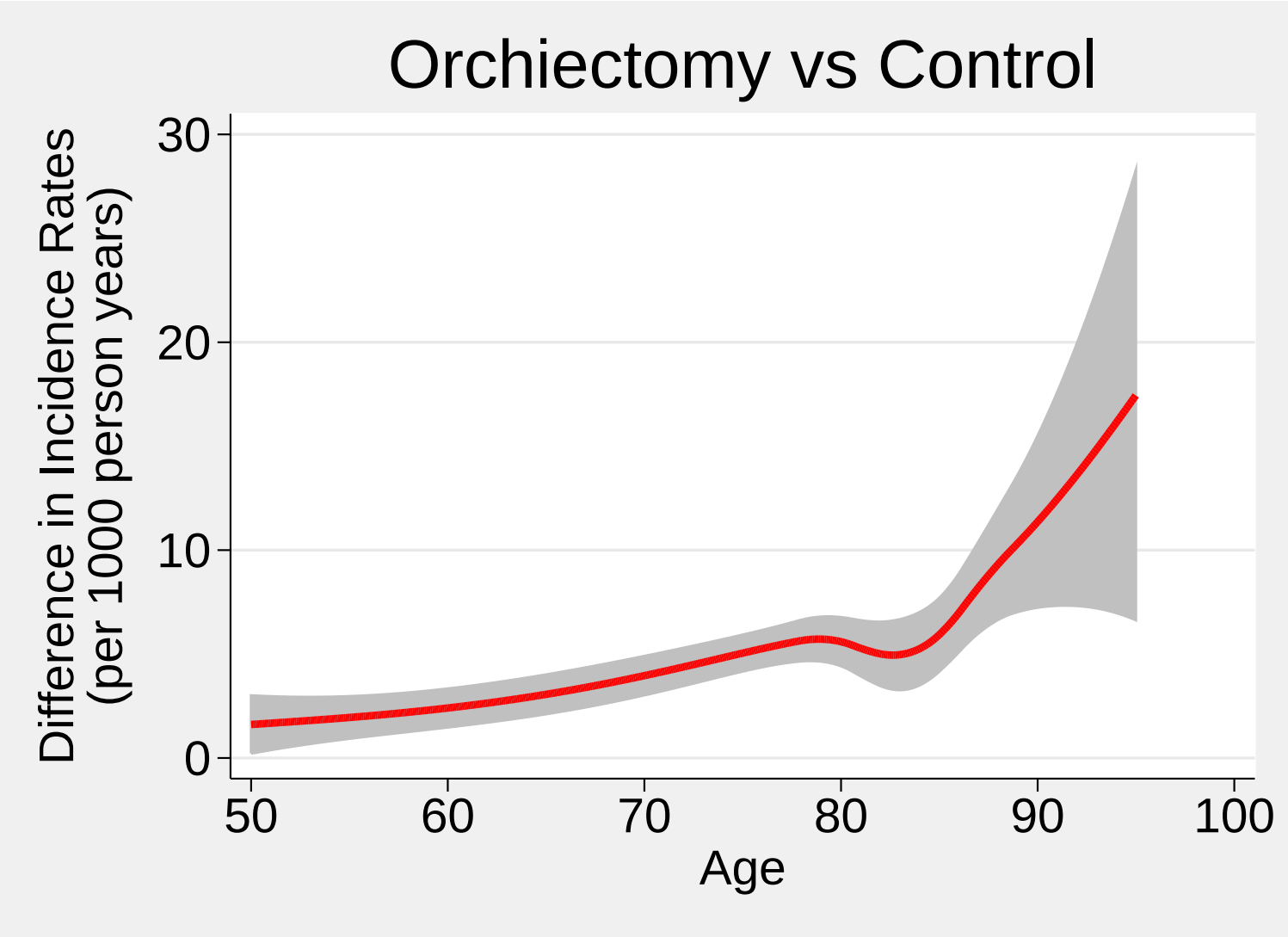
Non Proportional Hazards



Incidence Rate Ratio



Incidence Rate Difference



Multiple time-scales; ongoing research

- Both attained age and time since diagnosis can be modelled simultaneously, i.e., two time-scales[14]. Main time-scale is age.
- Better to use hazard scale.
- Model for PH, but can be extended to time-dependent effects.

$$\ln[h(a|\mathbf{x}_i, a_{0i})] = s\left(a|\boldsymbol{\gamma_0}, \mathbf{k}_0\right) + \mathbf{x}_i\boldsymbol{\beta} + s\left(a - a_{0i}|\boldsymbol{\gamma_1}, \mathbf{k}_1\right)$$

- a_{0i} is age at diagnosis
- Numerical integration required to obtain cumulative hazard for each individual at each iteration.

$$\ln L_i = d_i \ln[h(t_i)] - \int_{t_{0i}}^{t_i} h(u)du$$

- For orchiectomy data (N=423,315) takes 3-4 minutes

Marginal measures and standardisation [5]

International Epidemiological Association

International Journal of Epidemiology, 2020, 1–10
doi: 10.1093/ije/dyz268
Original article

Original article

Marginal measures and causal effects using the relative survival framework

Elisavet Syriopoulou ,[1]* Mark J Rutherford[1] and Paul C Lambert[1,2]

[1]Biostatistics Research Group, Department of Health Sciences, University of Leicester, Leicester, UK and [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

*Corresponding author. Biostatistics Research Group, Department of Health Sciences, University of Leicester University Road, Leicester LE1 7RH, UK. E-mail: e.syriopoulou@leicester.ac.uk

Marginal survival curves: average of individual predictions

- The predicted survival for individual i is

$$\widehat{S}_i(t) = \exp\left(-H_0(t)\exp\left(\beta_1 x_{1i} + \beta_2 x_{2i}\right)\right)$$

- We then average over all predicted survival curves

$$\widehat{S}^P(t) = \frac{1}{N}\sum_{i=1}^{N}\widehat{S}_i(t)$$

- The model can be as complex as required (continuous covariates, interactions, non-linear functions, non-proportional hazards).
- Note that we are predicting a curve, not $S(t)$ evaluated at a single time point.

Standardized survival curves

- When interest lies in comparing the survival of (two) exposure groups we need to standardize to the same covariate distribution.
- Let X be the exposure of interest.
- Let Z denote the set of measured covariates.

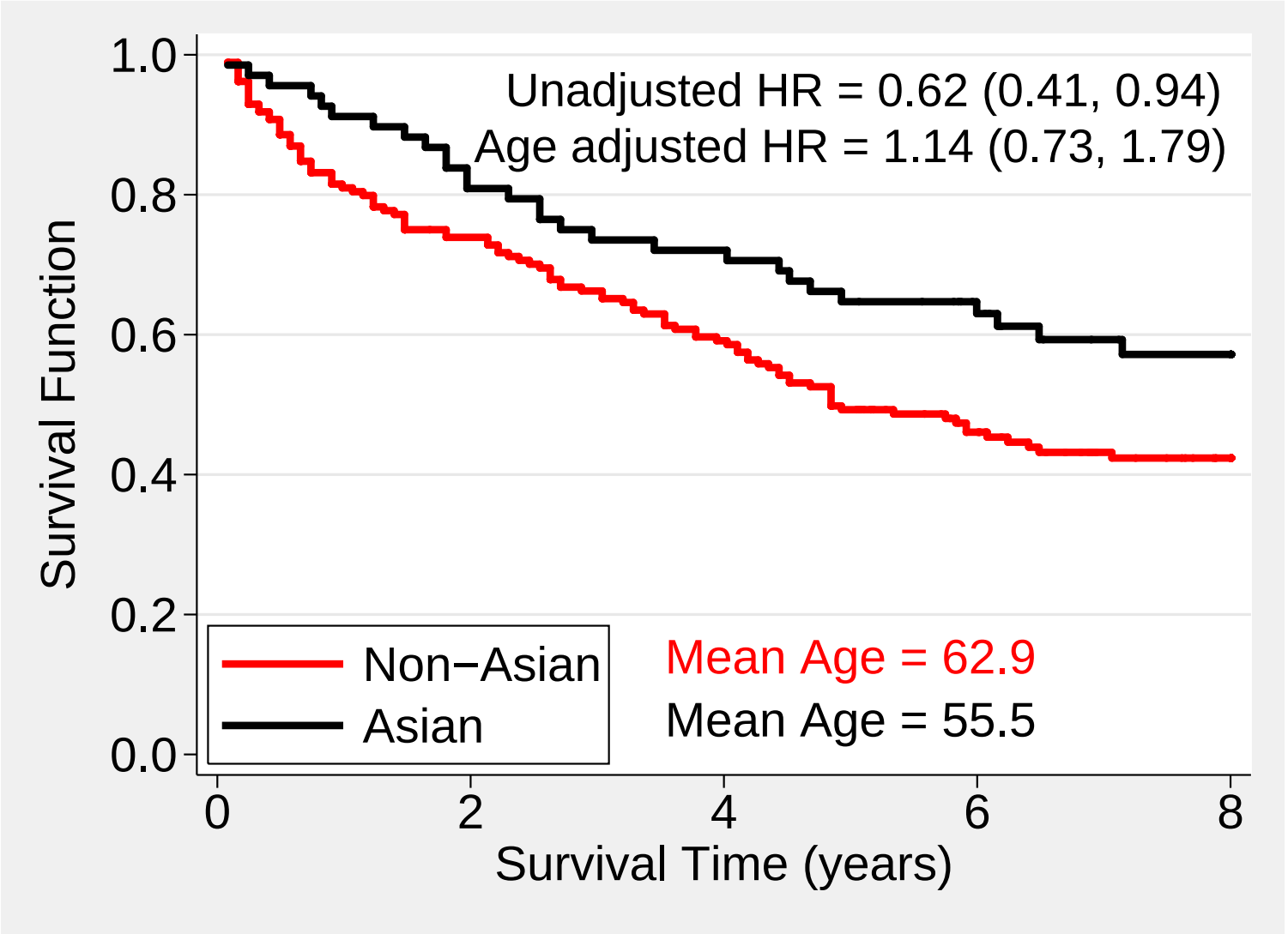
$$\widehat{S}^{P}(t|X=x,Z)=\frac{1}{N}\sum_{i=1}^{N}S_i\left(t|X=x,Z\right)$$

- Note that the average is over the marginal distribution of Z, not over the conditional distribution of Z among those with $X=x$.
- We are forcing the same covariate distribution on both exposure groups.

Example: Renal dialysis

- 252 patients entering a renal dialysis program in Leicestershire, England 1982-1991 with follow-up to the end of 1994.
- Interest in difference in survival by ethnicity (Non-South Asian vs South Asian).
- At the time of the study, approximately 25% of the population were of South Asian origin.

Kaplan-Meier Curves - Renal Replacement Therapy



Predictions for Standardised Survival Curves

The meansurv option

```
stpm2 asian age, df(3) scale(hazard)
/* Age distribution for study population as a whole */
predict meansurv_pop0, meansurv at(asian 0)
predict meansurv_pop1, meansurv at(asian 1)

/* Age distribution for non-asians */
predict meansurv_pop0b if asian == 0, meansurv at(asian 0)
predict meansurv_pop1b if asian == 0, meansurv at(asian 1)

/* Age distribution for asians */
predict meansurv_pop0c if asian == 1, meansurv at(asian 0)
predict meansurv_pop1c if asian == 1, meansurv at(asian 1)
```

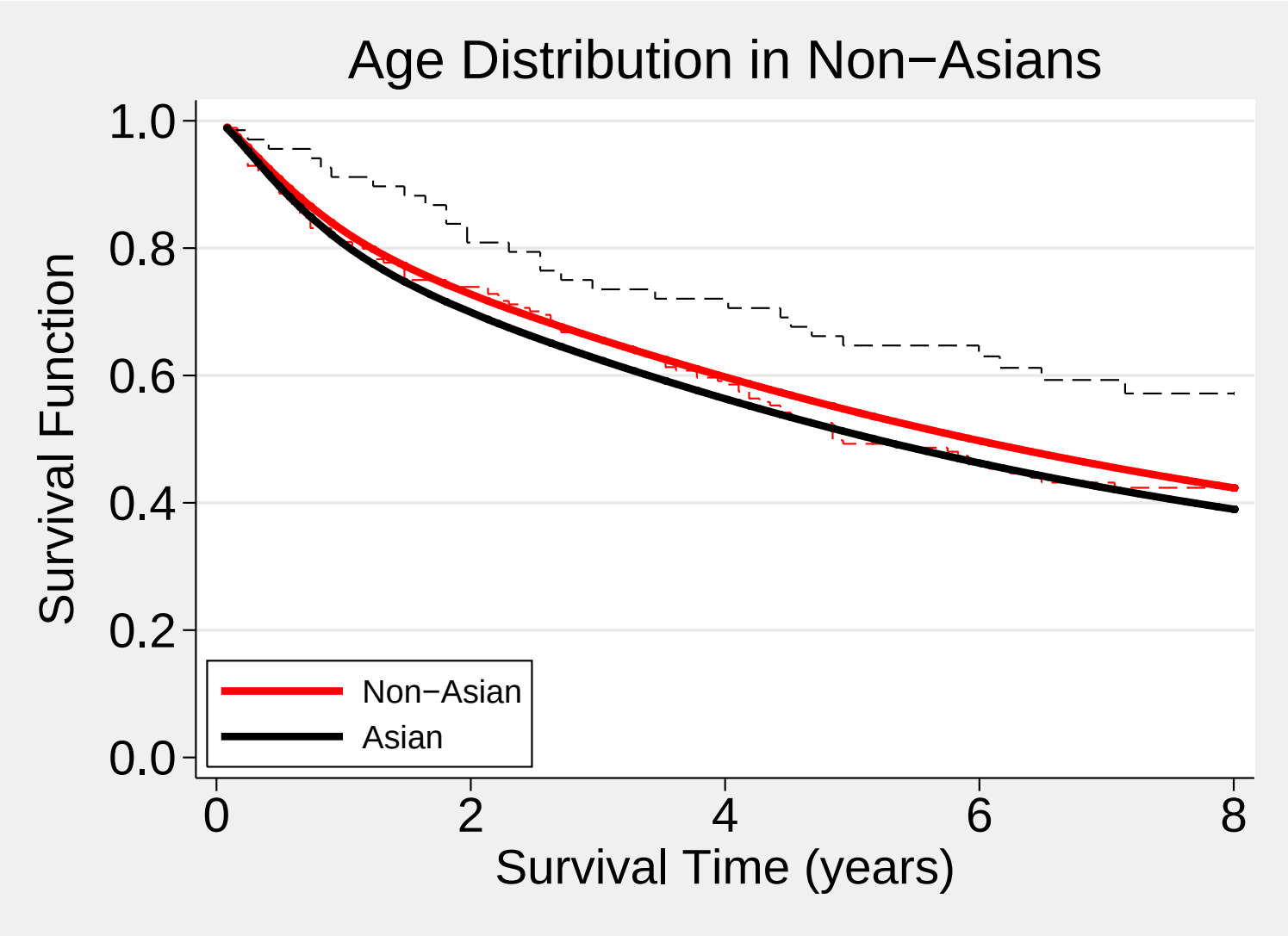
- $S(t)$ calculated for each subject in the study population and averaged.

Predictions for Standardised Survival Curves 2

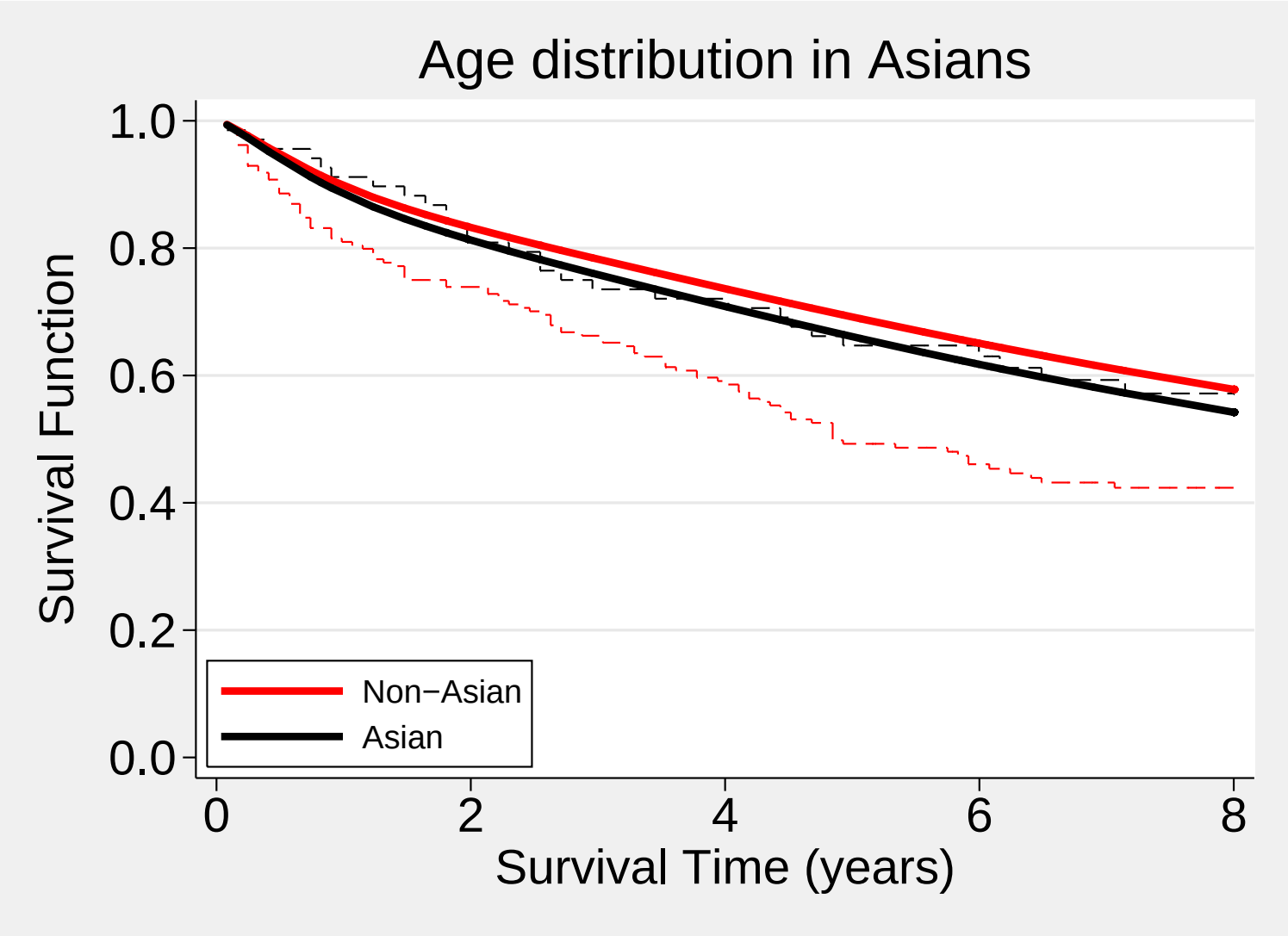
- The adjusted curves show the survival we would expect to see in both groups if each had the age distribution of the study population as a whole.

Standardized Survival Curve 1

Standardized Survival Curve 2



Standardized Survival Curve 3



Difference in standardised survival

- We can estimate the difference in standardised survival,

$$\widehat{\theta}(t,x) = \frac{1}{N}\sum_{i=1}^{N} S(t|X=1,\mathbf{z_i}) - \frac{1}{N}\sum_{i=1}^{N} S(t|X=0,\mathbf{z_i})$$

- If we have controlled for all confounders then this is a causal survival function difference.
- The model can be as complex as we like. It is just as easy to predict survival functions if we have non-linear effects, various interactions, including interactions with time (non proportional hazards). These interactions could include our exposure variable.

Software for regression standardisation

- With Stata `stpm2`, the `meansurv` option to `predict` produces an average of predicted survival curves for each observation.
- `stpm2_standsurv` and `standsurv` (under development) are faster.
- R users can use the stdReg package (Arvid Sjölander).

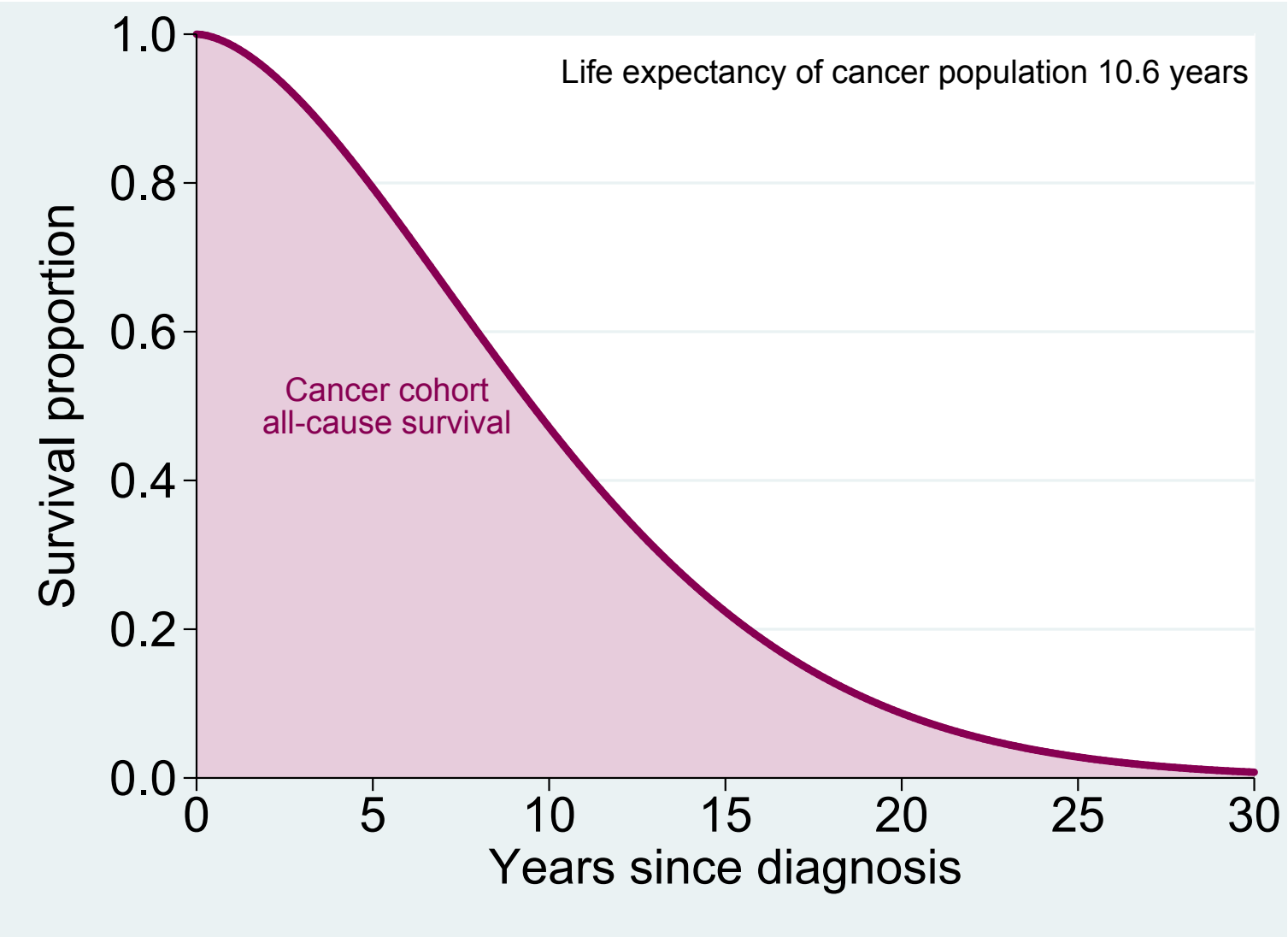
Extrapolation

- We are often warned about the dangers of extrapolation.
- However, sometimes to be able say something useful we need to extrapolate.
- For example, to estimate prevalence of cancer in 2030 we need to extrapolate incidence, survival, changing demographics and potentially changing risk factors [15].
- Common to extrapolate survival to end of life in economic evaluations [16].
  - Often done badly, making simple assumptions.
- Assumptions should be transparent. Good practice to show sensitivity analysis.
- We need parametric methods to extrapolate.

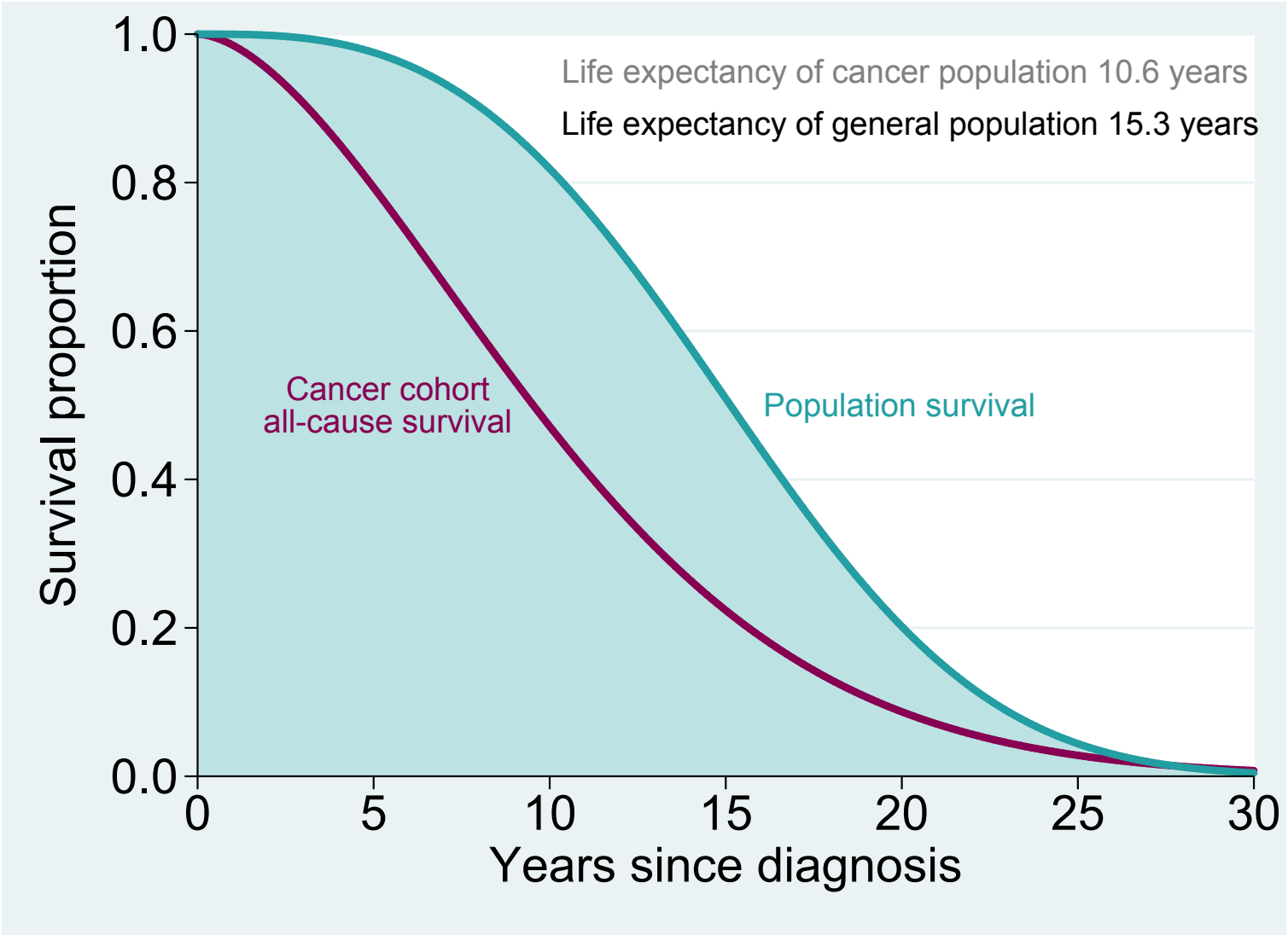
Impact of a cancer diagnosis on life expectancy

- We have promoted a number of alternative metrics in population-based cancer studies. One of which is reduction in live expectancy associated with a diagnosis of cancer.
- Extrapolation to end of life is needed for this.
- We know a lot about how mortality rates vary by demographic factors. We can utilise this external information to help with our extrapolation.

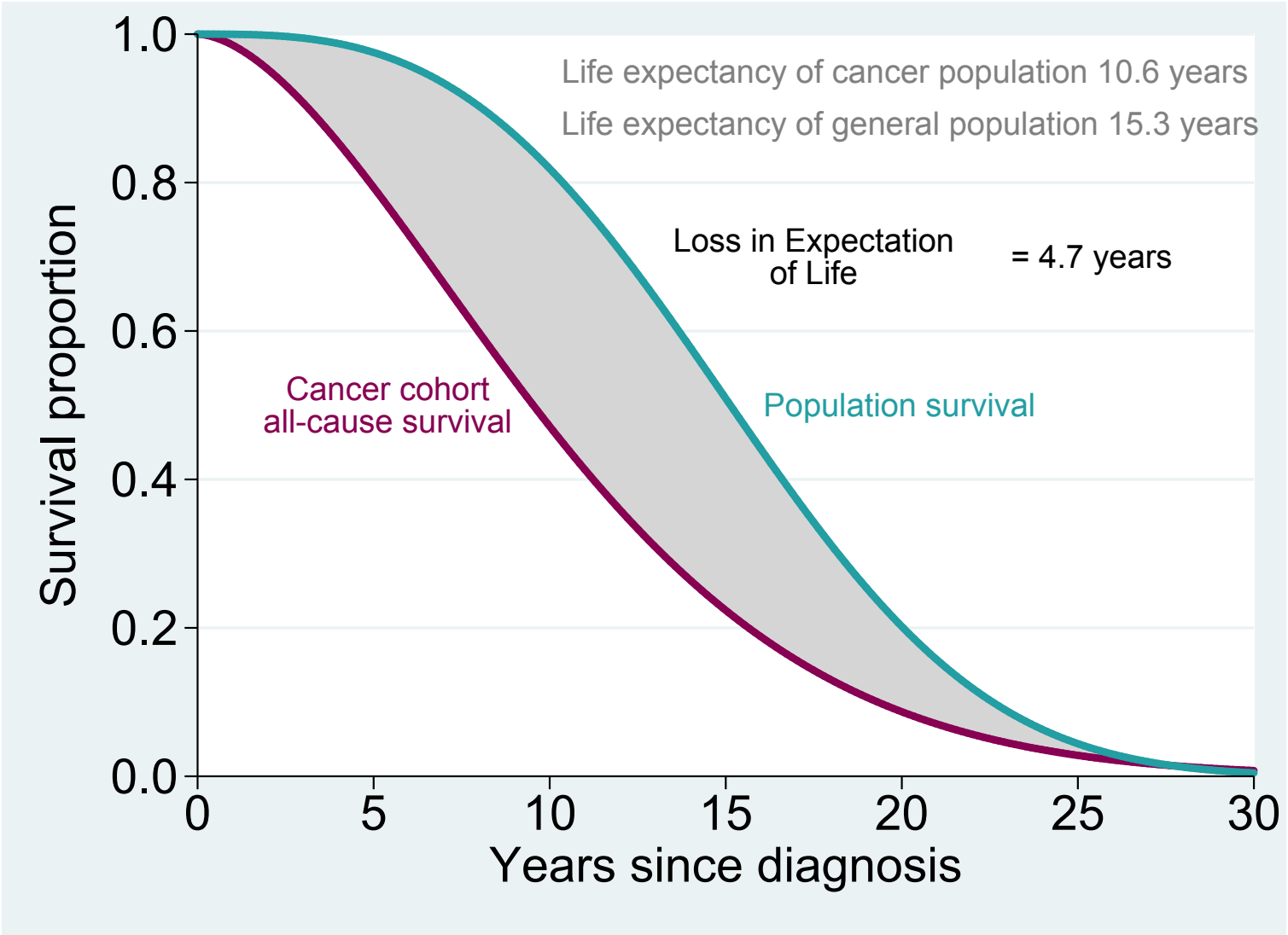
Expectation of life



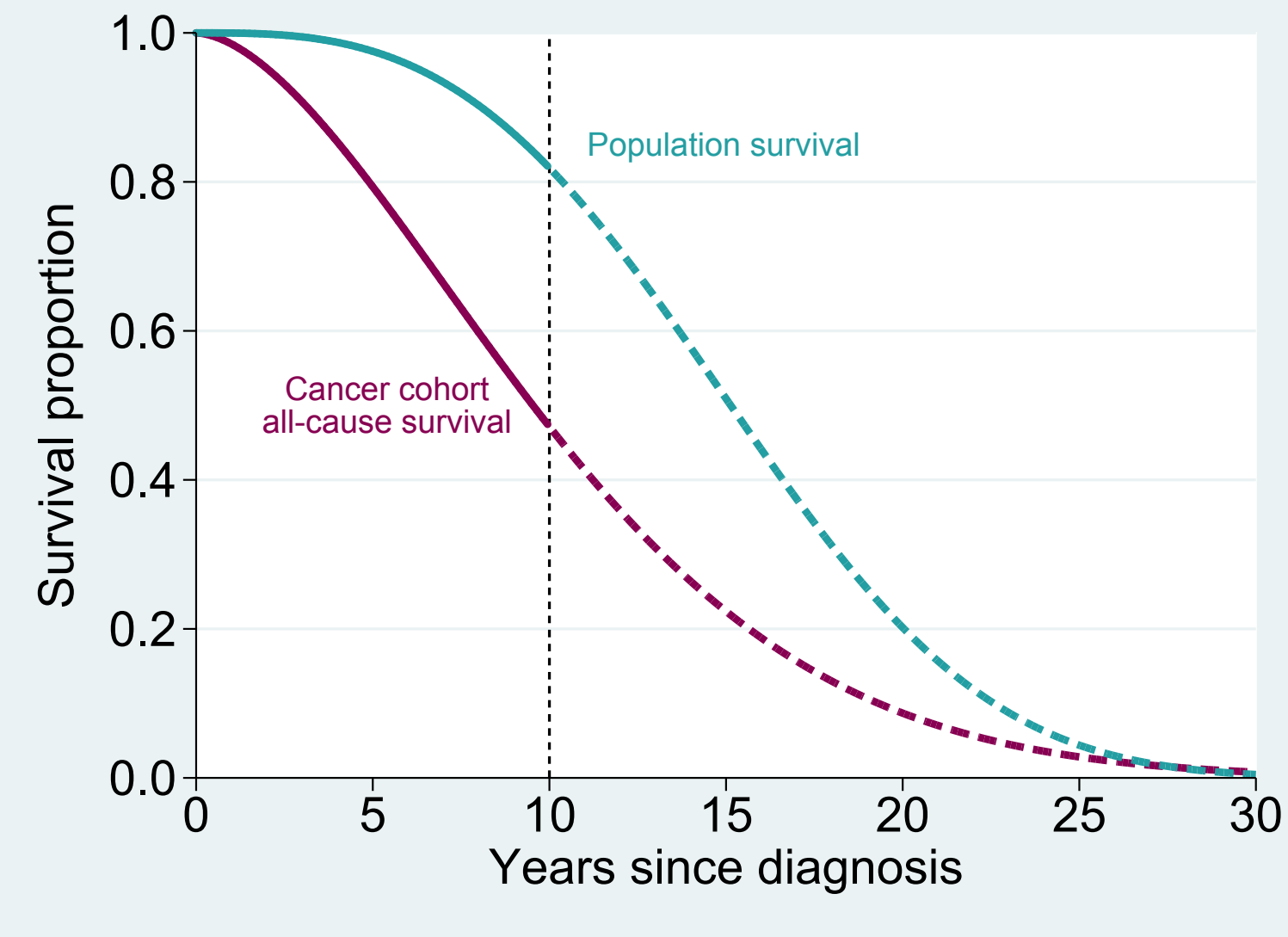
Expectation of life



Loss in expectation of life



Limited follow-up



How do we extrapolate all-cause survival?

Partitioning the mortality rate

- In population based cancer survival we make use of relative survival methods.

$$\textbf{All-cause survival} = \textbf{Expected survival} \times \textbf{Relative survival}$$

$$S(t) = S^*(t)R(t)$$

- The total mortality (hazard) rate is the sum of two components.

All cause Mortality Rate	=	Expected Mortality Rate	+	*Excess Mortality Rate*
$h(t)$	=	$h^*(t)$	+	$\lambda(t)$

- Fairly easy to extrapolate expected survival.
- Can make simple assumptions about excess mortality rate.

Assumptions when we extrapolate

- Simple assumptions about excess mortality when extrapolating.

1. Cure: no excess mortality after a certain point in time[17]
2. Constant excess mortality after a certain point in time
3. Excess mortality estimated from the model (linear with log time)

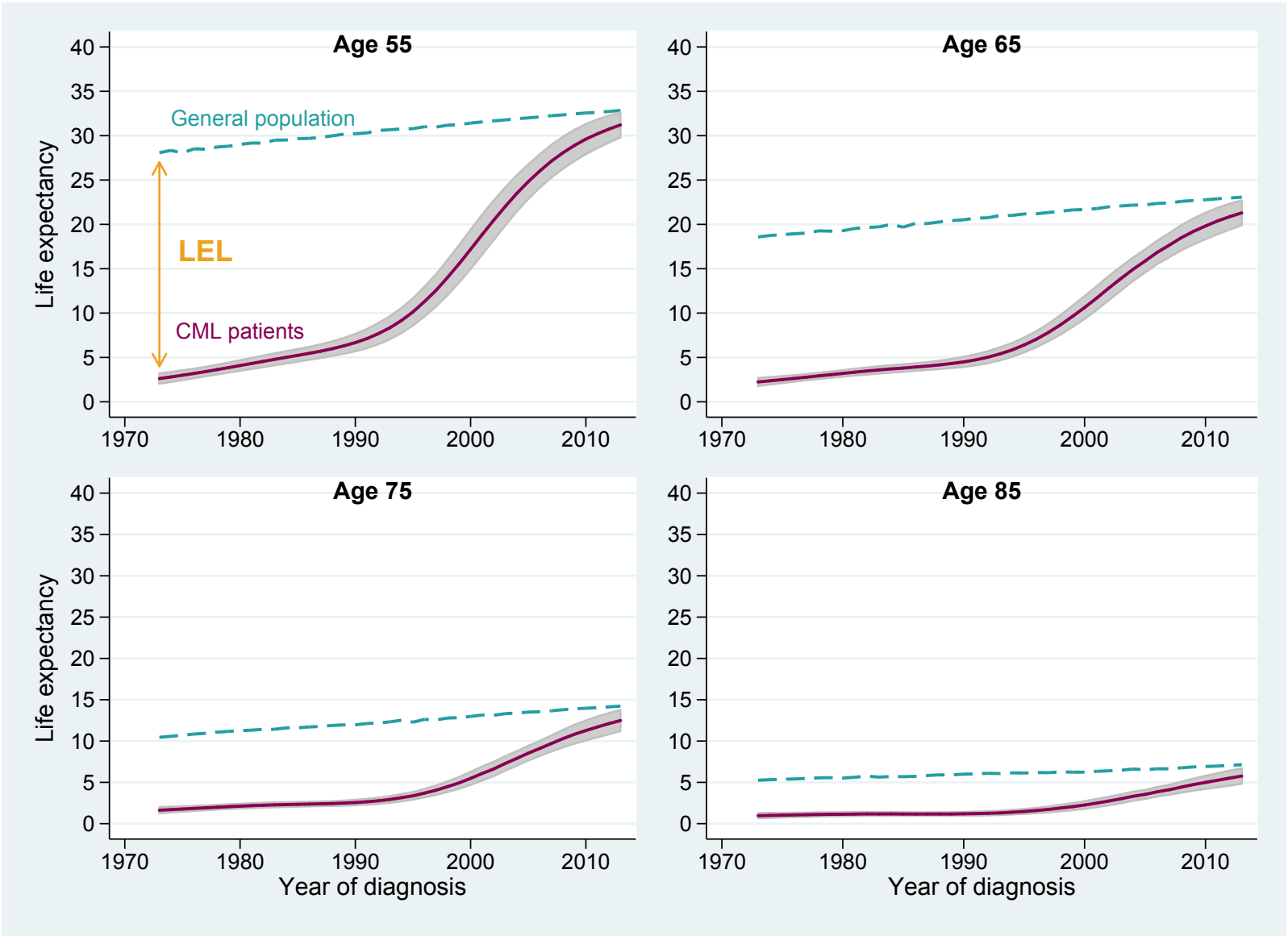
Loss in expectation of life

$$LEL(z) = \int_0^{t_{max}} S^*(t, z')dt - \int_0^{t_{max}} S(t, z)$$

$$LEL(z) = \int_0^{t_{max}} S^*(t, z')dt - \int_0^{t_{max}} S^*(t, z') \times R(t, z)dt$$

- As time since diagnosis increases, the expected mortality rate dominates.

Loss in expectation of life: CML (Sweden)[6]



Even more

- Cure models [17].
- Random effect models [18].
- Joint models [19].
- Multi-state models
- Competing Risks
  - Cause-specific models [20]
  - Direct modelling (subhazards) [21, 21].
- Restricted mean survival time [22].
- Prognostic modelling.

Conclusion

- There is nothing wrong with using a Cox model.
- If you only want to estimate a hazard ratio and that you 'know' you have proportional hazards then a Cox model is ideal.
- Can relax the PH assumption in the Cox model, and can estimate quantities other than HR.
- However, a parametric approach makes it easier to estimate quantities that provide more insight and may be more relevant to your research question.
- You will get the same hazard ratio, but a whole lot more.

References

[1] Weibull CE, Johansson ALV, Eloranta S, Smedby KE, Björkholm M, Lambert PC, *et al.*. Contemporarily treated patients with Hodgkin lymphoma have childbearing potential in line with matched comparators. *Journal of Clinical Oncology* 2018;**36**:2718–2725.

[2] Radkiewicz C, Johansson AL, Dickman PW, Lambe M, Edgren G. Sex differences in cancer risk and survival: A Swedish cohort study. *European Journal of Cancer* 2017;**84**:130–140.

[3] Andreassen BK, Grimsrud TK, Haug ES. Bladder cancer survival: Women better off in the long run. *European Journal of Cancer* 2018;**95**:52–58.

[4] Hernán MA. The hazards of hazard ratios. *Epidemiology* 2010;**21**:13–15.

[5] Syriopoulou E, Rutherford MJ, Lambert PC. Marginal measures and causal effects using the relative survival framework. *International Journal of Epidemiology* 2020;**49**:619–628.

[6] Bower H, Björkholm M, Dickman PW, Höglund M, Lambert PC, Andersson TML. Life expectancy of chronic myeloid leukemia patients is approaching the life expectancy of the general population. *Journal of Clinical Oncology* 2016;**34**:2851–2857.

[7] Cox DR. Regression models and life-tables (with discussion). *JRSSB* 1972;**34**:187–220.

[8] Reid N. A conversation with Sir David Cox. *Statistical Science* 1994;**9**:439–455.

References 2

[9] Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 2002;**21**:2175–2197.

[10] Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *The Stata Journal* 2009;**9**:265–290.

[11] Royston P, Lambert PC. *Flexible parametric survival analysis in Stata: Beyond the Cox model*. Stata Press, 2011.

[12] Rutherford MJ, Hinchliffe SR, Abel GA, Lyratzopoulos G, Lambert PC, Greenberg DC. How much of the deprivation gap in cancer survival can be explained by variation in stage at diagnosis: An example from breast cancer in the East of England. *International Journal of Cancer* 2013;**133**:2192–2200.

[13] Syriopoulou E, Mozumder SI, Rutherford MJ, Lambert PC. Robustness of individual and marginal model-based estimates: A sensitivity analysis of flexible parametric models. *Cancer Epidemiology* 2018;**58**:17–24.

[14] Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. *Stat Med* 2004;**23**:51–64.

[15] Bray F, Møller B. Predicting the future burden of cancer. *Nat Rev Cancer* 2006;**6**:63–74.

References 3

[16] Latimer NR. Survival analysis for economic evaluations alongside clinical trials–extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. *Medical decision making an international journal of the Society for Medical Decision Making* 2013;**33**:743–754.

[17] Andersson TML, Dickman PW, Eloranta S, Lambert PC. Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. *BMC Medical Research Methodology* 2011;**11**:96.

[18] Crowther MJ, Look MP, Riley RD. Multilevel mixed effects parametric survival models using adaptive gauss-hermite quadrature with application to recurrent events and individual participant data meta-analysis. *Stat Med* 2014;**33**:3844–3858.

[19] Crowther MJ, Abrams KR, Lambert PC. Flexible parametric joint modelling of longitudinal and survival data. *Stat Med* 2012;**31**:4456–4471.

[20] Hinchliffe SR, Lambert PC. Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. *BMC Medical Research Methodology* 2013;**13**:13.

[21] Lambert PC, Wilkes SR, Crowther MJ. Flexible parametric modelling of the cause-specific cumulative incidence function. *Statistics in Medicine* 2017;**36**:1429–1446.

References 4

[22] Royston P, Parmar MKB. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC medical research methodology* 2013;**13**:152.