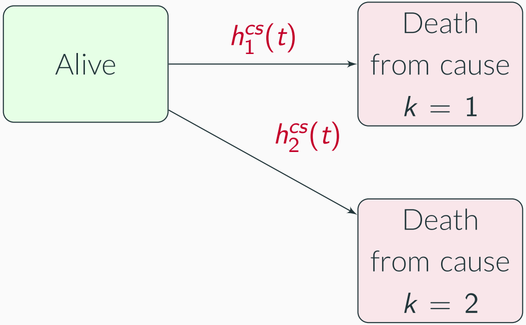
Analysing competing risks data using flexible parametric survival models: what tools are available in Stata, which ones to use and when?

Sarwar Islam Mozumder (sarwar.islam@le.ac.uk)

Biostatistics Research Group, Dept. of Health Sciences, University of Leicester

2018 Nordic and Baltic SUGM | Oslo | 12 September 2018

Competing risks



Cause-specific hazard (CSH) rate, $h_k^{cs}(t)$

Instantaneous mortality (failure) rate from cause k, given that the individual is still alive up to time t

CSH relationship with cause-specific CIF

Cause-specific CIF, $F_k(t)$

Probability a patient will die from cause $D = k$ by time t whilst also being at risk of dying from other competing causes of death

CSH relationship with cause-specific CIF

Cause-specific CIF, $F_k(t)$

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

$$S(t) = \prod_{k=1}^{K} S_k^{cs}(t) = \exp\left(- \sum_{k=1}^{K} \int_0^t h_k^{cs}(u) \mathrm{d}u \right)$$

Approaches for modelling (all) CSHs in Stata

Flexible parametric survival models (FPMs) [Royston and Parmar, 2002]

- Models and more accurately captures complex shapes of the (log-cumulative) baseline hazard function
- A generalisation of the Weibull distribution is used with restricted cubic splines (RCS) that allows for more flexibility

Cause-specific log-cumulative PH FPM

$$\ln\left(H_k^{cs}(t \mid \mathbf{x}_k)\right) = s_k(\ln t; \boldsymbol{\gamma}_k, \mathbf{m}_{0k}) + \beta_k^{cs}\mathbf{x}_k$$

$s_k(\ln t; \boldsymbol{\gamma}_k, \mathbf{m}_{0k})$: baseline restricted cubic spline function on log-time

Flexible parametric survival models (FPMs) [Royston and Parmar, 2002]

- Models and more accurately captures complex shapes of the (log-cumulative) baseline hazard function
- A generalisation of the Weibull distribution is used with restricted cubic splines (RCS) that allows for more flexibility
- Can also easily include time-dependent effects (TDE)

Cause-specific log-cumulative non-PH FPM

$$\ln\left(H_k^{cs}(t \mid \mathbf{x}_k)\right) = s_k(\ln t; \gamma_k, \mathbf{m}_{0k}) + \beta_k^{cs}\mathbf{x}_k + \sum_{l=1}^{E} s_k(\ln t; \boldsymbol{\alpha}_{lk}, \mathbf{m}_{lk})\mathbf{x}_{lk}$$

$s_k(\ln t; \boldsymbol{\alpha}_{lk}, \mathbf{m}_{lk})\mathbf{x}_{lk}$: interaction between spline variables and covariates for TDEs

Example dataset

Load public-use prostate cancer dataset:

```
. use "http://www.stata-journal.com/software/sj4-2/st0059/prostatecancer", clear

. tab status
```

status	Freq.	Percent	Cum.
Censor	150	29.64	29.64
Cancer	155	30.63	60.28
CVD	141	27.87	88.14
Other	60	11.86	100.00
Total	506	100.00	

stpm2 [Lambert and Royston, 2009]

```
. stset time, failure(status == 1) id(id) scale(12) exit(time 60)

. stpm2 treatment, scale(hazard) df(4) eform nolog

Log likelihood =   -440.316                    Number of obs     =        506
```

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
treatment	.6594084	.111509	-2.46	0.014	.4733827	.9185368
_rcs1	3.389716	.4258797	9.72	0.000	2.649838	4.336179
_rcs2	.8879662	.0724157	-1.46	0.145	.7567963	1.041871
_rcs3	1.06315	.0411503	1.58	0.114	.9854806	1.146942
_rcs4	1.016818	.0199075	0.85	0.394	.9785387	1.056594
_cons	.229559	.0272468	-12.40	0.000	.1819129	.2896844

```
Note: Estimates are transformed only in the first equation.

. stcox treatment, nolog noshow
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
treatment	.6602897	.1116672	-2.45	0.014	.4740025	.9197894

stpm2 [Lambert and Royston, 2009]

```
. stset time, failure(status == 2) id(id) scale(12) exit(time 60)

. stpm2 treatment, scale(hazard) df(4) eform nolog

Log likelihood = -448.73758                    Number of obs     =        506
```

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
treatment	1.202808	.2047249	1.08	0.278	.8616223	1.679097
_rcs1	2.82908	.2642265	11.13	0.000	2.355841	3.397384
_rcs2	.8685486	.0544436	-2.25	0.025	.7681357	.9820878
_rcs3	.9529595	.0319403	-1.44	0.151	.8923696	1.017663
_rcs4	1.027927	.0213538	1.33	0.185	.986915	1.070644
_cons	.17767	.0237024	-12.95	0.000	.1367912	.2307651

```
Note: Estimates are transformed only in the first equation.

. stcox treatment, nolog noshow
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
treatment	1.20334	.2048509	1.09	0.277	.8619538	1.679937

stpm2 [Lambert and Royston, 2009]

```
. stset time, failure(status == 3) id(id) scale(12) exit(time 60)

. stpm2 treatment, scale(hazard) df(4) eform nolog

Log likelihood = -231.45608                    Number of obs     =        506
```

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
treatment	.6432149	.1737196	-1.63	0.102	.3788467	1.092066
_rcs1	2.638735	.3351586	7.64	0.000	2.057219	3.384628
_rcs2	.7913665	.0590788	-3.13	0.002	.683647	.9160589
_rcs3	.9369818	.0467358	-1.30	0.192	.8497164	1.033209
_rcs4	1.029843	.031817	0.95	0.341	.9693337	1.09413
_cons	.097687	.0179093	-12.69	0.000	.0681998	.1399235

```
Note: Estimates are transformed only in the first equation.

. stcox treatment, nolog noshow
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
treatment	.6460519	.1745103	-1.62	0.106	.3804893	1.096964

Estimating cause-specific CIFs after fitting FPMs

Cause-specific CIF, $F_k(t)$

$$F_k(t) = \int_0^t \exp\left(-\sum_{k=1}^{K}\int_0^t h_k^{cs}(u)\mathrm{d}u\right) h_k^{cs}(u)\mathrm{d}u$$

Estimating cause-specific CIFs after fitting FPMs

Cause-specific CIF, $F_k(t)$

$$F_k(t) = \int_0^t \exp\left(-\sum_{k=1}^{K}\int_0^t h_k^{cs}(u)\mathrm{d}u\right) h_k^{cs}(u)\mathrm{d}u$$

Must be obtained by numerical approximation:

- Trapezoid method - `stpm2cif` [Hinchliffe and Lambert, 2013]
- Gauss-Legendre quadrature - `stpm2cr` [Mozumder et al., 2017]

stpm2cif: Data setup

```
. expand 3 // augment data k = 3 times

. bysort id: gen _cause=_n

. //create dummy variables for each cause of death
. gen _cvd=_cause==2

. gen _other=_cause==3

. gen _cancer=_cause==1

. //create cause of death event indicator variable
. gen _event=(_cause==status)

. label values _cause status

. foreach cause in _cancer _cvd _other {
  2.          gen treatment`cause´ = treatment*`cause´
  3. }
```

stpm2cif: Data setup

```
. list id status time treatment _cause _event in 1/9, sep(9)
```

	id	status	time	treatm~t	_cause	_event
1.	1	Censor	72	0	1	0
2.	1	Censor	72	0	2	0
3.	1	Censor	72	0	3	0
4.	2	Cancer	1	0	1	1
5.	2	Cancer	1	0	2	0
6.	2	Cancer	1	0	3	0
7.	3	CVD	40	1	1	0
8.	3	CVD	40	1	2	1
9.	3	CVD	40	1	3	0

stpm2cif: Data setup

```
. local knotstvc_opt

. local bknotstvc_opt

. local k = 1

. foreach cause in _cancer _cvd _other {
  2.         stset time, failure(status == `k´) exit(time 60) scale(12)
  3.         cap stpm2 treatment, df(4) scale(h) eform nolog
  4.         estimates store stpm2`cause´
  5.         local bhknots`cause´ `e(bhknots)´
  6.         local boundknots`cause´ `e(boundary_knots)´
  7.         local knotstvc_opt `knotstvc_opt´ `cause´ `bhknots`cause´´
  8.         local bknotstvc_opt `bknotstvc_opt´ `cause´ `boundknots`cause´´
  9.         local k = `k´ + 1
 10. }
```

stpm2cif: Fitting the model

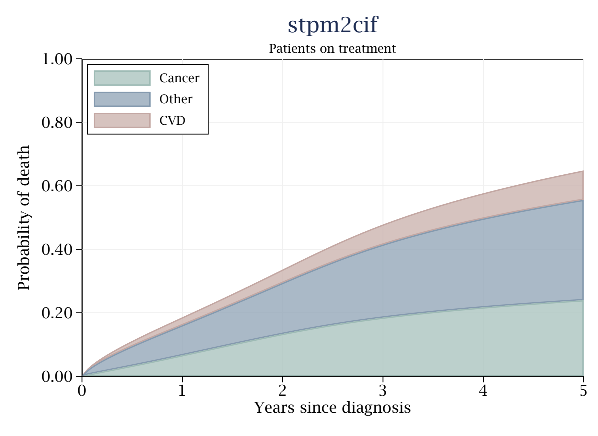
```
. stset time, failure(_event == 1) exit(time 60) scale(12)

. stpm2 treatment_cancer _cancer treatment_cvd _cvd treatment_other _other ///
> , scale(h) knotstvc(`knotstvc_opt´) bknotstvc(`bknotstvc_opt´) ///
> tvc(_cancer _cvd _other) rcsbaseoff nocons eform nolog

Log likelihood = -1120.5192                    Number of obs     =      1,518
```

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
treatment_cancer	.6593781	.111504	-2.46	0.014	.4733607	.9184951
_cancer	.2295677	.0272475	-12.40	0.000	.1819204	.2896945
treatment_cvd	1.202808	.2047249	1.08	0.278	.8616223	1.679097
_cvd	.17767	.0237024	-12.95	0.000	.1367912	.2307651
treatment_other	.6432149	.1737196	-1.63	0.102	.3788467	1.092066
_other	.097687	.0179093	-12.69	0.000	.0681998	.1399235
(*output omitted*)						

```
Note: Estimates are transformed only in the first equation.
```

stpm2cif: Post-estimation

```
. stpm2cif cancer cvd other, cause1(treatment_cancer 1 _cancer 1) ///
> cause2(treatment_cvd 1 _cvd 1) cause3(treatment_other 1 _other 1) ci
```

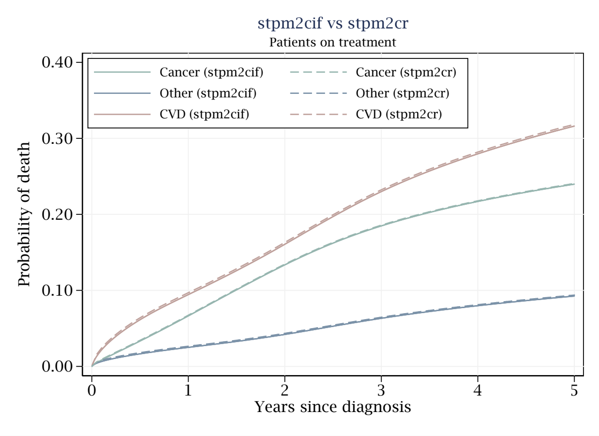


stpm2cr

```
. stset time, failure(status == 1,2,3) exit(time 60) scale(12)

. stpm2cr [cancer: treatment, scale(hazard) df(4)] ///
> [cvd: treatment, scale(hazard) df(4)] ///
> [other: treatment, scale(hazard) df(4)], ///
> events(status) cause(1 2 3) cens(0) eform model(csh)
```

stpm2cr: Post-estimation

```
. range newt 0 5 100

. predict cifgq_trt1, cif at(treatment 1) timevar(newt) ci
```



Note on computational time

```
. expand 500 //now 253,000 observations

. replace time = time + runiform()*0.0001

. replace id = _n
variable id was int now long
```

	Time (secs)
stpm2cr model	52.60
stpm2 (stacked data)	76.59
stpm2cr predict (w/ CIs)	2.56
stpm2cif (w/ CIs)	11.10

multistate [Crowther and Lambert, 2017]

- Written mainly by Michael (& Paul) for more complex multi-state models e.g. illness-death models
- Competing risks is a special case of multi-state models
- Can use `multistate` package to obtain equivalent non-parametric estimates and fit parametric models in presence of competing risks
- Uses a simulation approach for calculating transition probabilities i.e. cause-specific CIFs

Summary of FPM tools for estimating cause-specific CIFs using CSHs

- Post-estimation command, `stpm2cif`
  - Requires augmenting data before `stpm2`
  - Fitting a single model means interpretation is difficult and more room for errors
  - Uses a basic numerical integration method - slow for larger datasets

Summary of FPM tools for estimating cause-specific CIFs using CSHs

- Post-estimation command, `stpm2cif`
  - Requires augmenting data before `stpm2`
  - Fitting a single model means interpretation is difficult and more room for errors
  - Uses a basic numerical integration method - slow for larger datasets
- Using `stpm2cr` as a wrapper followed by `predict`
  - Fits separate `stpm2` models for each cause of death without data augmentation
  - Uses quicker numerical integration method
  - Can obtain other useful predictions e.g. restricted mean lifetime/comparative predictions

Summary of FPM tools for estimating cause-specific CIFs using CSHs

- Via the `predictms` command provided as a part of the `multistate` package
  - Uses a simulation approach. Can alternatively use AJ estimator to save on computational time
  - Can also be used without requiring `msset`
  - Extremely versatile - has some very useful features and post-estimation options

What about modelling covariate effects on the risk of dying from a particular cause?

Subdistribution hazards

$$\text{Alive } \textit{or} \text{ Death from cause } k = 2 \xrightarrow{h_1^{sd}(t)} \text{Death from cause } k = 1$$

$$\text{Alive } \textit{or} \text{ Death from cause } k = 1 \xrightarrow{h_2^{sd}(t)} \text{Death from cause } k = 2$$

Subdistribution hazard (SDH) rate, $h_k^{sd}(t)$

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 $D \neq k$

SDH relationship with cause-specific CIF

Cause-specific CIF, $F_k(t)$

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

SDH relationship with cause-specific CIF

Cause-specific CIF, $F_k(t)$

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

Note

$$1 - F_k(t) = P(D \neq k) + S_k^{sd}(t) \neq S_k^{cs}(t)$$

FPMs on (log-cumulative) SDH scale

Log-cumulative SDH FPM

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

1. Apply time-dependent censoring weights to the likelihood function for each cause k (`stcrprep`) [Lambert et al., 2017]
2. Model all k causes of death simultaneously directly using the full likelihood function (`stpm2cr`) [Mozumder et al., 2017; Jeong and Fine, 2007]

Time-dependent censoring weights

- Need to consider those who have already died from other competing causes of death in risk-set
- Calculate missing censoring times for those that died from other causes by applying time-dependent weights to likelihood
- Influence of weights decreases over-time as the probability of being censored increases
- Further details given by Lambert et al. [2017] and Geskus [2011]

stcrprep

```
. stset time, failure(status == 1,2,3) exit(time 60) scale(12) id(id)
. gen cod2 = cond(_d==0,0,status)
. stcrprep, events(cod2) keep(treatment ) trans(1 2 3) wtstpm2 censcov(treatment) every(1)
. gen event = cod2 == failcode
. stset tstop [iw=weight_c], failure(event) enter(tstart) noshow
```

(*output omitted*)

stcrprep

```
. stpm2 treatment_cancer _cancer treatment_cvd _cvd treatment_other _other ///
> , scale(h) knotstvc(`knotstvc_opt´) bknotstvc(`bknotstvc_opt´) ///
> tvc(_cancer _cvd _other) rcsbaseoff nocons eform nolog
note: delayed entry models are being fitted

Log likelihood =  -1228.025                     Number of obs     =      3,688

------------------------------------------------------------------------------------
                 |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------------
xb               |
treatment_cancer |   .6408643   .1083623    -2.63   0.009     .4600852    .8926761
         _cancer |   .3060732   .0335208   -10.81   0.000     .2469463    .3793569
   treatment_cvd |   1.329932   .2263497     1.68   0.094     .9527038    1.856525
            _cvd |   .2029639   .0262824   -12.32   0.000     .1574686    .2616034
 treatment_other |   .6740861   .1819979    -1.46   0.144     .3970979    1.144282
          _other |   .1034306   .0183681   -12.78   0.000     .0730273    .1464916
  (output omitted)
------------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation.

. predict cif_stcrprep_cancer, at(treatment_cancer 1 _cancer 1) zeros failure timevar(tempt)

. predict cif_stcrprep_cvd, at(treatment_cvd 1 _cvd 1) zeros failure timevar(tempt)

. predict cif_stcrprep_other, at(treatment_other 1 _other 1) zeros failure timevar(tempt)
```

stpm2cr

```
. stset time, failure(status == 1,2,3) exit(time 60) scale(12)

. stpm2cr [cancer: treatment, scale(hazard) df(4)] ///
> [cvd: treatment, scale(hazard) df(4)] ///
> [other: treatment, scale(hazard) df(4)], ///
> events(status) cause(1 2 3) cens(0) eform
  (output omitted)

. predict cifgq_trt1, cif at(treatment 1) timevar(tempt)
Calculating predictions for the following causes: 1 2 3
```

stpm2cr

```
. stset time, failure(status == 1,2,3) exit(time 60) scale(12)

. stpm2cr [cancer: treatment, scale(hazard) df(4)] ///
> [cvd: treatment, scale(hazard) df(4)] ///
> [other: treatment, scale(hazard) df(4)], ///
> events(status) cause(1 2 3) cens(0) eform
  (output omitted)

. predict cifgq_trt1, cif at(treatment 1) timevar(tempt)
Calculating predictions for the following causes: 1 2 3
```

Above is not comparable with time-dependent censoring weights approach as we assume proportionality for the competing causes of death.

stpm2cr

```
. stpm2cr [cancer: treatment, scale(hazard) df(4)] ///
> [cvd: treatment, scale(hazard) df(4) tvc(treatment) dftvc(3)] ///
> [other: treatment, scale(hazard) df(4) tvc(treatment) dftvc(3)], ///
> events(status) cause(1 2 3) cens(0) eform
```

(*output omitted*)

```
Log likelihood = -1117.3418                    Number of obs     =        506
```

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cancer						
treatment	.647454	.1094638	-2.57	0.010	.464834	.9018201
(*output omitted*)						
_cons	.1889881	.0229604	-13.71	0.000	.1489433	.2397993
(*output omitted*)						

stpm2cr

```
. stpm2cr [cancer: treatment, scale(hazard) df(4) tvc(treatment) dftvc(3)] ///
> [cvd: treatment, scale(hazard) df(4)] ///
> [other: treatment, scale(hazard) df(4) tvc(treatment) dftvc(3)], ///
> events(status) cause(1 2 3) cens(0) eform
```

(output omitted)

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(output omitted)						
cvd						
treatment	1.336129	.2273682	1.70	0.089	.9571939	1.865077
(output omitted)						
_cons	.1366028	.0187788	-14.48	0.000	.1043385	.178844
(output omitted)						

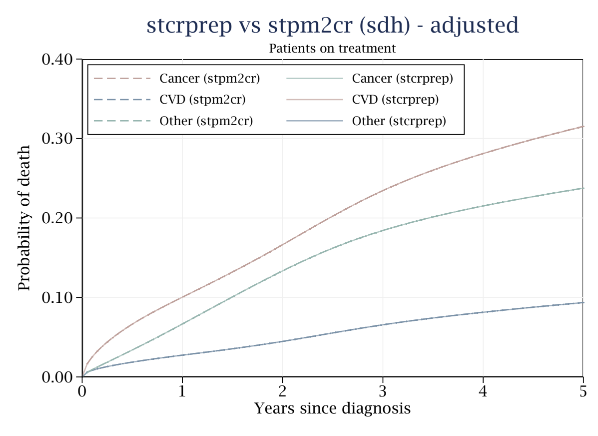
stpm2cr

```
. stpm2cr [cancer: treatment, scale(hazard) df(4) tvc(treatment) dftvc(3)] ///
> [cvd: treatment, scale(hazard) df(4) tvc(treatment) dftvc(3)] ///
> [other: treatment, scale(hazard) df(4)], ///
> events(status) cause(1 2 3) cens(0) eform
```

(*output omitted*)

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(*output omitted*)						
other						
treatment	.6771057	.1827954	-1.44	0.149	.3988974	1.149349
(*output omitted*)						
_cons	.0720086	.0138407	-13.69	0.000	.0494056	.1049525

Comparing stcrprep and stpm2cr



Comparison of computational time (to all k causes)

```
. expand 100 //now 50,060 observations

. replace time = time + runiform()*0.0001

. replace id = _n
variable id was int now long
```

	Time
stcrreg (total)	53 mins
stcrprep (total)	1 min
stpm2cr	17 secs

Summary of FPM tools for estimating cause-specific CIFs on (log-cumulative) SDH scale

- Using `stpm2` with time-dependent censoring weights
  - Need to prepare data first using `stcrprep`
  - Can use standard post-estimation commands such as`predict` (and `stpm2_standsurv`) as usual after `stpm2`
  - Computationally intensive for larger datasets
  - Requires more work for the user - increases room for error
- Post-estimation after `stpm2cr` for models on cause-specific CIF scale with `predict`
  - A single line of code to fit model
  - Does not require restructuring of data
  - Other predictions easy to obtain e.g. restricted mean lifetime
  - SEs/CIs obtained with analytically derived derivatives for the delta method - computationally quicker

References i

P. K. Andersen. Decomposition of number of life years lost according to causes of death. ***Statistics in Medicine***, 32:5278--85, Jul 2013.

J. Beyersmann, M. Dettenkofer, H. Bertz, and M. 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, Dec. 2007.

J. Beyersmann, A. Latouche, A. Buchholz, and M. Schumacher. Simulating competing risks data in survival analysis. ***Stat Med***, 28(6):956--971, 2009.

M. J. Crowther and P. C. Lambert. Parametric multistate survival models: Flexible modelling allowing transition-specific distributions with application to estimating clinically useful measures of effect differences. ***Statistics in medicine***, 36(29):4719--4742, 2017.

J. P. Fine and R. J. Gray. A proportional hazards model for the subdistribution of a competing risk. ***Journal of the American Statistical Association***, 446:496--509., 1999.

References ii

R. B. Geskus. Cause-specific cumulative incidence estimation and the Fine and Gray model under both left truncation and right censoring. ***Biometrics***, 67(1):39--49, Mar 2011.

S. R. Hinchliffe and P. C. Lambert. Extending the flexible parametric survival model for competing risks. ***The Stata Journal***, 13:344--355, 2013.

J.-H. Jeong and J. P. Fine. Parametric regression on cumulative incidence function. ***Biostatistics***, 8(2): 184--196, Apr 2007.

P. C. Lambert and P. Royston. Further development of flexible parametric models for survival analysis. ***The Stata Journal***, 9:265--290, 2009.

P. C. Lambert, S. R. Wilkes, and M. J. Crowther. Flexible parametric modelling of the cause-specific cumulative incidence function. ***Statistics in medicine***, 36(9):1429--1446, 2017.

A. Latouche, A. Allignol, J. Beyersmann, M. Labopin, and J. P. Fine. A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. ***J Clin Epidemiol***, 66(6): 648--653, Jun 2013.

References iii

S. I. Mozumder, M. J. Rutherford, P. C. Lambert, et al. A flexible parametric competing-risks model using a direct likelihood approach for the cause-specific cumulative incidence function. ***Stata Journal***, 17(2): 462--489, 2017.

P. Royston and M. K. B. Parmar. 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***, 21(15):2175--2197, Aug 2002.

P. Royston and M. K. B. Parmar. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. ***Stat Med***, 30(19): 2409--2421, Aug 2011.

P. Royston and M. K. B. Parmar. 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***, 13:152, 2013. ISSN 1471-2288.