UNIVERSITY OF LEICESTER

Understanding inequalities in cancer prognosis: An extension of mediation analysis to the relative survival framework.

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

[1]Biostatistics Research Group, University of Leicester

[2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet

[*]*es303@.le.ac.uk*

Funding: NIHR Doctoral Research Fellowship Programme

8 October 2019

Motivation

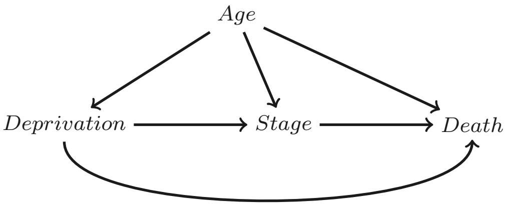
Survival after a cancer diagnosis varies considerably across population groups e.g socioeconomic groups.

Deprivation Group	5-year RS	Mean Years w/o Cancer	Mean Years with Cancer	Prop (%)
Age-at-diagnosis: 60				
Least Deprived	64.83	27.06	16.57	38.75
Most Deprived	56.74	23.08	12.36	46.44
Age-at-diagnosis: 70				
Least Deprived	63.57	18.26	11.38	37.65
Most Deprived	53.96	15.39	8.23	46.52

[1] Understanding the impact of socioeconomic differences in colorectal cancer survival: potential gain in life-years. *Brit J Cancer* 2019; 20:1052–1058

Understanding variation

Is there a third variable that can partly explain these differences?



Complex mechanisms contribute towards disparities: all-cause survival differences are the result of both cancer-related and other cause factors.

Marginal estimates

Let us assume we are interested on the effect of an exposure X on the survival time while adjusting for confounders Z.

A summary of the population prognosis can be obtained by the standardised relative survival function $E\left[R(t|Z)\right]$ that is estimated by:

$$E\left[\widehat{R}(t|Z)\right] = \frac{1}{N}\sum_{i=1}^{N}\widehat{R}(t|Z = z_i).$$

Regression standardisation

1. Fit a survival model such as flexible parametric model.
2. Obtain survival predictions for each individual in the population.
3. Calculate an average of the survival predictions.

Forming contrasts

If interested in **relative** survival:

$$E\left[R(t|X=1,Z)\right]-E\left[R(t|X=0,Z)\right]$$

- Refers to a hypothetical world where the cancer of interest is the only possible cause of death.

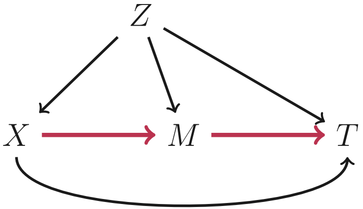
If interested in **all-cause** survival:

$$E\left[S(t|X=1,Z)\right]-E\left[S(t|X=0,Z)\right]$$
$$E\left[S^{*}(t|X=1,Z)R(t|X=1,Z)\right]-E\left[S^{*}(t|X=0,Z)R(t|X=0,Z)\right]$$

- Differences may be due to either cancer of interest or other cause mortality or both.

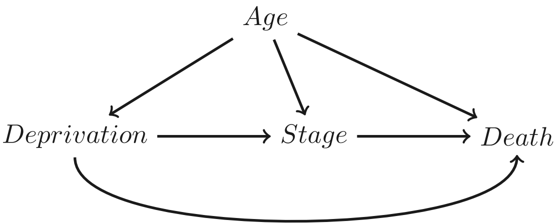
Exploring the effect of a mediator

How much of the differences between exposure groups can be explained by differences at the mediator M distribution?



Let M^x denote the counterfactual mediator distribution when intervening to set $X = x$.

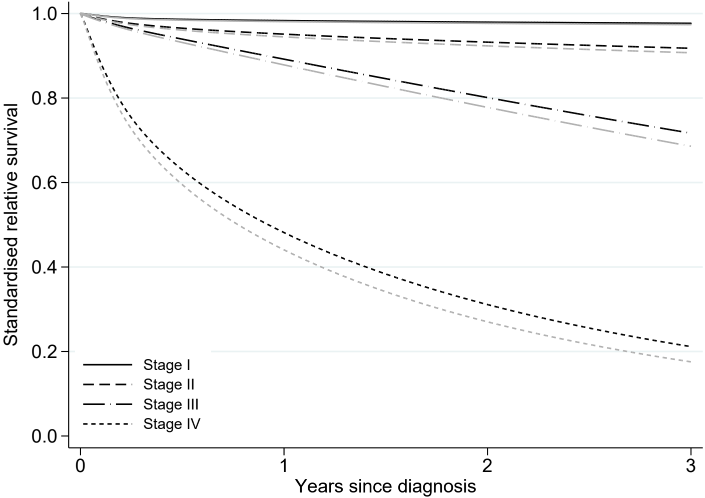
Illustration data - colon cancer in England



Data on 15,630 patients diagnosed between 2011-2013 (57.6% in the least deprived group).

Stage at diagnosis	Least Deprived	Most deprived
I	1338(**14.86%**)	912(**13.76%**)
II	2644(29.37%)	1950(29.42%)
III	2435(27.05%)	1716(25.89%)
IV	2585(**28.72%**)	2050(**30.93%**)

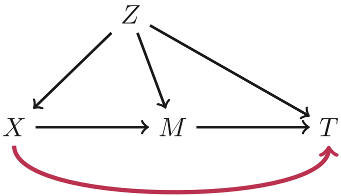
Relative survival by stage



Direct & indirect effects - relative survival framework

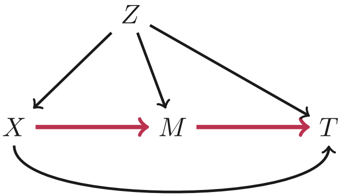
Natural direct effect

$$NDE_{RS} = E\left[R(t|\boldsymbol{X=1}, Z, M^0)\right] - E\left[R(t|\boldsymbol{X=0}, Z, M^0)\right]$$



Natural indirect effect

$$NIE_{RS} = E\left[R(t|X=1, Z, \boldsymbol{M^1})\right] - E\left[R(t|X=1, Z, \boldsymbol{M^0})\right]$$



Estimation

Step 1. Fit a survival model including X, M, Z.

```
stpm2 dep5 rcsa1 rcsa2 rcsa3 gender stage2 stage3 stage4 ///
      st2dep5 st3dep5 st4dep5, df(5) scale(h) bhaz(rate) ///
      tvc(rcsa1 rcsa2 rcsa3 dep5 stage2 stage3 stage4)  dftvc(3)


estimates store surv
```

Estimation

Step 1. Fit a survival model including X, M, Z.

```
stpm2 dep5 rcsa1 rcsa2 rcsa3 gender stage2 stage3 stage4 ///
      st2dep5 st3dep5 st4dep5, df(5) scale(h) bhaz(rate) ///
      tvc(rcsa1 rcsa2 rcsa3 dep5 stage2 stage3 stage4)  dftvc(3)


estimates store surv
```

Step 2. Fit a separate model for the mediator including X, Z.

```
//Fit a multinomial regression model for the most deprived
mlogit cancer_stage rcsa1 rcsa2 rcsa3 gender if dep5==1
estimates store ph1

//Fit a multinomial regression model for the least deprived
mlogit cancer_stage rcsa1 rcsa2 rcsa3 gender if dep5==0
estimates store ph0
```

Estimation

Step 3. For each individual in the study population obtain predictions for $\widehat{P}(M = m|X = x, Z = z_i)$, at each $X = x$.

```
preserve
   estimates restore pho
   matrix b0 = e(b)
   matrix V0= e(V)
   drawnorm b1_rcsa1 b1_rcsa2 b1_rcsa3 b1_gender b1_cons ///
            b2_rcsa1 b2_rcsa2 b2_rcsa3 b2_gender b2_cons ///
            b3_rcsa1 b3_rcsa2 b3_rcsa3 b3_gender b3_cons ///
            b4_rcsa1 b4_rcsa2 b4_rcsa3 b4_gender b4_cons, mean(b0) cov(V0) n(1) clear
   local cnames:  colfullnames b0
   local rnames:  rowfullnames b0
   mkmat  b1_rcsa1 b1_rcsa2 b1_rcsa3 b1_gender b1_cons ///
          b2_rcsa1 b2_rcsa2 b2_rcsa3 b2_gender b2_cons ///
          b3_rcsa1 b3_rcsa2 b3_rcsa3 b3_gender b3_cons ///
          b4_rcsa1 b4_rcsa2 b4_rcsa3 b4_gender b4_cons, matrix(b0_tmp)
   matrix colnames b0_tmp = ‘cnames’
   matrix rownames b0_tmp = ‘rnames’
   erepost b = b0_tmp V=V0, noesample
restore
//Obtain predictions for stages 1,2,3 and 4 (for least deprived)
predict p01 p02 p03 p04
//Repeat for the most deprived group: p11 p12 p13 p14
```

Estimation

Step 4. Obtain predictions of $\widehat{R}(t|X=x, Z=z_i, M=m)$ at $X=x$, using the predictions of Step 2 as weights.

$$TCE_{RS} = E\left[R(t|X=1, Z, M^1)\right] - E\left[R(t|X=0, Z, M^0)\right]$$

```
//First draw the model parameters from a multivariate normal distribution for the
survival model (similar to Step 3).

//Obtain predictions for the TCE
standsurv, failure timevar(timevar)   ///
  at1(dep5 1 stage2 0 stage3 0 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p11))
  at2(dep5 1 stage2 1 stage3 0 stage4 0 st2dep5 1 st3dep5 0 st4dep5 0, atindweights(p12))
  at3(dep5 1 stage2 0 stage3 1 stage4 0 st2dep5 0 st3dep5 1 st4dep5 0, atindweights(p13))
  at4(dep5 1 stage2 0 stage3 0 stage4 1 st2dep5 0 st3dep5 0 st4dep5 1, atindweights(p14))
  at5(dep5 0 stage2 0 stage3 0 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p01))
  at6(dep5 0 stage2 1 stage3 0 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p02))
  at7(dep5 0 stage2 0 stage3 1 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p03))
  at8(dep5 0 stage2 0 stage3 0 stage4 1 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p04))
  lincom(1 1 1 1 -1 -1 -1 -1) lincomvar(tce)
```

Estimation

$$NDE_{RS} = E\left[R(t|\boldsymbol{X} = \mathbf{1}, Z, M^0)\right] - E\left[R(t|\boldsymbol{X} = \mathbf{0}, Z, M^0)\right]$$

```
//First draw the model parameters from a multivariate normal distribution for the
survival model (similar to Step 3).

//Obtain predictions for the NDE
standsurv, failure timevar(timevar)   ///
  at1(dep5 1 stage2 0 stage3 0 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p01))
  at2(dep5 1 stage2 1 stage3 0 stage4 0 st2dep5 1 st3dep5 0 st4dep5 0, atindweights(p02))
  at3(dep5 1 stage2 0 stage3 1 stage4 0 st2dep5 0 st3dep5 1 st4dep5 0, atindweights(p03))
  at4(dep5 1 stage2 0 stage3 0 stage4 1 st2dep5 0 st3dep5 0 st4dep5 1, atindweights(p04))
  at5(dep5 0 stage2 0 stage3 0 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p01))
  at6(dep5 0 stage2 1 stage3 0 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p02))
  at7(dep5 0 stage2 0 stage3 1 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p03))
  at8(dep5 0 stage2 0 stage3 0 stage4 1 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p04))
  lincom(1 1 1 1 -1 -1 -1 -1) lincomvar(nde)
```

Estimation

$$NIE_{RS} = E\left[R(t|X=1, Z, \boldsymbol{M^1})\right] - E\left[R(t|X=1, Z, \boldsymbol{M^0})\right]$$

```
//First draw the model parameters from a multivariate normal distribution for the
survival model (similar to Step 3).

//Obtain predictions for the NIE
standsurv, failure timevar(timevar)   ///
  at1(dep5 1 stage2 0 stage3 0 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p11))
  at2(dep5 1 stage2 1 stage3 0 stage4 0 st2dep5 1 st3dep5 0 st4dep5 0, atindweights(p12))
  at3(dep5 1 stage2 0 stage3 1 stage4 0 st2dep5 0 st3dep5 1 st4dep5 0, atindweights(p13))
  at4(dep5 1 stage2 0 stage3 0 stage4 1 st2dep5 0 st3dep5 0 st4dep5 1, atindweights(p14))
  at5(dep5 1 stage2 0 stage3 0 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p01))
  at6(dep5 1 stage2 1 stage3 0 stage4 0 st2dep5 1 st3dep5 0 st4dep5 0, atindweights(p02))
  at7(dep5 1 stage2 0 stage3 1 stage4 0 st2dep5 0 st3dep5 1 st4dep5 0, atindweights(p03))
  at8(dep5 1 stage2 0 stage3 0 stage4 1 st2dep5 0 st3dep5 0 st4dep5 1, atindweights(p04))
  lincom(1 1 1 1 -1 -1 -1 -1) lincomvar(nie)
```

Estimation

Step 5. Repeat from Step 2, k times, while performing parametric bootstrap for the parameter estimates for both models.

Estimation

Step 5. Repeat from Step 2, k times, while performing parametric bootstrap for the parameter estimates for both models.

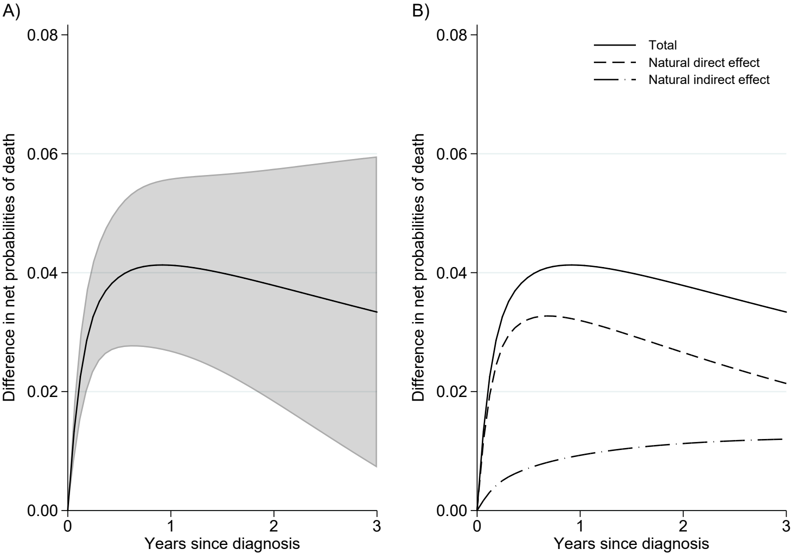
Step 6. Calculate 95% confidence intervals either by taking the 2.5% and 97.5% percentiles of the estimates across the bootstrapped samples or by using their standard deviation.

Estimation

$$\widehat{NDE_{RS}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{m} \boldsymbol{\widehat{R}(t|X=1, Z=z_i, M=m)}\widehat{P}(M=m|X=0, Z=z_i) - \frac{1}{N}\sum_{i=1}^{N}\sum_{m} \boldsymbol{\widehat{R}(t|X=0, Z=z_i, M=m)}\widehat{P}(M=m|X=0, Z=z_i)$$

$$\widehat{NIE_{RS}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{m} \widehat{R}(t|X=1, Z=z_i, M=m)\boldsymbol{\widehat{P}(M=m|X=1, Z=z_i)} - \frac{1}{N}\sum_{i=1}^{N}\sum_{m} \widehat{R}(t|X=1, Z=z_i, M=m)\boldsymbol{\widehat{P}(M=m|X=0, Z=z_i)}$$

Colon cancer



Direct & indirect effects - all cause setting

- Compare $S^*(t|X=1,Z)$ with $S^*(t|X=0,Z)$:

$$NDE_{AC1} = E\left[\boldsymbol{S^*(t|X=1,Z)}R(t|X=1,Z,M^0)\right] - E\left[\boldsymbol{S^*(t|X=0,Z)}R(t|X=0,Z,M^0)\right]$$

Differences may be due to either the cancer of interest or other causes or both.

```
standsurv, failure timevar(timevar)
  at1(dep5 1 stage2 0 stage3 0 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p01))
  at2(dep5 1 stage2 1 stage3 0 stage4 0 st2dep5 1 st3dep5 0 st4dep5 0, atindweights(p02))
  at3(dep5 1 stage2 0 stage3 1 stage4 0 st2dep5 0 st3dep5 1 st4dep5 0, atindweights(p03))
  at4(dep5 1 stage2 0 stage3 0 stage4 1 st2dep5 0 st3dep5 0 st4dep5 1, atindweights(p04))
  at5(dep5 0 stage2 0 stage3 0 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p01))
  at6(dep5 0 stage2 1 stage3 0 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p02))
  at7(dep5 0 stage2 0 stage3 1 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p03))
  at8(dep5 0 stage2 0 stage3 0 stage4 1 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p04))
  lincom(1 1 1 1 -1 -1 -1 -1) lincomvar(nde_ac1)
  expsurv(using(popmort.dta)
  datediag(dx) agediag(agediag) pmrate(rate) pmage(age) pmyear(year)  pmother(dep sex)
    at1(dep 1) at2(dep 1)
    at3(dep 1) at4(dep 1)
    at5(dep 0) at6(dep 0)
    at7(dep 0) at8(dep 0))
```

Direct & indirect effects - all cause setting

- Use the observed distribution of the exposure, $S^*(t|X,Z)$:

$$NDE_{AC2} = E\left[\boldsymbol{S^*(t|X,Z)}R(t|X=1,Z,M^0)\right] - E\left[\boldsymbol{S^*(t|X,Z)}R(t|X=0,Z,M^0)\right]$$

Differences can only be due to the cancer of interest.

```
standsurv, failure timevar(timevar)
  at1(dep5 1 stage2 0 stage3 0 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p01))
  at2(dep5 1 stage2 1 stage3 0 stage4 0 st2dep5 1 st3dep5 0 st4dep5 0, atindweights(p02))
  at3(dep5 1 stage2 0 stage3 1 stage4 0 st2dep5 0 st3dep5 1 st4dep5 0, atindweights(p03))
  at4(dep5 1 stage2 0 stage3 0 stage4 1 st2dep5 0 st3dep5 0 st4dep5 1, atindweights(p04))
  at5(dep5 0 stage2 0 stage3 0 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p01))
  at6(dep5 0 stage2 1 stage3 0 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p02))
  at7(dep5 0 stage2 0 stage3 1 stage4 0 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p03))
  at8(dep5 0 stage2 0 stage3 0 stage4 1 st2dep5 0 st3dep5 0 st4dep5 0, atindweights(p04))
  lincom(1 1 1 1 -1 -1 -1 -1) lincomvar(nde_ac2)
  expsurv(using(popmort.dta)
  datediag(dx) agediag(agediag) pmrate(rate) pmage(age) pmyear(year)  pmother(dep sex)
    at1(dep .) at2(dep .)
    at3(dep .) at4(dep .)
    at5(dep .) at6(dep .)
    at7(dep .) at8(dep .))
```

Direct & indirect effects - all cause setting

- It might also be of interest to estimate the effect, within subsets of the whole population e.g. NDE among the exposed using $S^*(t|X=1, Z_{X=1})$.

Avoidable deaths under hypothetical interventions

“What if we could eliminate differences in the mediator distribution between exposed and unexposed groups?”

Avoidable deaths under hypothetical interventions

“What if we could eliminate differences in the mediator distribution between exposed and unexposed groups?”

- The predicted number of deaths for the exposed:

$$D_1(t|X=1,M^1) = N^* \times \left(1 - E\left[S^*(t|X=1,Z_{X=1})R(t|X=1,Z,M^1)\right]\right)$$

Avoidable deaths under hypothetical interventions

“What if we could eliminate differences in the mediator distribution between exposed and unexposed groups?”

- The predicted number of deaths for the exposed:

$$D_1(t|X=1,M^1) = N^* \times \left(1 - E\left[S^*(t|X=1,Z_{X=1})R(t|X=1,Z,M^1)\right]\right)$$

- The expected number of deaths if the exposed had the same **mediator distribution** as the unexposed:

$$D_M(t|X=1,M^0) = N^* \times \left(1 - E\left[S^*(t|X=1,Z_{X=1})R(t|X=0,Z,\boldsymbol{M^0})\right]\right)$$

Avoidable deaths under hypothetical interventions

“What if we could eliminate differences in the mediator distribution between exposed and unexposed groups?”

- The predicted number of deaths for the exposed:

$$D_1(t|X=1,M^1) = N^* \times \left(1 - E\left[S^*(t|X=1,Z_{X=1})R(t|X=1,Z,M^1)\right]\right)$$

- The expected number of deaths if the exposed had the same **mediator distribution** as the unexposed:

$$D_M(t|X=1,M^0) = N^* \times \left(1 - E\left[S^*(t|X=1,Z_{X=1})R(t|X=0,Z,\boldsymbol{M^0})\right]\right)$$

- The avoidable deaths are:

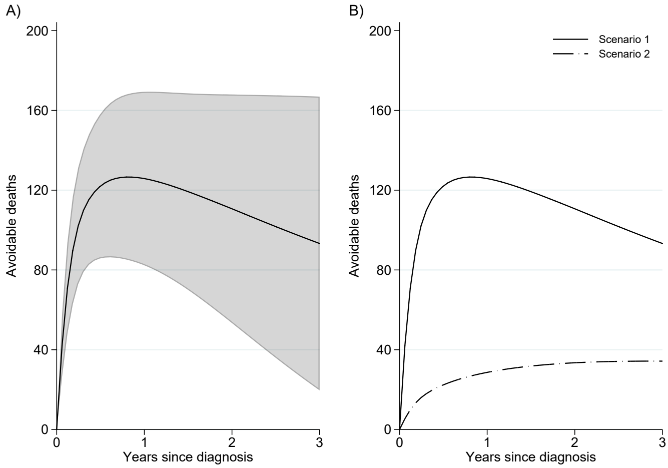
$$D_1(t|X=1,M^1) - D_M(t|X=1,M^0)$$

Avoidable deaths under hypothetical interventions

How many avoidable deaths would be observed if the most deprived patients had the same stage distribution as the least deprived?

```
standsurv, failure timevar(timevar) per(3228)
  at1(dep5 1 stage2 0 ... st4dep5 0, atif(dep5==1) atindweights(p11))
  at2(dep5 1 stage2 1 ... st4dep5 0, atif(dep5==1) atindweights(p12))
  at3(dep5 1 stage2 0 ... st4dep5 0, atif(dep5==1) atindweights(p13))
  at4(dep5 1 stage2 0 ... st4dep5 1, atif(dep5==1) atindweights(p14))
  at5(dep5 1 stage2 0 ... st4dep5 0, atif(dep5==1) atindweights(p01))
  at6(dep5 1 stage2 1 ... st4dep5 0, atif(dep5==1) atindweights(p02))
  at7(dep5 1 stage2 0 ... st4dep5 0, atif(dep5==1) atindweights(p03))
  at8(dep5 1 stage2 0 ... st4dep5 1, atif(dep5==1) atindweights(p04))
  lincom(1 1 1 1 -1 -1 -1 -1) lincomvar(AD)
  expsurv(using(popmort.dta)
    datediag(dx) agediag(agediag) pmrate(rate) pmage(age) pmyear(year) pmother(dep sex)
    at1(dep 5) at2(dep 5)
    at3(dep 5) at4(dep 5)
    at5(dep 5) at6(dep 5)
    at7(dep 5) at8(dep 5))
```

Avoidable deaths for colon cancer



*Out of 3228 patients (N^*) from the most deprived group diagnosed in 2013 the most recent year in our data.

Conclusions

- Mediation analysis within the relative survival framework allows to focus on cancer-related factors.
- Need to be careful when interpreting the results as a number of assumption need to hold:
    - Well-defined interventions assumption is probably violated but quantifying the impact of such a conceptual intervention in a formalised causal framework gives a firm basis to improve our understanding on cancer disparities.
    - Achieving conditional exchangeability for the other cause mortality depends on the availability of relevant life tables.
- Marginal estimates can also obtained with IPW or doubly robust standardisation (future work).

Selected References I

Glymour MM and Spiegelman D

Evaluating Public Health Interventions: 5. Causal Inference in Public Health Research—Do Sex, Race, and Biological Factors Cause Health Outcomes?

Am J Public Health, 107(1):81–85, 2017.

Pohar Perme M, Stare J & Estève J

On Estimation in Relative Survival.

Biometrics,, 68, 113-120, 2012.

Royston P and Lambert PC

Flexible parametric survival analysis in Stata: Beyond the Cox model.

Stata Press, 2011.

Sjölander A

Regression standardization with the R package stdReg.

Eur J Epidemiol, 31(6):563-574, 2016.

VanderWeele TJ

Causal mediation analysis with survival data.

Epidemiology, 22(4):582-585, 2011.