Marginal estimates through regression standardization in competing risks and relative survival models

Paul C Lambert[1,2]

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

2019 Nordic and Baltic Stata Users Group meeting
Stockholm, 30 August 2019

Regression Standardization

1. Fit a statistical model that contains exposure, X, and potential confounders, Z.
2. Predict outcome for all individuals assuming they are all exposed (set $X = 1$).
3. Take mean to give marginal estimate of outcome.
4. Repeat by assuming all are unexposed (set $X = 0$).
5. Take the difference/ratio in means to form contrasts.

- Key point is the distribution of confounders, Z, is the same for the exposed and unexposed.
- If the model is sufficient for confounding control then such contrasts can be interpreted as causal effects.
- Also known as direct/model based standardization. G-formula (with no time-dependent confounders)[1].

Why not `margins`?

- `margins` does regression standardization, so why not use this?
- It is an excellent command, but does not do what I wanted for survival data.
- In particular, extensions to competing risks and relative survival.

Marginal survival time

- With survival data

X	-	is a binary exposure: 0 (unexposed) and 1 (exposed).
T	-	is a survival time.
T^0	-	is the potential survival time if X is set to 0.
T^1	-	is the potential survival time if X is set to 1.

- The average causal difference in mean survival time

$$E[T^1] - E[T^0]$$

- This is what `stteffects` can estimate.
- We often have limited follow-up and calculating the mean survival requires extrapolation and makes very strong distributional assumptions.

Marginal Survival functions

- Rather than use mean survival we can define our causal effect in terms of the marginal survival function.

$$E[T^1 > t] - E[T^0 > t]$$

- We can limit t within observed follow-up time.
- For confounders, $\boldsymbol{Z}$, we can write this as,

$$E[S(t|X = 1, \boldsymbol{Z})] - E[S(t|X = 0, \boldsymbol{Z})]$$

- Note that this is the expectation over the distribution of $\boldsymbol{Z}$.

Estimation

- Fit a survival model for exposure X and confounders $\boldsymbol{Z}$.
- Predict survival function for each individual setting $X = x$ and then average.
- Force everyone to be exposed and then unexposed.

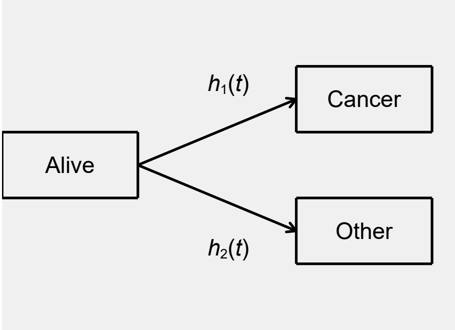
$$\frac{1}{N}\sum_{i=1}^{N}\widehat{S}\left(t|X = 1, \boldsymbol{Z} = z_i\right) - \frac{1}{N}\sum_{i=1}^{N}\widehat{S}\left(t|X = 0, \boldsymbol{Z} = z_i\right)$$

- Use their observed covariate pattern, $\boldsymbol{Z} = z_i$.
- We can standardize to an external (reference) population.

$$\frac{1}{N}\sum_{i=1}^{N} w_i \widehat{S}_i(t|X = x, \boldsymbol{Z} = z_i)$$

- `standsurv` will perform these calculation.

Competing risks



Separate models for each cause, e.g.

$$h_1(t|Z) = h_{0,1}(t)\exp\left(\beta_1 Z\right)$$
$$h_2(t|Z) = h_{0,2}(t)\exp\left(\beta_2 Z\right)$$

Two types of probability

- We may be interested in cause-specific survival/failure.

(1) *In the absence of other causes (net)*

$$F_k(t) = 1 - S_k(t) = P(T_k \leq t) = \int_0^t S_k(u)h_k(u)du$$

- We may be interested in cumulative incidence functions.

(2) *In the presence of other causes (crude)*

$$CIF_k(t) = P\left(T \leq t, \text{event} = k\right) = \int_0^t S(u)h_k(u)du$$

- Both are of interest - depends on research question.
- (1) Needs conditional independence assumption to interpret as net probability of death.

Description of Example

- 102,062 patients with bladder cancer in England (2002-2013).
- Death due to cancer and other causes.
- Covariates age, sex and deprivation in five groups.
- Restrict here to most and least deprived.

Models

- Flexible parametric (Royston-Parmar) models[2]
- Separate model for cancer and other causes.
- Age modelled using splines (3 df)
- 2-way interactions
- Time-dependent effects for all covariates.

Two separate cause-specific models

Cancer Model

```
stset dod, failure(status==1) exit(time min(dx+365.24*10,mdy(12,31,2013))) ///
          origin(dx) id(patid) scale(365.24)

stpm2 dep5 male agercs* dep_agercs*, df(5) scale(hazard) ///
          tvc(agercs* male dep5) dftvc(3)

estimates store cancer
```

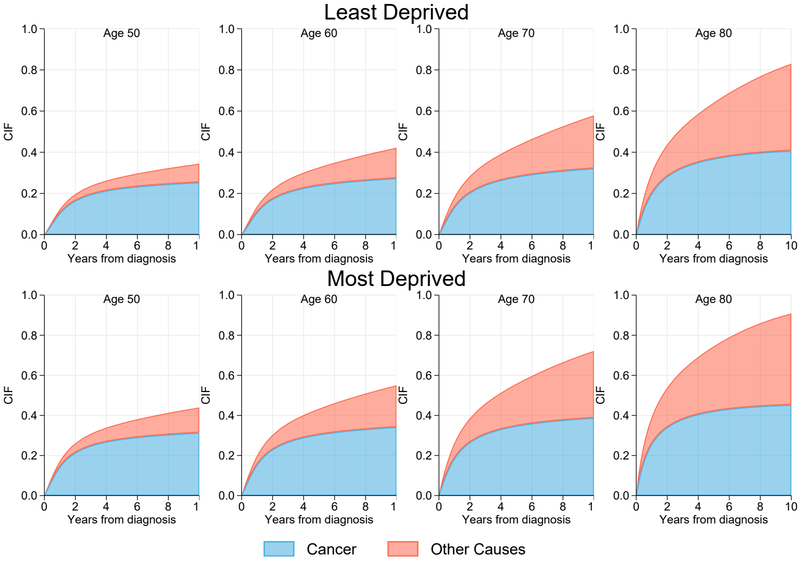
Other cause Model

```
stset dod, failure(status==2) exit(time min(dx+365.24*10,mdy(12,31,2013))) ///
          origin(dx) id(patid) scale(365.24)

stpm2 dep5 male agercs* dep_agercs*, df(5) scale(hazard) ///
          tvc(agercs* male dep5) dftvc(3)

estimates store other
```

Conditional cause-specific CIFs (Females)



Standardized cause-specific survival/failure

- Probability of death in the absence of other causes.
- Consider a single cause: standardize and form contrasts.

Cancer specific survival/failure

$$F_1(t) = 1 - S_1(t)$$

$$E\left[F_1(t)|X = 1, \boldsymbol{Z}\right] - E\left[F_1(t)|X = 0, \boldsymbol{Z}\right]$$

$$\frac{1}{N}\sum_{i=1}^{N}\widehat{F}_1(t|X = 1, \boldsymbol{Z} = z_i) - \frac{1}{N}\sum_{i=1}^{N}\widehat{F}_1(t|X = 0, \boldsymbol{Z} = z_i)$$

- Not a ‘real world’ probability, but comparisons between exposures where differential other cause mortality is removed is of interest.

Using `standsurv`

- Take mean of 102,062 survival functions where all individuals forced to be unexposed.
- Take mean of 102,062 survival functions where all individuals forced to be exposed.

Using standsurv

- Take mean of 102,062 survival functions where all individuals forced to be unexposed.
- Take mean of 102,062 survival functions where all individuals forced to be exposed.

```
. estimates restore cancer
. range tt 0 10 101

. standsurv, timevar(tt) failure ci                                          ///
   at1(dep5 0 dep_agercs1 0 dep_agercs2 0 dep_agercs3 0)                     ///
   at2(dep5 1 dep_agercs1=agercs1 dep_agercs2=agercs2 dep_agercs3=agercs3)   ///
   contrast(difference)                                                      ///
   atvar(F_cancer_s_dep1 F_cancer_s_dep5)                                    ///
   contrastvar(F_cancer_diff)
```

Using standsurv

- Take mean of 102,062 survival functions where all individuals forced to be unexposed.
- Take mean of 102,062 survival functions where all individuals forced to be exposed.

```
. estimates restore cancer
. range tt 0 10 101

. standsurv, timevar(tt) failure ci                                           ///
   at1(dep5 0 dep_agercs1 0 dep_agercs2 0 dep_agercs3 0)                      ///
   at2(dep5 1 dep_agercs1=agercs1 dep_agercs2=agercs2 dep_agercs3=agercs3)   ///
   contrast(difference)                                                       ///
   atvar(F_cancer_s_dep1 F_cancer_s_dep5)                                     ///
   contrastvar(F_cancer_diff)
```

Using standsurv

- Take mean of 102,062 survival functions where all individuals forced to be unexposed.
- Take mean of 102,062 survival functions where all individuals forced to be exposed.

```
. estimates restore cancer
. range tt 0 10 101

. standsurv, timevar(tt) failure ci                                       ///
   at1(dep5 0 dep_agercs1 0 dep_agercs2 0 dep_agercs3 0)                  ///
   at2(dep5 1 dep_agercs1=agercs1 dep_agercs2=agercs2 dep_agercs3=agercs3) ///
   contrast(difference)                                                   ///
   atvar(F_cancer_s_dep1 F_cancer_s_dep5)                                 ///
   contrastvar(F_cancer_diff)
```

Using standsurv

- Take mean of 102,062 survival functions where all individuals forced to be unexposed.
- Take mean of 102,062 survival functions where all individuals forced to be exposed.

```
.  estimates restore cancer
.  range tt 0 10 101

.  standsurv, timevar(tt) failure ci                                           ///
    at1(dep5 0 dep_agercs1 0 dep_agercs2 0 dep_agercs3 0)                      ///
    at2(dep5 1 dep_agercs1=agercs1 dep_agercs2=agercs2 dep_agercs3=agercs3)    ///
    contrast(difference)                                                       ///
    atvar(F_cancer_s_dep1 F_cancer_s_dep5)                                     ///
    contrastvar(F_cancer_diff)
```

Using standsurv

- Take mean of 102,062 survival functions where all individuals forced to be unexposed.
- Take mean of 102,062 survival functions where all individuals forced to be exposed.

```
. estimates restore cancer
. range tt 0 10 101

. standsurv, timevar(tt) failure ci                                        ///
   at1(dep5 0 dep_agercs1 0 dep_agercs2 0 dep_agercs3 0)                  ///
   at2(dep5 1 dep_agercs1=agercs1 dep_agercs2=agercs2 dep_agercs3=agercs3) ///
   contrast(difference)                                                    ///
   atvar(F_cancer_s_dep1 F_cancer_s_dep5)                                  ///
   contrastvar(F_cancer_diff)
```

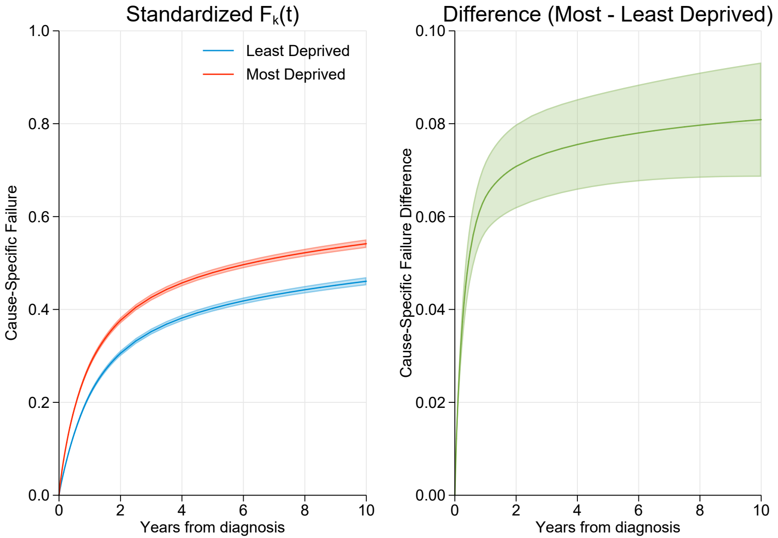
Using standsurv

- Take mean of 102,062 survival functions where all individuals forced to be unexposed.
- Take mean of 102,062 survival functions where all individuals forced to be exposed.

```
. estimates restore cancer
. range tt 0 10 101

. standsurv, timevar(tt) failure ci                                            ///
   at1(dep5 0 dep_agercs1 0 dep_agercs2 0 dep_agercs3 0)                      ///
   at2(dep5 1 dep_agercs1=agercs1 dep_agercs2=agercs2 dep_agercs3=agercs3)    ///
   contrast(difference)                                                        ///
   atvar(F_cancer_s_dep1 F_cancer_s_dep5)                                      ///
   contrastvar(F_cancer_diff)
```

Standardized cause-specific Failure $(1 - S_k(t))$



Standardized cause-specific CIF

- Probability of death in the presence of other causes.
- We can standardize the cause-specific CIF in the same way.
- These requires combining K different models

$$E\left[CIF_k(t)|X = x, Z\right]$$

$$\frac{1}{N}\sum_{i=1}^{N}\int_0^t \widehat{S}(u|X = x, \boldsymbol{Z} = z_i)\widehat{h}_k(u|X = x, \boldsymbol{Z} = , z_i)du$$

- Calculate for X=1 and X=0 and then obtain contrast.
- Can be interpreted as causal effects under assumptions[3].

Using `standsurv`

- Take mean of 102,062 CIFs where all individuals forced to be unexposed.
- Take mean of 102,062 CIFs where all individuals forced to be exposed.

Using standsurv

- Take mean of 102,062 CIFs where all individuals forced to be unexposed.
- Take mean of 102,062 CIFs where all individuals forced to be exposed.

```
. standsurv, crmodels(cancer other) timevar(tt) cif ci                ///
   at1(dep5 0 dep_agercs1 0 dep_agercs2 0 dep_agercs3 0)              ///
   at2(dep5 1 dep_agercs1=agercs1 dep_agercs2=agercs2 dep_agercs3=agercs3)  ///
   contrast(difference)                                               ///
   atvar(CIF_s_dep1 CIF_s_dep5))                                      ///
   contrastvar(CIF_diff)
```

Using standsurv

- Take mean of 102,062 CIFs where all individuals forced to be unexposed.
- Take mean of 102,062 CIFs where all individuals forced to be exposed.

```
. standsurv, crmodels(cancer other) timevar(tt) cif ci                    ///
   at1(dep5 0 dep_agercs1 0 dep_agercs2 0 dep_agercs3 0)                  ///
   at2(dep5 1 dep_agercs1=agercs1 dep_agercs2=agercs2 dep_agercs3=agercs3) ///
   contrast(difference)                                                   ///
   atvar(CIF_s_dep1 CIF_s_dep5))                                          ///
   contrastvar(CIF_diff)
```

Using standsurv

- Take mean of 102,062 CIFs where all individuals forced to be unexposed.
- Take mean of 102,062 CIFs where all individuals forced to be exposed.

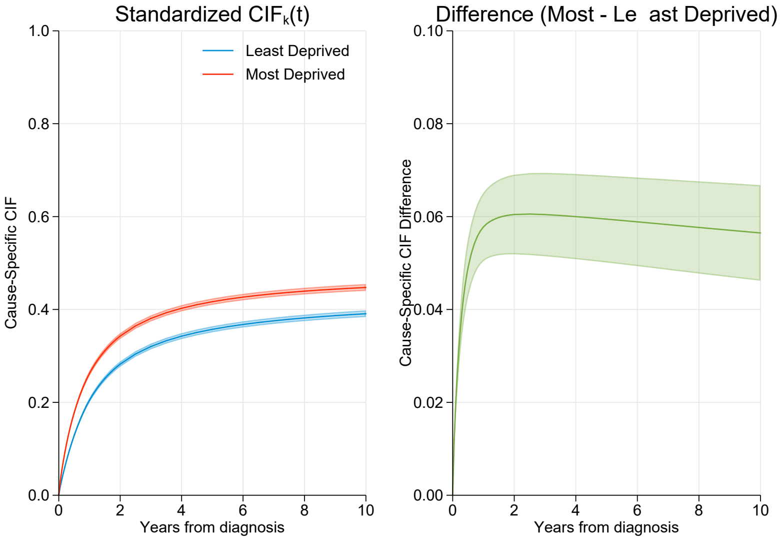
```
. standsurv, crmodels(cancer other) timevar(tt) cif ci                    ///
   at1(dep5 0 dep_agercs1 0 dep_agercs2 0 dep_agercs3 0)                  ///
   at2(dep5 1 dep_agercs1=agercs1 dep_agercs2=agercs2 dep_agercs3=agercs3) ///
   contrast(difference)                                                   ///
   atvar(CIF_s_dep1 CIF_s_dep5))                                          ///
   contrastvar(CIF_diff)
```

Using standsurv

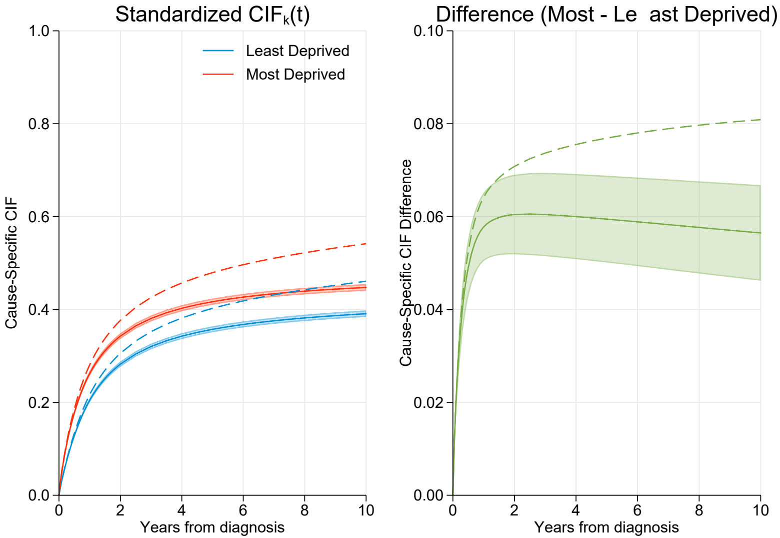
- Take mean of 102,062 CIFs where all individuals forced to be unexposed.
- Take mean of 102,062 CIFs where all individuals forced to be exposed.

```
. standsurv, crmodels(cancer other) timevar(tt) cif ci                    ///
   at1(dep5 0 dep_agercs1 0 dep_agercs2 0 dep_agercs3 0)                 ///
   at2(dep5 1 dep_agercs1=agercs1 dep_agercs2=agercs2 dep_agercs3=agercs3) ///
   contrast(difference)                                                   ///
   atvar(CIF_s_dep1 CIF_s_dep5))                                          ///
   contrastvar(CIF_diff)
```

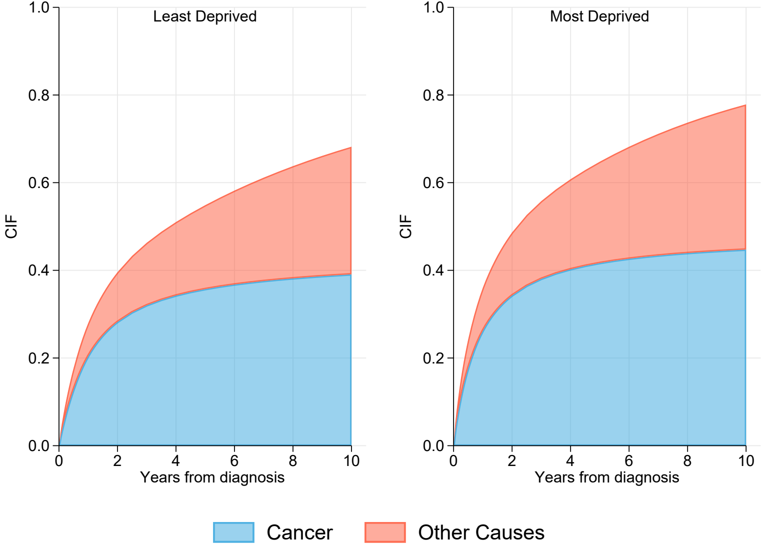
Standardized cause-specific CIF



Standardized cause-specific CIF



Stacked standardized cause-specific CIF



Timings for standardized survival/failure functions

- N individuals, 1 event , exposure X, 10 confounders $\boldsymbol{Z}$.
- Fit model: Standardized $S(t|X=x,\boldsymbol{Z})$ for $X=0$ & $X=1$ and contrasts with CIs.
- Calculate time for Weibull models and FPMs.

N	Weibull		FPM	
	Point Estimate	Confidence Interval	Point Estimate	Confidence Interval
1,000	0.02	0.03	0.03	0.05
10,000	0.04	0.1	0.09	0.1
100,000	0.4	0.7	0.6	0.9
250,000	1.0	1.8	1.6	2.6
500,000	2.0	3.5	2.5	4.5
1,000,000	3.9	4.6	5.5	11.1

Times in seconds on standard issue University of Leicester laptop.

Timings for standardized cause-specific CIF

- N individuals, 2 events , exposure X, 10 confounders $\boldsymbol{Z}$.
- Fit 2 models: standardized CIF for $X = 0$ & $X = 1$ and contrast with CIs.
- Calculate time for Weibull models and FPMs.

N	Weibull		FPM	
	Point Estimate	Confidence Interval	Point Estimate	Confidence Interval
1,000	0.1	0.3	0.3	1.4
10,000	0.2	2.1	2.1	8.6
100,000	13.2	16.8	20.6	93.9
250,000	5.8	48.1	56.1	246.4
500,000	10.1	97.7	117.2	521.2
1,000,000	24.2	159.0	225.6	1018.9

Times in seconds on standard issue University of Leicester laptop.

Relative Survival

- Relative survival models used with large population cancer registry data when cause of death not available or not reliable.

$$h(t|X, Z) = h^*(t|X, Z) + \lambda(t|X, Z)$$

$h(t\|X, Z)$	-	All-cause mortality rate
$h^*(t\|X, Z)$	-	Expected mortality rate
$\lambda(t\|X, Z)$	-	Excess mortality rate

- Expected mortality rates obtained from national lifetables.
- On survival scale.

$$S(t|X, Z) = S^*(t|X, Z)R(t|X, Z)$$

- The equivalent of a CIF is know as a crude probability in the relative survival framework.

Melanoma Example

Relative Survival Model

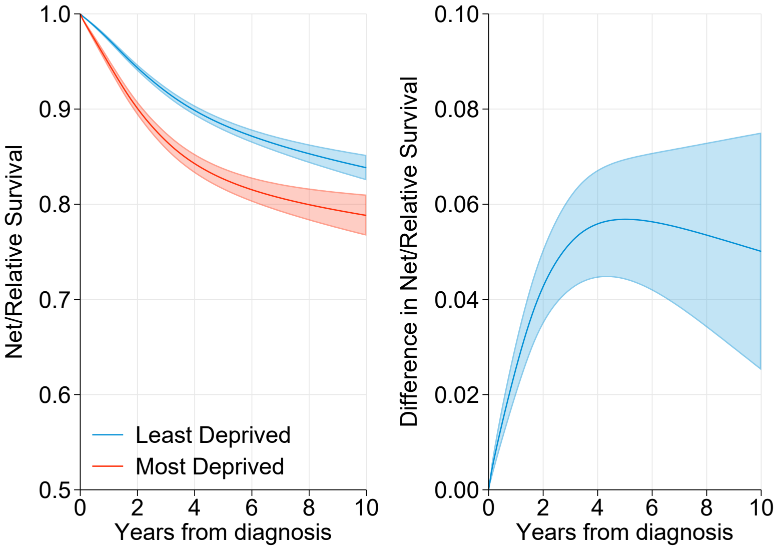
```
stpm2 dep5 agercs* , scale(hazard) df(5) tvc(dep5 agercs*) dftvc(3) bhazard(rate)
```

$$\overline{R}(t|X = x, Z) = \frac{1}{N}\sum_{i=1}^{N} R_i(t|X = x, Z = z_i)$$

Standardized Relative Survival

```
standsurv,  timevar(tt) ci                                                    ///
  at1(dep5 0 agercs1_dep5 0 agercs2_dep5 0 agercs3_dep5 0)                    ///
  at2(dep5 1 agercs1_dep5=agercs1 agercs2_dep5=agercs2 agercs3_dep5=agercs3)  ///
  contrast(difference)                                                        ///
  atvar(R_dep5 R_dep1)                                                        ///
  contrastvar(R_diff)
```

Standardized Relative Survival

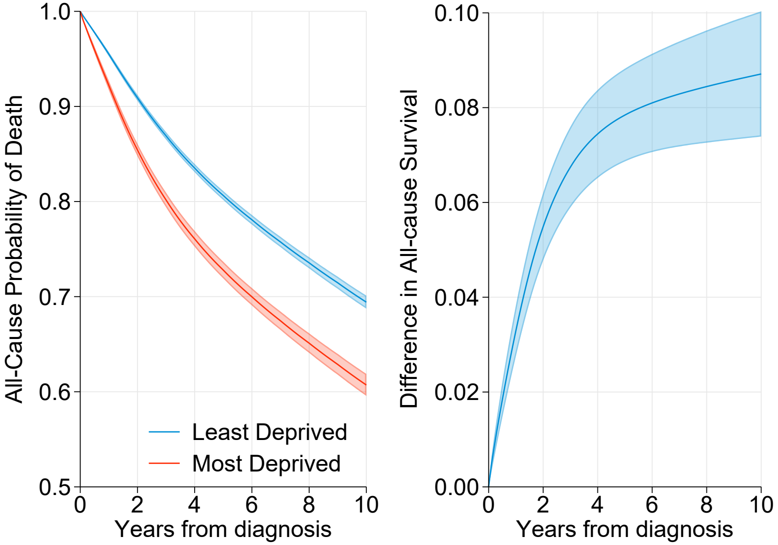


All-cause Survival

$$\bar{S}(t|X = x, Z) = \frac{1}{N}\sum_{i=1}^{N} S^*(t|X = x, Z = z_i)$$

```
standsurv,  timevar(tt) ci                                              ///
  at1(dep5 0 agercs1_dep5 0 agercs2_dep5 0 agercs3_dep5 0)              ///
  at2(dep5 1 agercs1_dep5=agercs1 agercs2_dep5=agercs2 agercs3_dep5=agercs3)  ///
  expsurv(using(popmort_uk_regions_2017.dta)                            ///
     datediag(dx)                                                       ///
     agediag(agediag)                                                   ///
     pmrate(rate)                                                       ///
     pmage(age)                                                         ///
     pmyear(year)                                                       ///
     pmother(sex dep region)                                            ///
     pmmaxyear(2016)                                                    ///
     at1(dep 1)                                                         ///
     at2(dep 5))                                                        ///
  contrast(difference)                                                  ///
  atvar(S_dep5 S_dep1)                                                  ///
  contrastvar(S_diff)
```

Standardized All-cause Survival

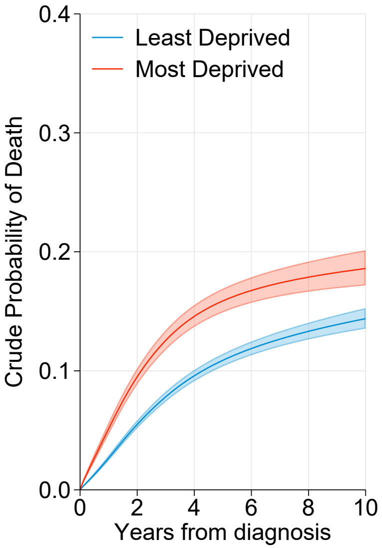


Standardized Crude Probabilities

$$\overline{F}_c(t|X=x,\boldsymbol{Z}) = \frac{1}{N}\sum_{i=1}^{N}\int_0^t S^*(u|X=x,\boldsymbol{Z}=\boldsymbol{z}_i)R(u|X=x,\boldsymbol{Z}=\boldsymbol{z}_i)\lambda(u|X=x,\boldsymbol{Z}=\boldsymbol{z}_i),$$

```
standsurv,  crudeprob timevar(tt) ci                                      ///
  at1(dep5 0 agercs1_dep5 0 agercs2_dep5 0 agercs3_dep5 0)               ///
  at2(dep5 1 agercs1_dep5=agercs1 agercs2_dep5=agercs2 agercs3_dep5=agercs3)  ///
  expsurv(using(popmort_uk_regions_2017.dta)                             ///
     datediag(dx)                                                        ///
     agediag(agediag)                                                    ///
     pmrate(rate)                                                        ///
     pmage(age)                                                          ///
     pmyear(year)                                                        ///
     pmother(sex dep region)                                             ///
     pmmaxyear(2016)                                                     ///
     at1(dep 1)                                                          ///
     at2(dep 5))                                                         ///
  contrast(difference)                                                   ///
  atvar(CP_dep5 CP_dep1)                                  ///
  contrastvar(CP_diff)
```

Standardized Crude Probabilities of Death



standsurv

- standsurv works for a many parametric models
  - streg:Exponential, Weibull, Gompertz, LogNormal, LogLogistic
  - Flexible parametric (Splines: strcs (log hazard) or stpm2 (log cumulative hazard))
- Standard, relative survival and competing risks models
  - Can use different models for different causes. E.g. Weibull for one cause and flexible parametric model for another
- Various Standardizations
  - Survival, restricted means, centiles, hazards... and more
- Standard errors calculated using delta-method or M-estimation with all analytical derivatives,so fast

More information on standsurv available at

https://pclambert.net/software/standsurv/

Summary

- Regression standardisation is a simple and underused tool
- Can also estimate causal effects using IPW.
- Advantages of regression adjustment
  - Not a big leap from what people doing at the moment - model may be the same, just report in a different way.
  - We often do not want to just report marginal effects - predictions for specific covariate patterns are still of interest.
- As long as we can predict survival function, models can be as complex as we like (non-linear effects, non-proportional hazards, interactions with exposure etc.)

References

[1] Vansteelandt S, Keiding N. Invited commentary: G-computation–lost in translation? *Am J Epidemiol* 2011;**173**:739–742.

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

[3] Young JG, Tchetgen Tchetgen EJ, Hernan MA. The choice to define competing risk events as censoring events and implications for causal inference. *arXiv preprint* 2018;.

[4] Sjölander A. Regression standardization with the R package stdReg. *European Journal of Epidemiology* 2016;**31**:563–574.

[5] Kipourou DK, Charvat H, Rachet B, Belot A. Estimation of the adjusted cause-specific cumulative probability using flexible regression models for the cause-specific hazards. *Statistics in medicine* 2019;.