UNIVERSITY OF LEICESTER

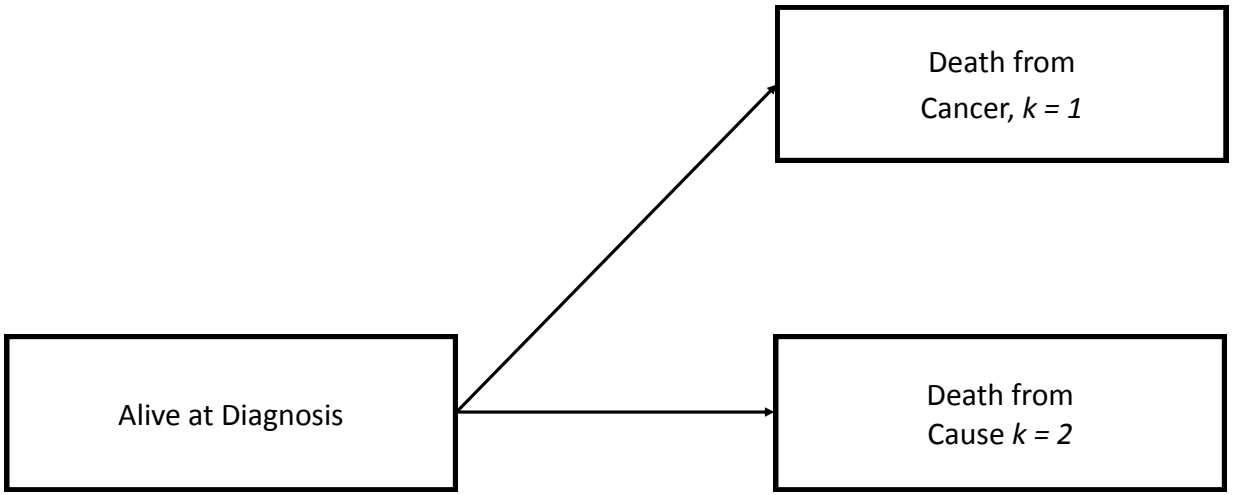
2016 Stata London Users Group Meeting

`stpm2cr`: A Stata module for direct likelihood inference on the cause-specific cumulative incidence function within the flexible parametric modelling framework

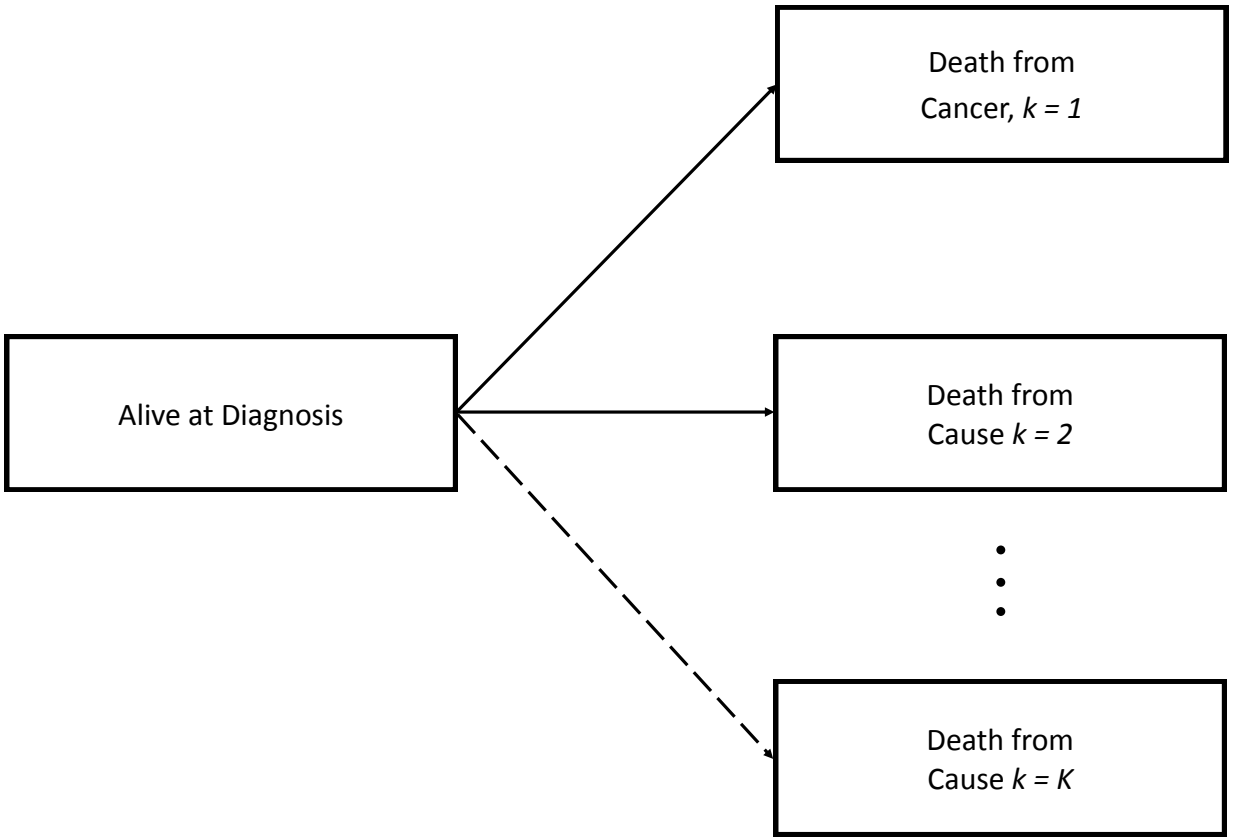
Sarwar Islam Mozumder[1], Mark J Rutherford[1]
& Paul C Lambert[1, 2]

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

Multi-state Model



Multi-state Model



The Cumulative Incidence Function (CIF)

Cause-specific CIF, $F_k(t)$

$$F_k(t) = P(T \leq t, D = k)$$

The Cumulative Incidence Function (CIF)

Cause-specific CIF, $F_k(t)$

The probability that a patient will die from cause $D = k$ by time t whilst also being at risk from dying of other causes

- We obtain this by either,
  1. Estimating using **all** cause-specific hazard functions, or
  2. Transforming using a direct relationship with the subdistribution hazard functions

The Cumulative Incidence Function (CIF)

Cause-specific CIF, $F_k(t)$

The probability that a patient will die from cause $D = k$ by time t whilst also being at risk from dying of other causes

- We obtain this by either,
  1. Estimating using **all** cause-specific hazard functions, or
  2. Transforming using a direct relationship with the subdistribution hazard functions

Approach (1)

Cause-specific Hazards, $h_k^{cs}(t)$

$$h_k^{cs}(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = k | T > t)}{\Delta t}$$

Approach (1)

Cause-specific Hazards, $h_k^{cs}(t)$

Instantaneous conditional rate of mortality from cause $D = k$ given that the patient is still alive at time t

Approach (1)

Cause-specific Hazards, $h_k^{cs}(t)$

Instantaneous conditional rate of mortality from cause $D = k$ given that the patient is still alive at time t

Estimating Cause-specific CIF using CSH

$$F_k(t) = \int_0^t S(u) h_k^{cs}(u) du$$

Approach (1)

Cause-specific Hazards, $h_k^{cs}(t)$

Instantaneous conditional rate of mortality from cause $D = k$ given that the patient is still alive at time t

Estimating Cause-specific CIF using CSH

$$F_k(t) = \int_0^t \exp\left[\int_0^s -\sum_{j=1}^{K} h_j^{cs}(u)du\right] h_k^{cs}(s)ds$$

Approach (2)

Subdistribution Hazards, $h_k^{sd}(t)$

$$h_k^{sd}(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = k | T > t \cup (T \leq t \cap cause \neq k)}{\Delta t}$$

Approach (2)

Subdistribution Hazards, $h_k^{sd}(t)$

The instantaneous rate of failure at time t from cause $D = k$ amongst those who have not died, or have died from any of the other causes, where $D \neq k$

Approach (2)

Subdistribution Hazards, $h_k^{sd}(t)$

The instantaneous rate of failure at time t from cause $D = k$ amongst those who have not died, or have died from any of the other causes, where $D \neq k$

Direct Transformation of the Cause-specific CIF

$$F_k(t) = 1 - \exp\left[-\int_0^t h_k^{sd}(u)du\right]$$

Regression Modelling

SDH Regression Model

$$h_k^{sd}(t|\mathbf{x}) = h_{0,k}^{sd}(t)\exp\left[\mathbf{x}_k\beta_k^{sd}\right]$$

- Subdistribution hazard ratio $= \exp\left[\beta_k^{sd}\right]$
- Association on the effect of a covariate on risk

Why Flexible Parametric Survival Models? [Royston and Lambert, 2011]

- Models baseline (log-cumulative) SDH function using restricted cubic splines

Log-Cumulative SDH Flexible Parametric Model

$$\ln(H_k^{sd}(t|\mathbf{x}_{ik})) = s_k(\ln(t)|\gamma_k, \mathbf{m}_{0k}) + \mathbf{x}_{ik}\beta_k$$

Why Flexible Parametric Survival Models? [Royston and Lambert, 2011]

- Models baseline (log-cumulative) SDH function using restricted cubic splines

Log-Cumulative SDH Flexible Parametric Model

$$\ln(H_k^{sd}(t|\mathbf{x}_{ik})) = s_k(\ln(t)|\gamma_k, \mathbf{m}_{0k}) + \mathbf{x}_{ik}\beta_k$$

- Easy to include time-dependent effects

Relaxing Assumption of Proportionality

$$\ln(H_k^{sd}(t)) = s_k(\ln(t); \gamma_k, \mathbf{m}_{0k}) + \mathbf{x}_k\beta_k + \sum_{l=1}^{E} s_k(\ln(t); \alpha_{lk}, \mathbf{m}_{lk})\mathbf{x}_{lk}$$

- Can predict time-dependent HRs, absolute differences and more...

The Likelihood Function [Jeong and Fine, 2006]

Direct Parametrisation (competing risks)

$$L = \prod_{i=1}^{n} \left[\prod_{j=1}^{K} \left[\left(f_j^s(t_i|\mathbf{x}_j) \right)^{\delta_{ij}} \right] \left[S(t|\mathbf{x}) \right]^{1-\sum_{j=1}^{K} \delta_{ij}} \right]$$

The Likelihood Function [Jeong and Fine, 2006]

Direct Parametrisation (competing risks)

$$L = \prod_{i=1}^{n} \left[\prod_{j=1}^{K} \left[\left(f_j^{s}(t_i|\mathbf{x}_j) \right)^{\delta_{ij}} \right] \left[S(t|\mathbf{x}) \right]^{1-\sum_{j=1}^{K} \delta_{ij}} \right]$$

The Likelihood Function [Jeong and Fine, 2006]

Direct Parametrisation (competing risks)

$$L = \prod_{i=1}^{n} \left[\prod_{j=1}^{K} \left[\left(f_j^s(t_i|\mathbf{x}_j) \right)^{\delta_{ij}} \right] [S(t|\mathbf{x})]^{1-\sum_{j=1}^{K} \delta_{ij}} \right]$$

CSH Approach

$$L = \prod_{i=1}^{n} \left[\prod_{j=1}^{K} \left[\left(S(t|\mathbf{x}) h_j^{cs}(t_i|\mathbf{x}_j) \right)^{\delta_{ij}} \right] [S(t|\mathbf{x})]^{1-\sum_{j=1}^{K} \delta_{ij}} \right]$$

- Estimates covariate effects on the cause-specific CIF rather than the CSH rate

The Likelihood Function [Jeong and Fine, 2006]

Direct Parametrisation (competing risks)

$$L = \prod_{i=1}^{n} \left[\prod_{j=1}^{K} \left[(f_j^s(t_i|\mathbf{x}_j))^{\delta_{ij}} \right] \left[1 - \sum_{j=1}^{K} F_j(t|\mathbf{x}_j) \right]^{1-\sum_{j=1}^{K} \delta_{ij}} \right]$$

Existing Approaches with Implementation in Stata

- `stcrreg`: Fine & Gray model
- `stcrprep`: Restructures the data and calculates appropriate weights [Lambert et al., 2016 (submitted]

Both commands above are computationally intensive due to the requirement of numerical integration and fitting models to an expanded dataset.

Quick Note on stcrreg

```
. stset survmm, failure(cause == 1) scale(12) id(id) exit(time 180)
  (output omitted)
. stcrreg i.stage, compete(cause == 2, 3)
        failure _d:  cause == 1
  analysis time _t:  survmm/12
  exit on or before:  time 180
                id:  id

Iteration 0:   log pseudolikelihood = -27756.886
Iteration 1:   log pseudolikelihood = -27756.795
Iteration 2:   log pseudolikelihood = -27756.795

Competing-risks regression                    No. of obs       =     14,162
                                              No. of subjects  =     14,162
Failure event   : cause == 1                  No. failed       =      3,042
Competing events: cause == 2 3                No. competing    =      4,803
                                              No. censored     =      6,317

                                              Wald chi2(1)     =     880.27
Log pseudolikelihood = -27756.795             Prob > chi2      =     0.0000

                                (Std. Err. adjusted for 14,162 clusters in id)
```

_t	SHR	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
stage						
Regional	3.502575	.1479802	29.67	0.000	3.224223	3.804958

Quick Note on stcrreg

```
. stset survmm, failure(cause == 1, 2, 3) scale(12) id(id) exit(time 180)
  (output omitted)
. gen cause2 = cond(_d==0,0,cause)
. stset survmm, failure(cause2 == 1) scale(12) id(id) exit(time 180)
  (output omitted)
. stcrreg i.stage, compete(cause2 == 2, 3)

        failure _d:  cause2 == 1
  analysis time _t:  survmm/12
  exit on or before:  time 180
                id:  id

Iteration 0:   log pseudolikelihood = -27756.886
Iteration 1:   log pseudolikelihood = -27756.795
Iteration 2:   log pseudolikelihood = -27756.795

Competing-risks regression                    No. of obs       =     14,162
                                              No. of subjects  =     14,162
Failure event   : cause2 == 1                 No. failed       =      3,042
Competing events: cause2 == 2 3               No. competing    =      4,795
                                              No. censored     =      6,325

                                              Wald chi2(1)     =     880.27
Log pseudolikelihood = -27756.795             Prob > chi2      =     0.0000

                                (Std. Err. adjusted for 14,162 clusters in id)
```

_t	SHR	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
stage						
Regional	3.502575	.1479802	29.67	0.000	3.224223	3.804958

The `stpm2cr` Command

- Fit log-cumulative SDH flexible parametric models simultaneously for all cause-specific CIFs
- Uses individual patient data - significant computational time gains
- Initial values obtained using `stcompet`, i.e. the Aalen-Johansen approach, and `reg`

Main Syntax

`stpm2cr` [*equation1*][*equation2*]...[*equationN*] [*if*] [*in*] ,
`events(`*varname*`)` [`censvalue(`#`)` `cause(`*numlist*`)` `level(`#`)`
`alleq` `noorthog` `eform` `oldest` `mlmethod(`*string*`)` `lininit`
maximise_options]

Equation Syntax

The syntax of each equation is:

causename: [*varlist*], scale(*scalename*) [df(#) knots(*numlist*) tvc(*varlist*) dftvc(*df_list*) knotstvc(*numlist*) bknots(*knotslist*) bknotstvc(*numlist*) noconstant cure]

U.S. SEER Colorectal Data

- Survival of males diagnosed with colorectal cancer from 1998 to 2013
- Localised and regional stage at diagnosis and ages 75 to 84 years old (14,215)
- Time to death from:
  1. Colorectal cancer
  2. Heart disease
  3. Other causes

Fitting a Model

```
. stset survmm, failure(cause == 1, 2, 3) scale(12) id(id) exit(time 180)
  (output omitted)
. stpm2cr [colrec_cancer: stage2, scale(hazard) df(5)] ///
>         [other_causes: stage2, scale(hazard) df(5)] ///
>         [heart_disease: stage2, scale(hazard) df(5)] ///
>         , events(cause) cause(1 2 3) cens(0) eform nolog
  (output omitted)
Obtaining Initial Values
Starting to Fit Model
```

Fitting a Model

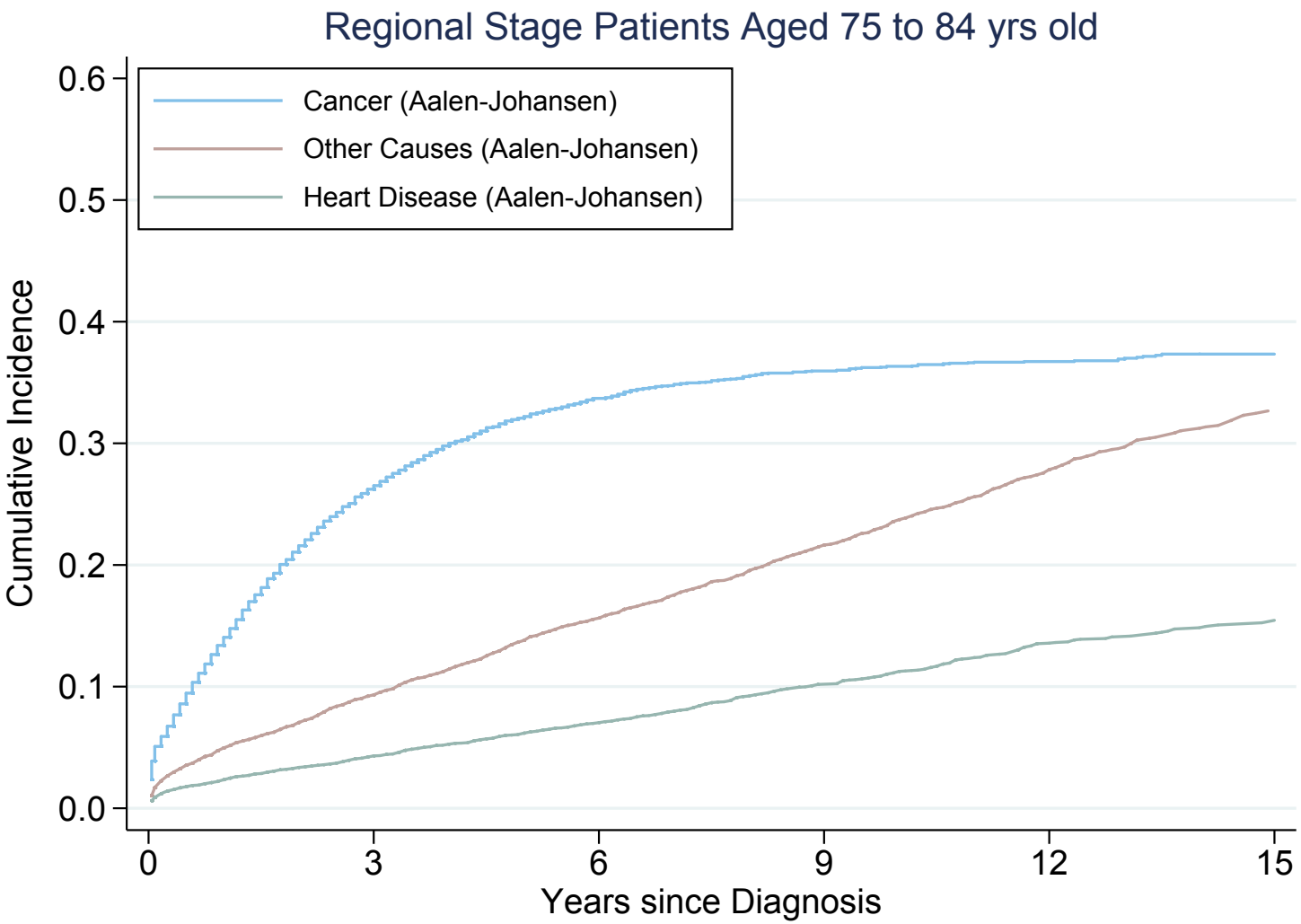
Log likelihood = -26795.633 Number of obs = 14,162

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
colrec_cancer						
stage2	3.429293	.1448444	29.18	0.000	3.156836	3.725265
cr_rcs_c1_1	2.413135	.0384796	55.24	0.000	2.338882	2.489744
cr_rcs_c1_2	1.14997	.0120123	13.38	0.000	1.126665	1.173756
cr_rcs_c1_3	1.029327	.0059401	5.01	0.000	1.01775	1.041035
cr_rcs_c1_4	1.066262	.004378	15.63	0.000	1.057716	1.074877
cr_rcs_c1_5	1.014686	.0030352	4.87	0.000	1.008754	1.020652
_cons	.0672821	.0025488	-71.24	0.000	.0624675	.0724678
other_causes						
stage2	.7203278	.0248887	-9.49	0.000	.673162	.7707984
cr_rcs_c2_1	2.977916	.0637639	50.96	0.000	2.855527	3.10555
cr_rcs_c2_2	.9215077	.0122849	-6.13	0.000	.8977415	.9459031
cr_rcs_c2_3	.9148877	.006712	-12.13	0.000	.9018266	.9281379
cr_rcs_c2_4	1.012339	.0052904	2.35	0.019	1.002023	1.022762
cr_rcs_c2_5	.996456	.0034676	-1.02	0.308	.9896827	1.003276
_cons	.1208134	.0033707	-75.75	0.000	.1143843	.1276038
heart_disease						
stage2	.686007	.0343982	-7.52	0.000	.6217948	.7568504
cr_rcs_c3_1	2.795411	.0817361	35.16	0.000	2.639715	2.960291
cr_rcs_c3_2	.9261574	.016438	-4.32	0.000	.8944935	.9589422
cr_rcs_c3_3	.9187738	.0092581	-8.41	0.000	.9008063	.9370997
cr_rcs_c3_4	.9981656	.0071221	-0.26	0.797	.9843037	1.012223
cr_rcs_c3_5	1.00047	.0047326	0.10	0.921	.9912372	1.009789
_cons	.0578301	.0023139	-71.23	0.000	.0534682	.0625479

stpm2cr Post-estimation

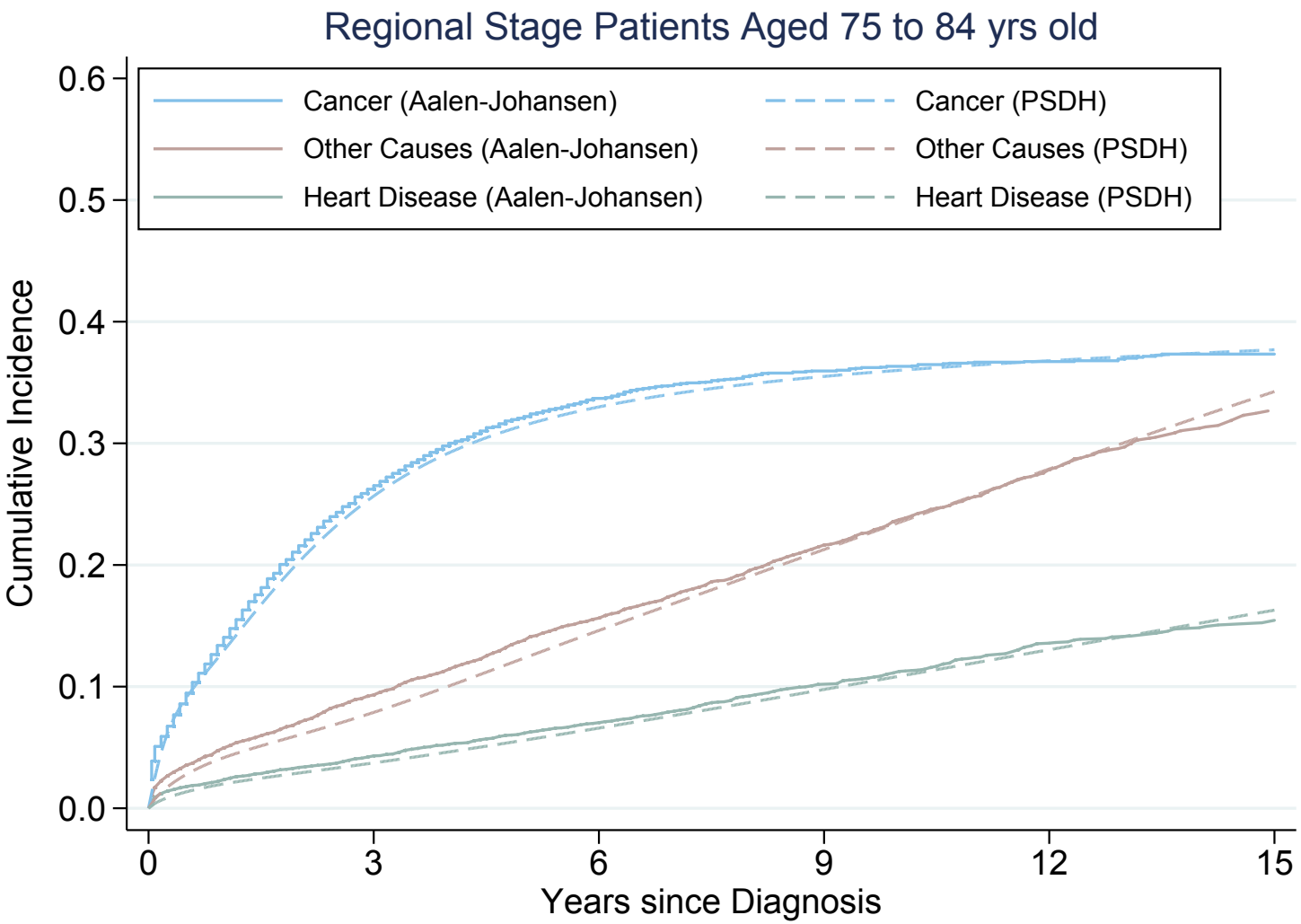
predict *newvarname* [*if*] [*in*] [, at(*varname* # [*varname* #]) cause(*numlist*) chrdenominator(*varname* # [*varname* # ...]) chrnumerator(*varname* # [*varname* # ...]) ci cif cifdiff1(*varname* # [*varname* # ...]) cifdiff2(*varname* # [*varname* # ...]) cifratio csh cumodds cumsubhazard cured shrdenominator(*varname* # [*varname* # ...]) shrnumerator(*varname* # [*varname* # ...]) subdensity subhazard survivor timevar(*varname*) uncured xb zeros deviance dxb level(#)]

Comparison with Aalen-Johansen Estimates



Comparison with Aalen-Johansen Estimates

```
. predict cif_reg, cif at(stage1 0 stage2 1) ci
Calculating predictions for the following causes: 1 2 3
```

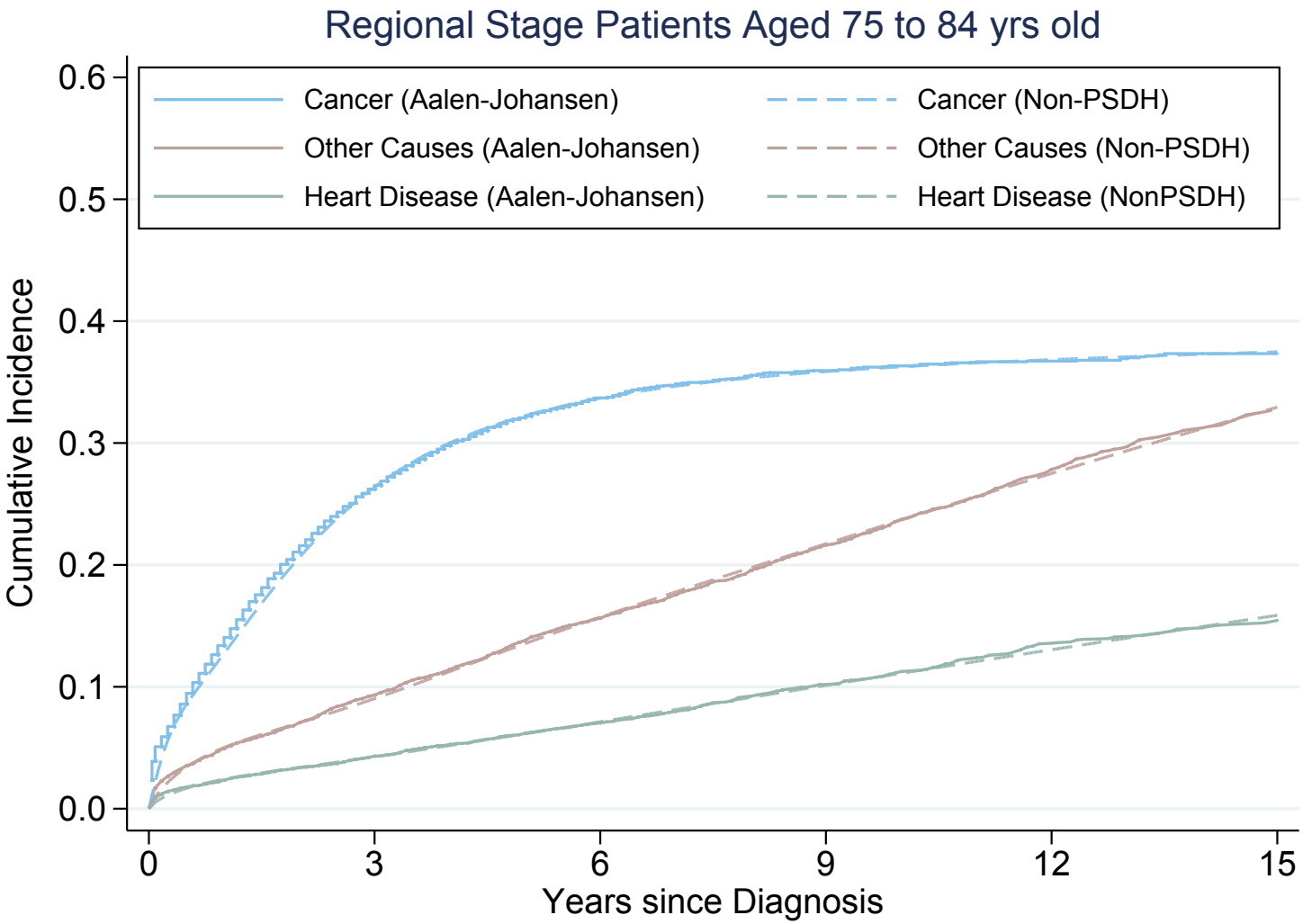


Relaxing the Proportionality Assumption

```
. stpm2cr [colrec_cancer: stage2, scale(hazard) df(5) tvc(stage2) dftvc(3)] ///
>         [other_causes: stage2, scale(hazard) df(5) tvc(stage2) dftvc(3)] ///
>         [heart_disease: stage2, scale(hazard) df(5) tvc(stage2) dftvc(3)] ///
>         , events(cause) cause(1 2 3) cens(0) eform nolog
  (output omitted)
Obtaining Initial Values
Starting to Fit Model
  (output omitted)
```

Comparison with Aalen-Johansen Estimates

```
. predict cif_reg_tvc, cif at(stage1 0 stage2 1) ci
Calculating predictions for the following causes: 1 2 3
```



Relationship with the CSH (Beyersmann & Schumacher, 2007)

$$h_k(t) = \lambda_k(t)\left[1 + \frac{\left[\sum_{j=1}^{K} F_j(t)\right] - F_k(t)}{1 - F(t)}\right]$$

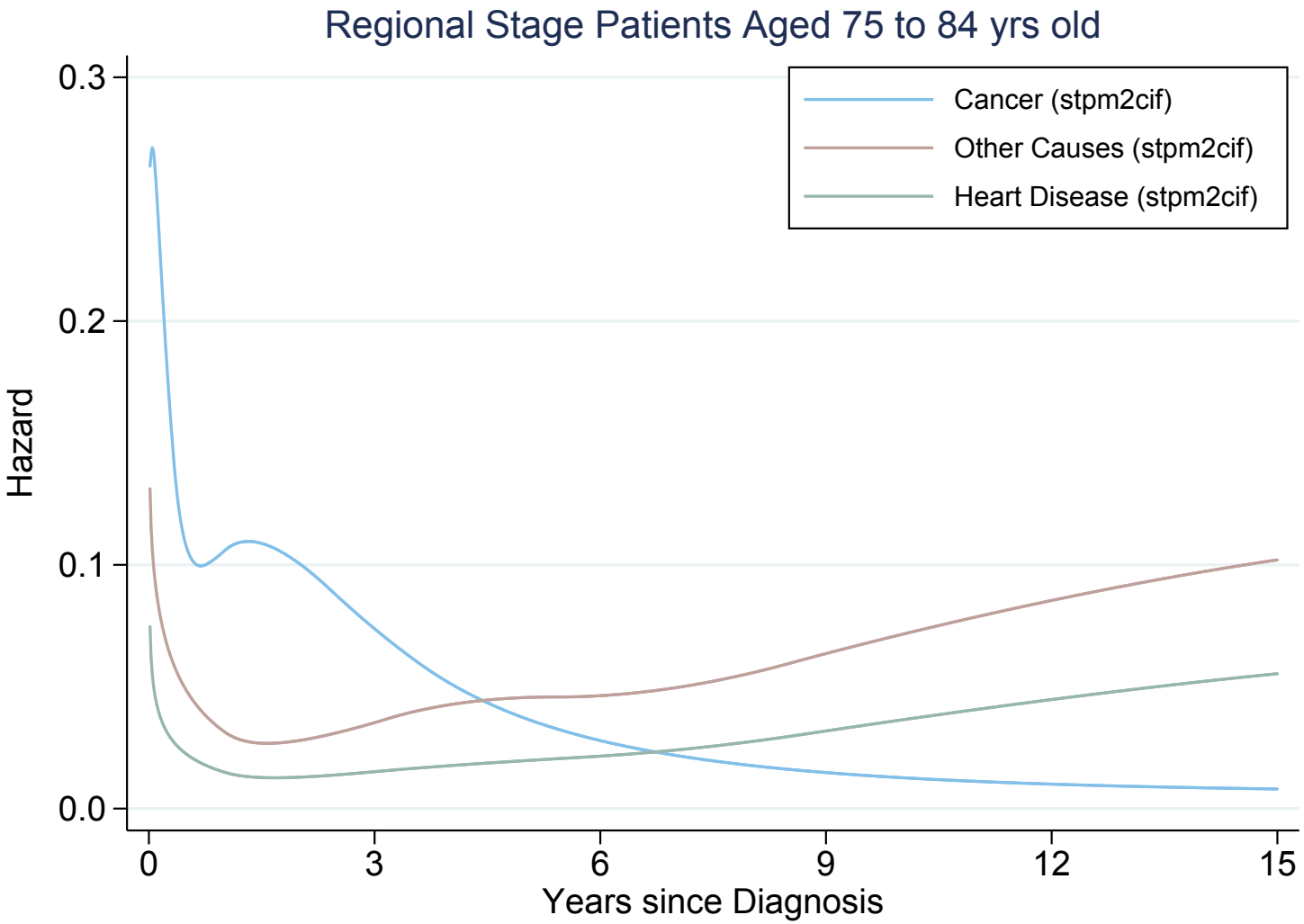
- Can also calculate the CSH from the model
- To calculate from Fine & Gray model, need to fit models for all causes separately (this could take a long, long time)

Relationship with the CSH (Beyersmann & Schumacher, 2007)

$$h_k(t) = \lambda_k(t)\left[1 + \frac{\left[\sum_{j=1}^{K} F_j(t)\right] - F_k(t)}{1 - F(t)}\right]$$

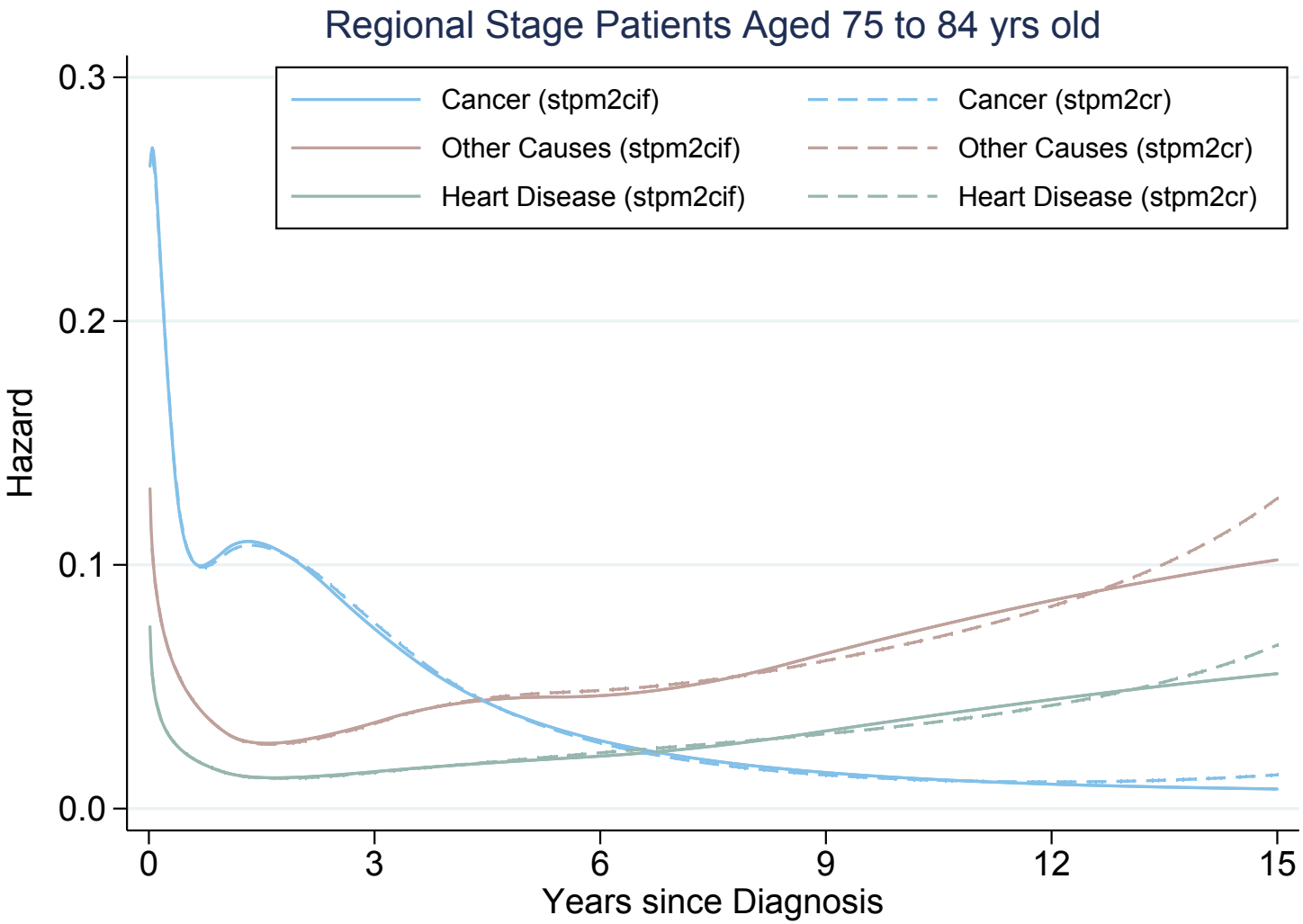
- Can also calculate the CSH from the model
- To calculate from Fine & Gray model, need to fit models for all causes separately (this could take a long, long time)

stpm2cif [Hinchliffe and Lambert, 2013] vs. stpm2cr



stpm2cif [Hinchliffe and Lambert, 2013] vs. stpm2cr

```
. predict csh_reg_tvc, csh at(stage1 0 stage2 1)
Calculating predictions for the following causes: 1 2 3
```



Which way should we model?

- If interest is just on the effect of one cause - no need to model all cause-specific CIFs
- Aetiological = CSH regression models
  - Directly model covariate effects on the hazard rate for those at risk
- Prognostic (decision-making) = SDH regression models
  - Understand why a covariate affects the cause-specific CIF in a certain way
- Make inferences on both scales for a better understanding [Latouche et al., 2013, Beyersmann et al., 2007]
- Advantage of FPMs: Computationally efficient, useful out-of-sample predictions . . .

Selected References I

S. I. Mozumder, M.J. Rutherford, and P.C. Lambert. Direct likelihood inference on the cause-specific cumulative incidence function: a flexible parametric regression modelling approach. *Statistics in Medicine*, 2016 (submitted).

P. Royston and P. C. Lambert. *Flexible parametric survival analysis in Stata: Beyond the Cox model*. Stata Press, 2011.

Jong-Hyeon Jeong and Jason Fine. Direct parametric inference for the cumulative incidence function. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(2):187–200, 2006.

P.C. Lambert, Wilkes S. R., and M.J. Crowther. Flexible parametric modelling of the cause-specific cumulative incidence function. *Statistics in Medicine*, 2016 (submitted).

Sally R Hinchliffe and Paul C Lambert. Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. *BMC medical research methodology*, 13(1):1, 2013.

Aurelien Latouche, Arthur Allignol, Jan Beyersmann, Myriam Labopin, and Jason P Fine. A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *Journal of clinical epidemiology*, 66(6):648–653, 2013.

Jan Beyersmann, Markus Dettenkofer, Hartmut Bertz, and Martin Schumacher. A competing risks analysis of bloodstream infection after stem-cell transplantation using subdistribution hazards and cause-specific hazards. *Statistics in medicine*, 26(30):5360–5369, 2007.

Hein Putter, M Fiocco, and RB Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*, 26(11):2389–2430, 2007.