An introduction to relative survival analysis using R

confeRence 2020

Tengku Muhd Hanis Mokhtar

PhD student, USM

November22, 2020

About myself

Background:

- PhD student in Department of Community Medicine, USM
- MSc (Medical Statistics) from USM, 2019
- MBBCh from Al-Azhar University, 2015

Interest:

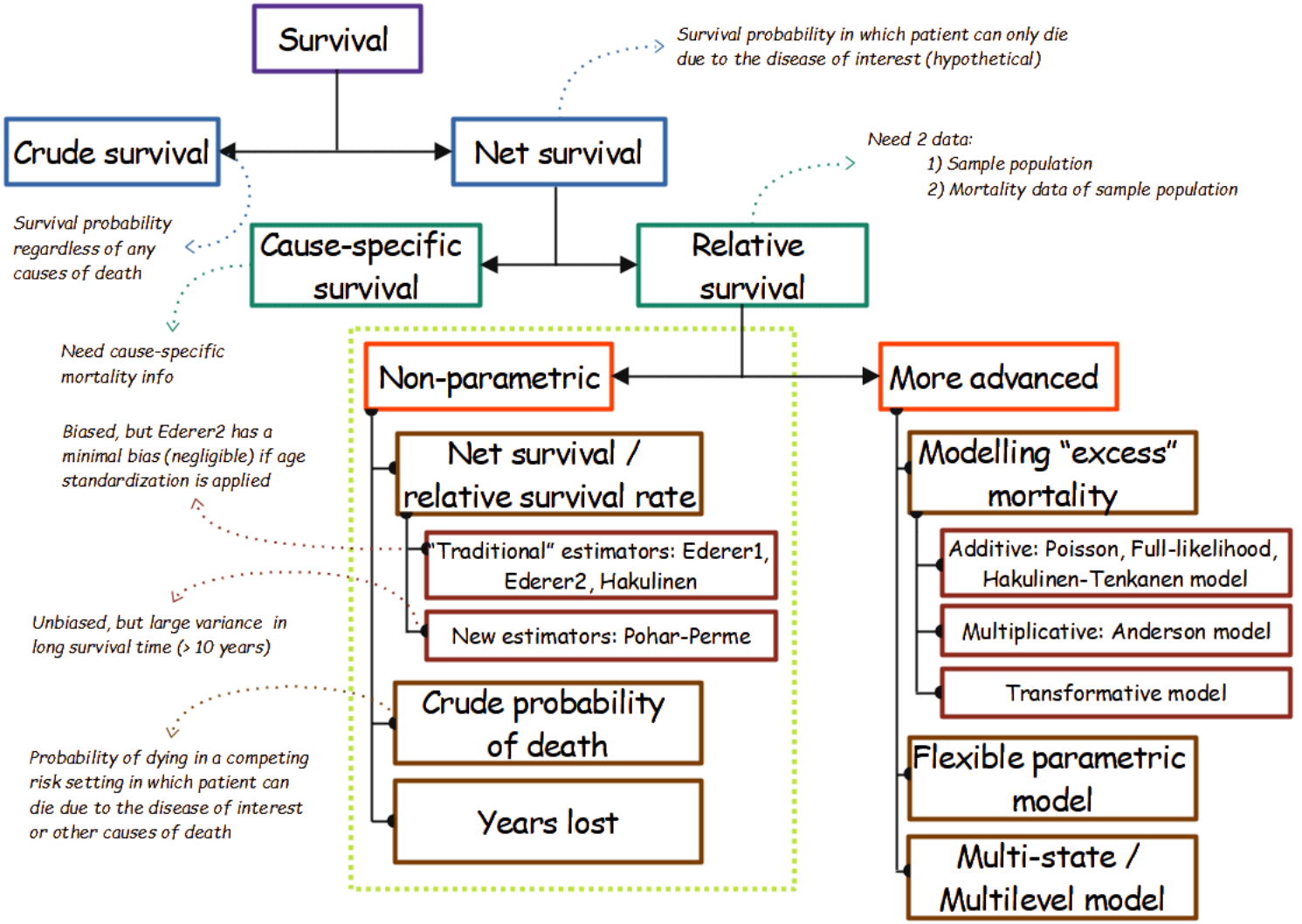
- Medical statistics, population-based study
- Machine learning application in medical sciences
- Application of R (and Python) in medical data

Contact me:

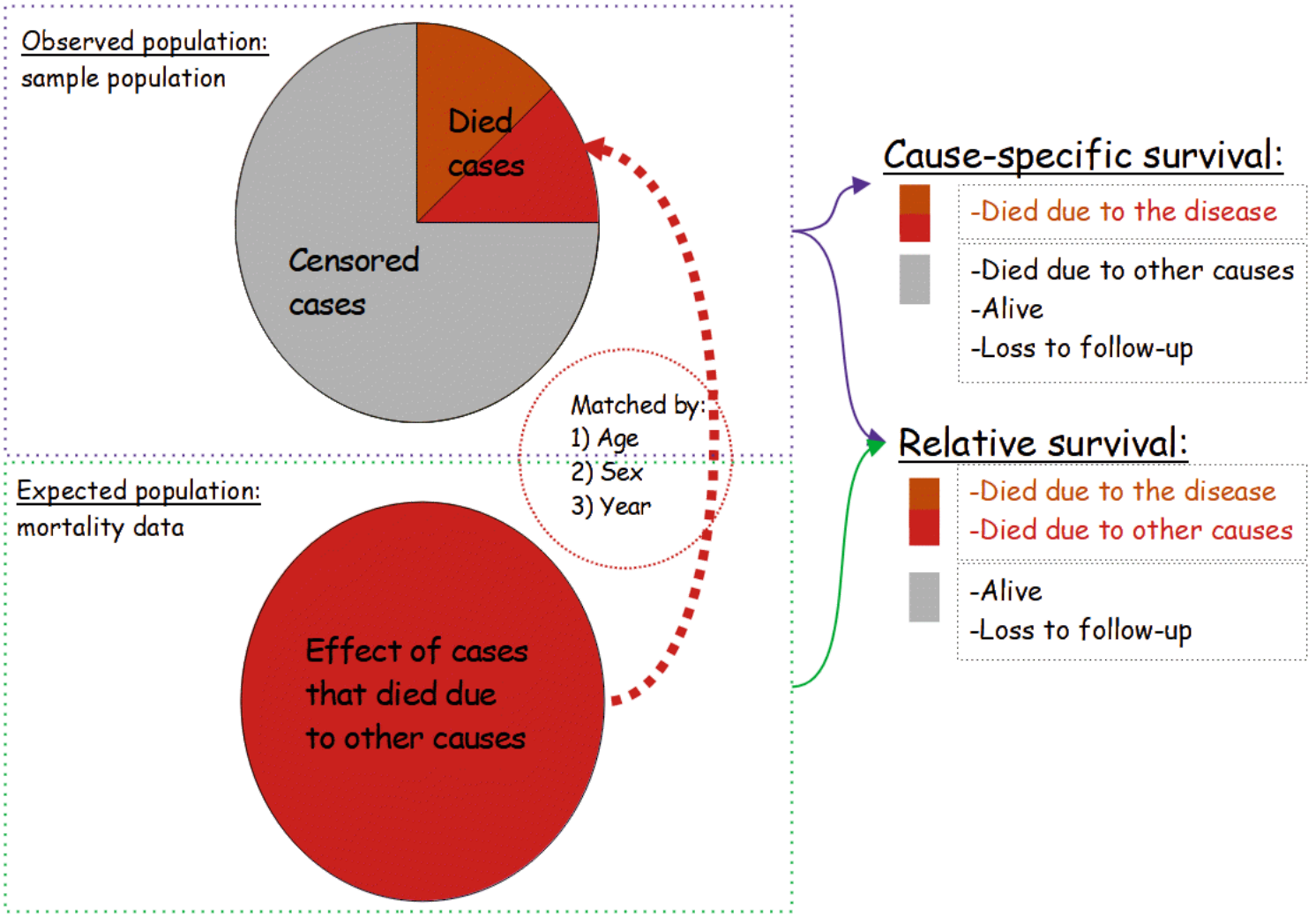
- tengkuhanismokhtar@gmail.com
- Linkedin: Tengku Muhammad Hanis
- Website: https://tengkuhanis.netlify.app/

Download material: https://is.gd/xDWRVn

Survival framework



The difference (ie. study design)



Relative survival

- General idea of relative survival (S_r):

$$S_r = \frac{S_o}{S_e}$$

S_o denotes overall survival in the observed population
S_e denotes overall survival in the expected population

- Relative survival rates (S_r) is the ratio of overall survival of an observed population to the overall survival of an expected population in which there is no event of interest

- We can summarise relative survival (though mathematically incorrect!) as:

$$\text{Relative survival} = \text{Observed survival} - \text{Expected survival}$$

Why and when?

Cause-specific survival VS relative survival

- Choose based on **data availability**
- **Population-based study**, should use relative survival
- Relative survival is better for **comparison** between populations and subpopulations
- Slight misclassification of death lead to **large bias** (in cause-specific survival)
- The use of different population mortality data lead to a **minor change** of survival estimate (in relative survival)

Application to colrec data

1. Expanding abridged life table to a complete life table

2. Convert expanded complete life table into a rate table

    - Use R (MortalityLaws package)
    - Use MortPak software

3. Application of relative survival analysis (non-parametric):

    - Estimation of the net survival or relative survival rate

        - Assumptions (theoretical):
            1. Independence between mortality due to the disease of interest and mortality due to other causes
            2. Comparability of the observed and expected population

    - Crude probability of death:

    - Expected number of years lost due to the disease

Example of abridged life table

Year	Age	mx	qx	ax	lx	dx	Lx	Tx	ex
1983	0	0.01429	0.01411	0.12	100000	1411	98759	6677791	66.78
1983	1-4	0.00066	0.00262	1.74	98589	258	393771	6579032	66.73
1983	5-9	0.00036	0.00182	2.66	98331	179	491234	6185261	62.90
1983	10-14	0.00026	0.00132	3.10	98152	129	490512	5694027	58.01
1983	15-19	0.00104	0.00518	2.94	98022	508	489063	5203515	53.09
1983	20-24	0.00155	0.00773	2.26	97515	754	485509	4714452	48.35

Example of a complete life table

Year	Age	mx	qx	ax	lx	dx	Lx	Tx	ex
1983	0	0.01429	0.01411	0.12	100000	1411	98759	6677791	66.78
1983	1	0.00109	0.00109	0.50	98589	107	98535	6579032	66.73
1983	2	0.00033	0.00033	0.50	98482	32	98465	6480497	65.80
1983	3	0.00070	0.00070	0.50	98449	69	98415	6382031	64.83
1983	4	0.00051	0.00051	0.50	98380	50	98355	6283617	63.87
1983	5	0.00013	0.00013	0.50	98331	13	98324	6185261	62.90

Part 1: Expand abridged life table

- Package: MortalityLaws
- The 5x1 abridged life table for Slovenia male and female was downloaded from human mortality database website (https://www.mortality.org/)

For male

```
## Expand male
age_int <- c(0, 1, seq(5, 110, by = 5))  # age interval in abridged life table
age_range <- 0:110  # range of age to be expanded

# filter 1994-2005
slovenia_male$Year <- as.factor(slovenia_male$Year)
by_yearM <- slovenia_male %>% filter(Year %in% 1994:2005) %>% group_by(Year) %>%
    nest()

# Separate mx in list
mx_male <- vector("list", 0)
for (i in seq_along(by_yearM$data)) {
    mx_male[[i]] <- by_yearM$data[[i]]$mx
}
```

Mortality law

- Parametric function that describes the dying-out process of individuals in a population during a significant portion of their life spans
- Which to choose depend on the literature or the expert opinion
- Use availableLaws() to see more laws

YEAR	NAME	MODEL
1871	Thiele	mu[x] = A exp(-Bx) + C exp[-.5D (x-E)^2] + F exp(Gx)
1883	Wittstein	q[x] = (1/B) A^-[(Bx)^N] + A^-[(M-x)^N]
1979	Siler	mu[x] = A exp(-Bx) + C + D exp(Ex)
1980	Heligman-Pollard	q[x]/p[x] = A^[(x + B)^C] + D exp[-E log(x/F)^2] + G H^x
1980	Heligman-Pollard	q[x] = A^[(x + B)^C] + D exp[-E log(x/F)^2] + GH^x / [1 + GH^x]
1980	Heligman-Pollard	q[x] = A^[(x + B)^C] + D exp[-E log(x/F)^2] + GH^x / [1 + KGH^x]
1980	Heligman-Pollard	q[x] = A^[(x + B)^C] + D exp[-E log(x/F)^2] + GH^(x^K) / [1 + GH^(x^K)]

```r
# Estimate coeffcient
models_male <- vector("list", 0)
for (i in seq_along(mx_male)) {
    models_male[[i]] <- MortalityLaw(age_int, mx = mx_male[[i]], law = "siler", op
}

# Expand life table
male_1994_2005 <- vector("list", 0)
for (i in seq_along(models_male)) {
    male_1994_2005[[i]] <- LawTable(age_range, par = models_male[[i]]$coefficients
        law = "siler", sex = "male")
}

# Combine life table into data frame
male_list <- vector("list", 0)
for (i in seq_along(male_1994_2005)) {
    male_list[[i]] <- data.frame(male_1994_2005[[i]]$lt)
}

male_lt <- male_list %>% enframe() %>% unnest(cols = value)
```

For female

```
## Expand female age interval and range of age as male

# filter 1994-2005
slovenia_female$Year <- as.factor(slovenia_female$Year)
by_yearF <- slovenia_female %>% filter(Year %in% 1994:2005) %>% group_by(Year) %>%
    nest()

# Separate mx in list
mx_female <- vector("list", 0)
for (i in seq_along(by_yearM$data)) {
    mx_female[[i]] <- by_yearM$data[[i]]$mx
}

# Estimate coeffcient
models_female <- vector("list", 0)
for (i in seq_along(mx_female)) {
    models_female[[i]] <- MortalityLaw(age_int, mx = mx_female[[i]], law = "siler"
        opt.method = "LF2")
}
```

```r
# Expand life table
female_1994_2005 <- vector("list", 0)
for (i in seq_along(models_female)) {
    female_1994_2005[[i]] <- LawTable(age_range, par = models_female[[i]]$coeffici
        law = "siler", sex = "female")
}

# Combine life table into data frame
female_list <- vector("list", 0)
for (i in seq_along(female_1994_2005)) {
    female_list[[i]] <- data.frame(female_1994_2005[[i]]$lt)
}

female_lt <- female_list %>% enframe() %>% unnest(cols = value)
```

Part 2: Make a rate table

- Package: relsurv
- We need survival probability (px) for a rate table:

$$px = 1 - qx$$

```
## Make a rate table select age(x) and probability of dying(qx) px (survival
## probability) = 1 - qx
pop_m <- male_lt %>% mutate(year = rep(1994:2005, each = 111), px = 1 - qx) %>% se
    year, px)
pop_f <- female_lt %>% mutate(year = rep(1994:2005, each = 111), px = 1 - qx) %>%
    select(x, year, px)

# long -> wide use px (survival probability)
pop_m.w <- pivot_wider(pop_m, names_from = year, values_from = px)
pop_f.w <- pivot_wider(pop_f, names_from = year, values_from = px)
```

```
pop_m.w[1:4, 1:9]
```

```
## # A tibble: 4 x 9
##       x `1994` `1995` `1996` `1997` `1998` `1999` `2000` `2001`
##   <int>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1     0  0.993  0.994  0.994  0.994  0.994  0.995  0.994  0.995
## 2     1  1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
## 3     2  1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
## 4     3  1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
```

```r
# delete age column (x)
pop_m.w$x <- NULL
pop_f.w$x <- NULL

# as matrix
pop_fwm <- as.matrix(pop_f.w)
pop_mwm <- as.matrix(pop_m.w)
str(pop_fwm)
```

```
##  num [1:111, 1:12] 0.993 1 1 1 1 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr [1:12] "1994" "1995" "1996" "1997" ...
```

```r
# ratetable
pop_rate <- transrate(men = pop_mwm, women = pop_fwm, yearlim = c(1994, 2005), int
is.ratetable(pop_rate)
```

```
## [1] TRUE
```

```r
summary(pop_rate)
```

```
##  Rate table with 3 dimensions:
##     age ranges from 0 to 40176.51; with 111 categories
##     sex has levels of: male female
##     year ranges from 12419 to 16437; with 12 categories
```

Part 3: Non-parametric relative survival

Package: relsurv

colrec data:

- Data in relsurv package provided by Slovene Cancer Registry
- Survival of patients with colon and rectal cancer diagnosed in 1994-2000.
- Format
  - A data frame with 5971 observations on the following 7 variables:
  - Sex: sex (1=male, 2=female)
  - Age*: age (in days)
  - Diag*: date of diagnosis (in date format)
  - Time*: survival time (in days)
  - Stat: censoring indicator (0=censoring, 1=death)
  - Stage: cancer stage (Values 1-3, code 99 stands for unknown)
  - Site: cancer site

**variables are randomly perturbed to make the identification of patients impossible*

Packages

```
library(relsurv)
library(survminer)
```

We are going to edit data, so to limit follow-up time to 5 years only

```
# limit follow-up time to 5 years
colrec$end <- colrec$diag + colrec$time
# recode end time
colrec$end2 <- ifelse(colrec$end > as.date("31Dec2005", order = "dmy"), as.date("3
    order = "dmy"), colrec$end)
# recode the status
colrec$stat2 <- ifelse(colrec$end > as.date("31Dec2005", order = "dmy"), 0, colrec
# edit follow-up time
colrec$time2 <- colrec$end2 - colrec$diag
```

Fit Pohar-perme

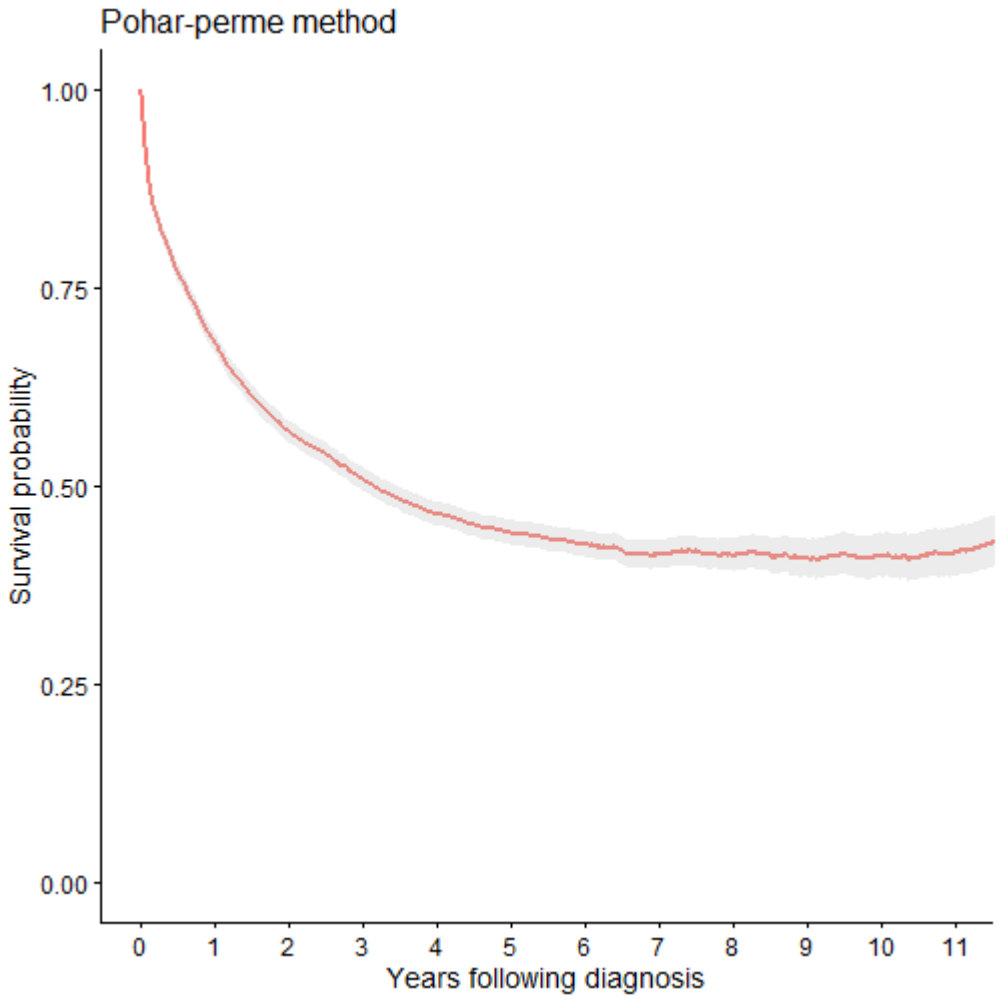
```
rs_PP <- rs.surv(Surv(time2, stat2) ~ 1,
                 rmap = list(age = age, sex = sex, year = diag),
                 method = "pohar-perme",
                 ratetable = slopop,
                 data = colrec)

summary(rs_PP, times = 1:5 * 365.241)
```

```
## Call: rs.surv(formula = Surv(time2, stat2) ~ 1, data = colrec, ratetable = slopop,
##     method = "pohar-perme", rmap = list(age = age, sex = sex,
##         year = diag))
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   365   3919    2048    0.682 0.00641        0.669        0.695
##   730   3144     774    0.568 0.00704        0.554        0.582
##  1096   2715     429    0.509 0.00738        0.494        0.523
##  1461   2387     328    0.466 0.00764        0.451        0.481
##  1826   2163     224    0.441 0.00791        0.426        0.457
```

We are going to use survminer package to plot instead the base R function

```
ggsurvplot(rs_PP, conf.int = T, xscale = "d_y", break.x.by = 365.24, xlab = "Years
    title = "Pohar-perme method", censor = F, legend = "none")
```

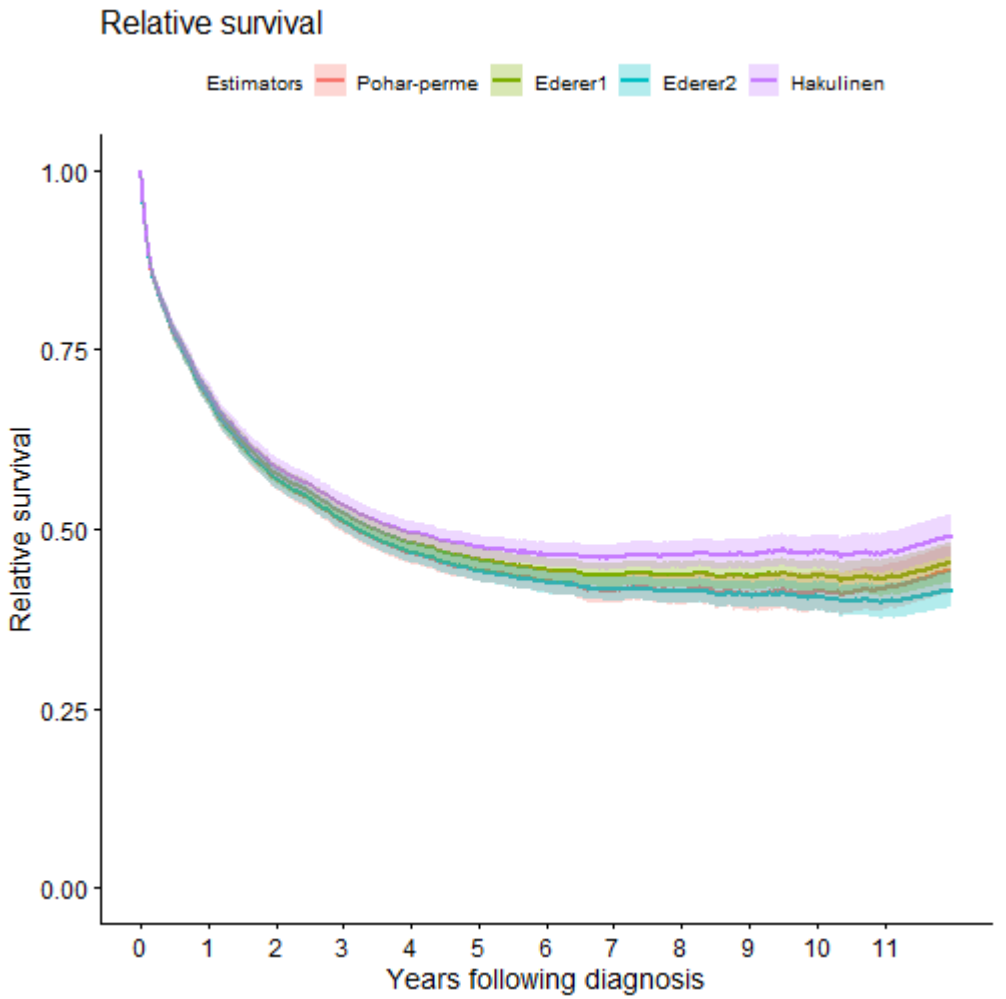


Fit other estimators

```
# Edere 1
rs_e1 <- rs.surv(Surv(time2, stat2) ~ 1, rmap = list(age = age, sex = sex, year =
    method = "ederer1", ratetable = slopop, data = colrec)
# Ederer2
rs_e2 <- rs.surv(Surv(time2, stat2) ~ 1, rmap = list(age = age, sex = sex, year =
    method = "ederer2", ratetable = slopop, data = colrec)
# Hakulinen
rs_h <- rs.surv(Surv(time2, stat2) ~ 1, rmap = list(age = age, sex = sex, year = d
    method = "hakulinen", ratetable = slopop, data = colrec)
```

Compare all estimators

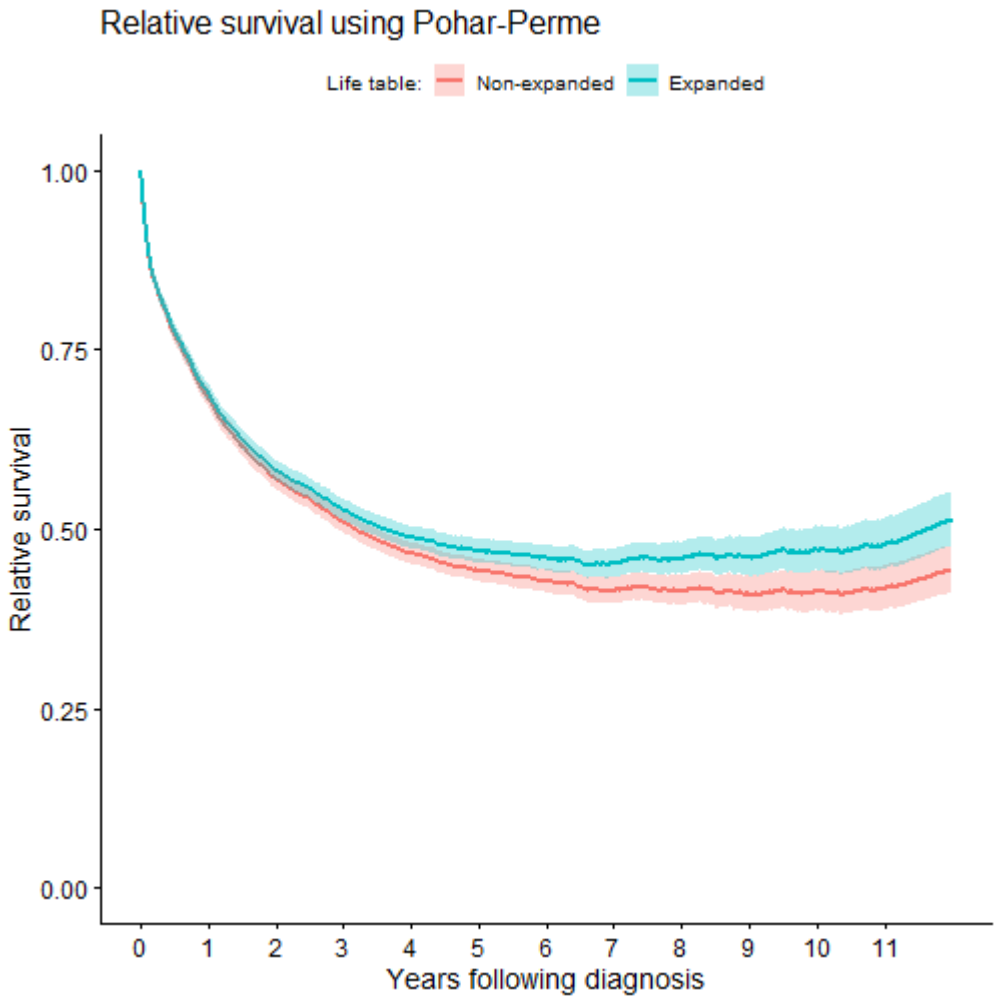
```
rs_list <- list(`Pohar-perme` = rs_PP, Ederer1 = rs_e1, Ederer2 = rs_e2, Hakulinen
ggsurvplot(rs_list, data = colrec, conf.int = T, censor = F, combine = T, xscale =
    break.x.by = 365.24, xlab = "Years following diagnosis", ylab = "Relative surv
    title = "Relative survival", legend = "top", legend.title = "Estimators", lege
        "Ederer1", "Ederer2", "Hakulinen"))
```


Relative survival
Estimators
Pohar-perme
Ederer1
Ederer2
Hakulinen
1.00
0.75
0.50
0.25
0.00
Relative survival
0
1
2
3
4
5
6
7
8
9
10
11
Years following diagnosis

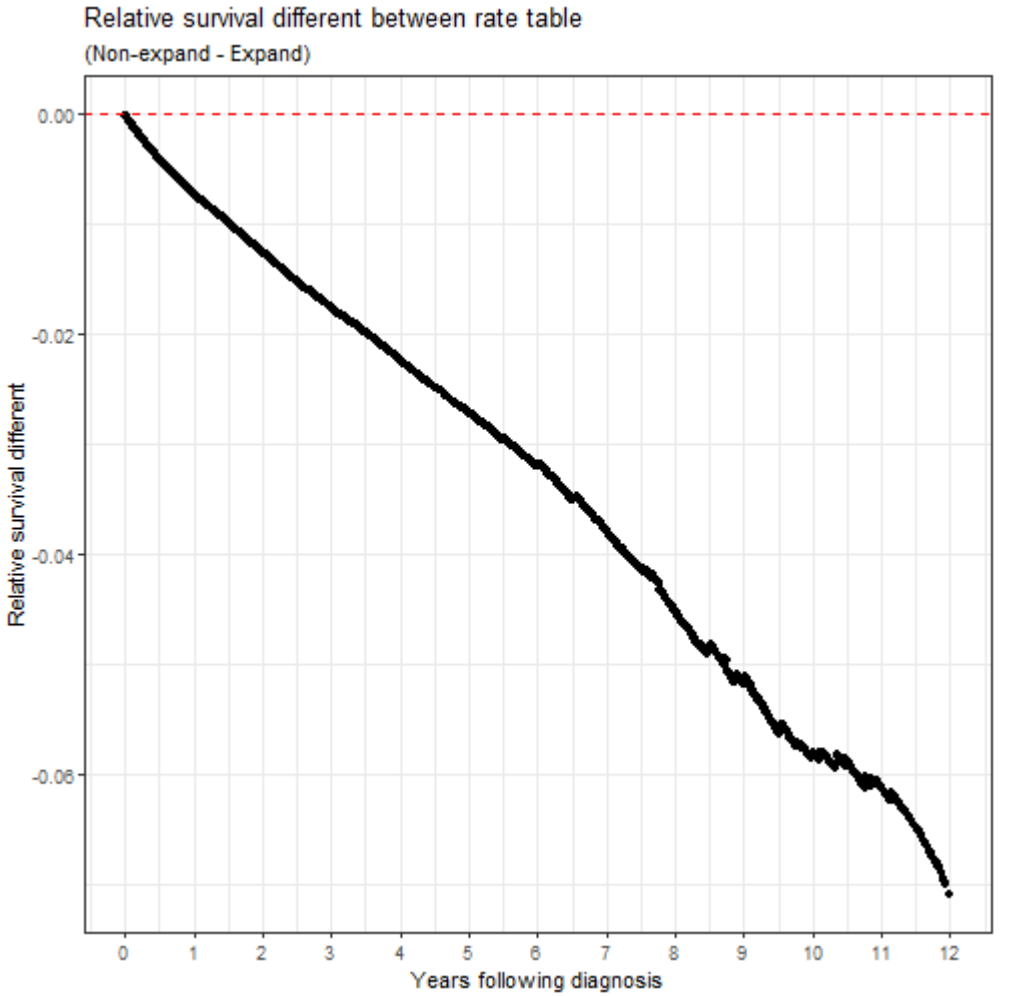
Compare life table

```r
# Pohar-perme with expanded life table
rs_PP_expand <- rs.surv(Surv(time2, stat2) ~ 1, rmap = list(age = age, sex = sex,
    year = diag), method = "pohar-perme", ratetable = pop_rate, data = colrec)

# Compare plot
compare_lt <- list(`Non-expanded` = rs_PP, Expanded = rs_PP_expand)
ggsurvplot(compare_lt, data = colrec, conf.int = T, combine = T, censor = F, xscal
    break.x.by = 365.24, xlab = "Years following diagnosis", ylab = "Relative surv
    title = "Relative survival using Pohar-Perme", legend = "top", legend.title =
    legend.labs = c("Non-expanded", "Expanded"))
```


Relative survival using Pohar-Perme
Life table:
Non-expanded
Expanded
1.00
0.75
0.50
0.25
0.00
Relative survival
0
1
2
3
4
5
6
7
8
9
10
11
Years following diagnosis

Differences between relative survival estimate of expanded and non-expanded life table

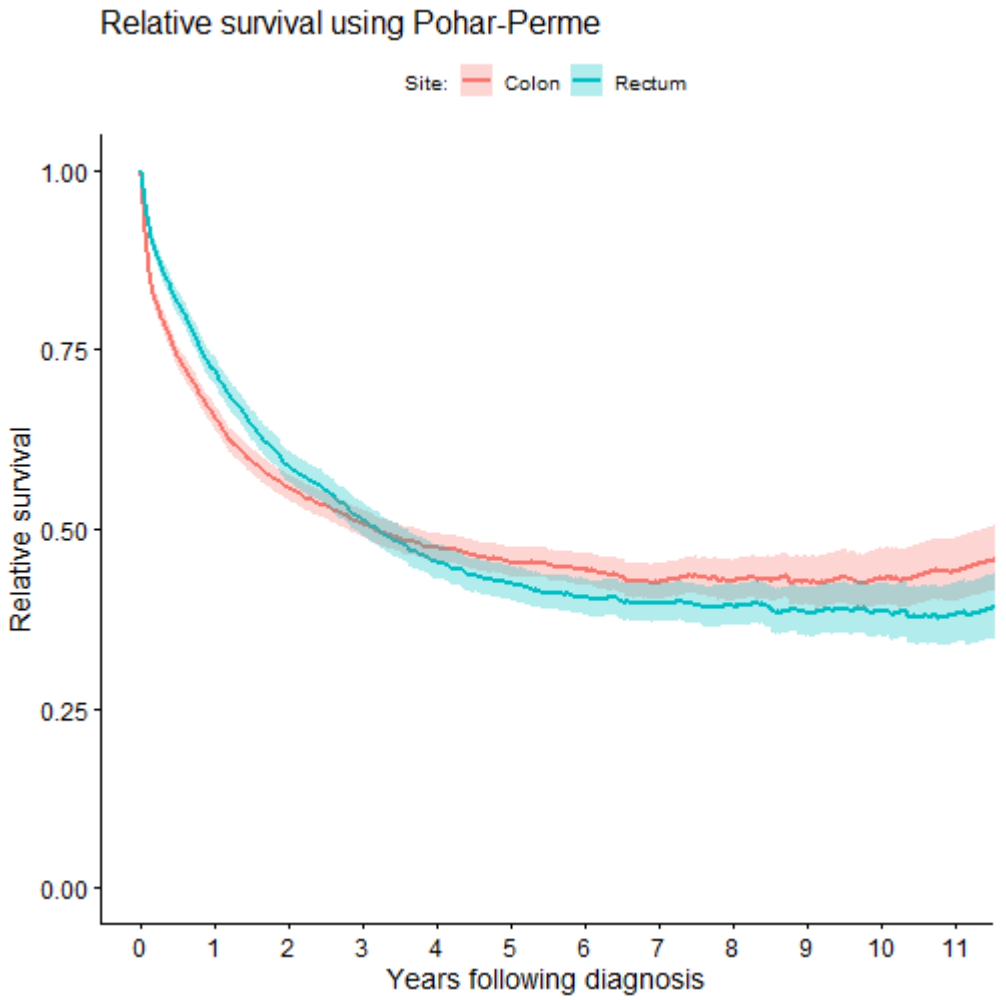


Log-rank type test for comparison of net survival curves

```
diff_site <- rs.diff(Surv(time2, stat2) ~ site, rmap = list(age = age, sex = sex,
    year = diag), data = colrec, ratetable = slopop)
diff_site
```

```
## Value of test statistic: 0.2355707
## Degrees of freedom: 1
## P value: 0.6274237
```

```
diff_rs <- rs.surv(Surv(time2, stat2) ~ site, rmap = list(age = age, sex = sex, ye
    data = colrec, ratetable = slopop)
ggsurvplot(diff_rs, data = colrec, conf.int = T, combine = T, censor = F, xscale =
    break.x.by = 365.24, xlab = "Years following diagnosis", ylab = "Relative surv
    title = "Relative survival using Pohar-Perme", legend = "top", legend.title =
    legend.labs = c("Colon", "Rectum"))
```


Relative survival using Pohar-Perme
Site: Colon Rectum
1.00
0.75
0.50
0.25
0.00
Relative survival
0 1 2 3 4 5 6 7 8 9 10 11
Years following diagnosis

Crude probability of death and year lost

```
cpdeath <- cmp.rel(Surv(time2, stat2) ~ site, rmap = list(age = age, sex = sex, ye
    ratetable = slopop, data = colrec, tau = 3652.41)
# tau = 3652.41, value after 10 years is censored
summary(cpdeath, times = c(1, 5, 10), scale = 365.24)$est  #scale default 1:day
```

```
##                                    1         5        10
## causeSpec siterectum=0  0.34003855 0.5269945 0.5445191
## population siterectum=0 0.02916328 0.1036957 0.1784933
## causeSpec siterectum=1  0.27453284 0.5438010 0.5843705
## population siterectum=1 0.03084100 0.1033123 0.1678205
```

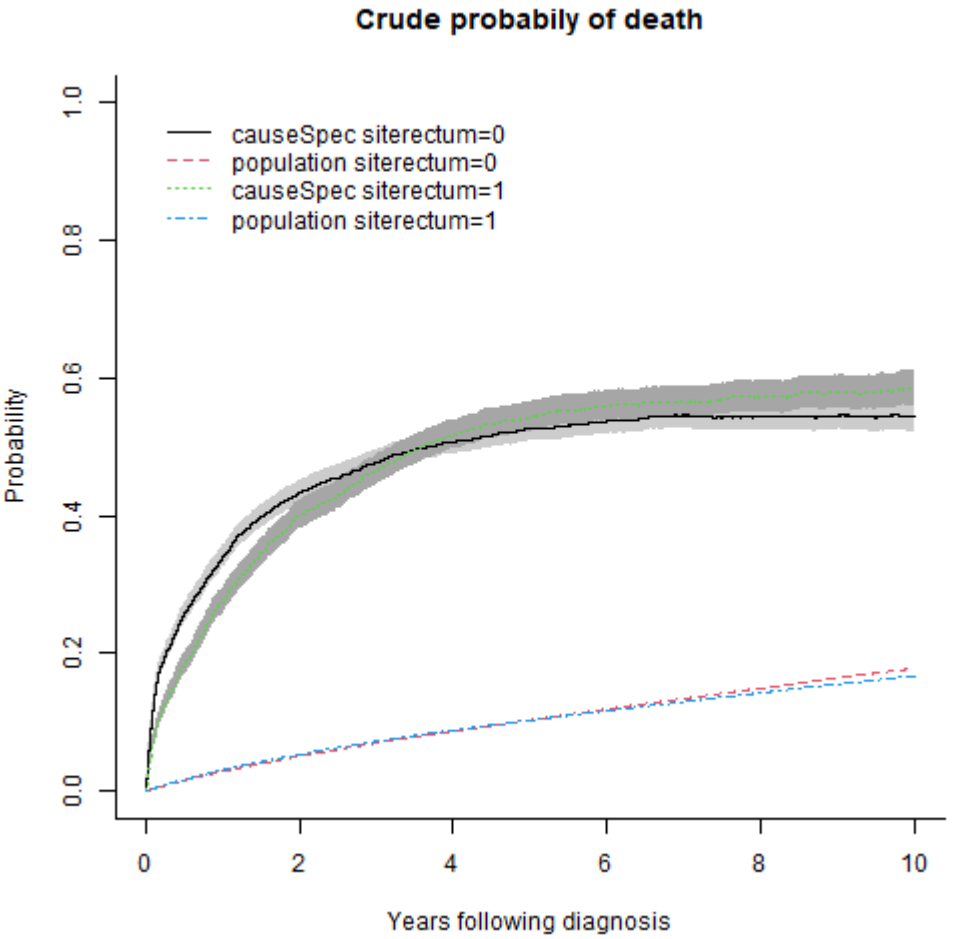
- Within 10 years after diagnosis 58.4% of patients have died due to rectal cancer and 54.5% of patients have died due to colon cancer

```
cpdeath
```

```
## Estimates, variances and area under the curves:
## $est
##                                    2          4         6         8
## causeSpec siterectum=0  0.43423535 0.50738517 0.5377674 0.5447054
## population siterectum=0 0.05099486 0.08736450 0.1195701 0.1497519
## causeSpec siterectum=1  0.40139481 0.51857321 0.5587463 0.5731538
## population siterectum=1 0.05364613 0.08864497 0.1169295 0.1431313
##
## $var
##                                      2            4            6            8
## causeSpec siterectum=0  7.855896e-05 8.542588e-05 9.140629e-05 1.040238e-04
## population siterectum=0 2.756566e-07 1.236647e-06 2.988433e-06 5.658865e-06
## causeSpec siterectum=1  1.117243e-04 1.193491e-04 1.245047e-04 1.389559e-04
## population siterectum=1 3.178646e-07 1.454460e-06 3.461529e-06 6.550556e-06
##
## $area
##                            Area at tau = 10
## causeSpec siterectum=0            4.8086656
## population siterectum=0           1.0017281
## causeSpec siterectum=1            4.8209210
## population siterectum=1           0.9814313
```

- area = year lost due to the disease
- Patients with rectal cancer lost 4.8 years due to the cancer
- Patients with colon cancer lost 4.8 years due to the cancer

```
plot(cpdeath, xscale = 365.24, col = 1:4, conf.int = c(1, 3), xlab = "Years follow
    main = "Crude probabily of death")
```



References

1. Mariotto, A.B.; Noone, A.M.; Howlader, N.; Cho, H.; Keel, G.E.; Garshell, J.; Woloshin, S.; Schwartz, L.M. Cancer survival: An overview of measures, uses, and interpretation. J. Natl. Cancer Inst. - Monogr. 2014, 2014, 145-186, doi:10.1093/jncimonographs/lgu024.
2. Lambert, P.C.; Dickman, P.W.; Rutherford, M.J. Comparison of different approaches to estimating age standardized net survival. BMC Med. Res. Methodol. 2015, 15, 1-13, doi:10.1186/s12874-015-0057-3.
3. Roche, L.; Danieli, C.; Belot, A.; Grosclaude, P.; Bouvier, A.M.; Velten, M.; Iwaz, J.; Remontet, L.; Bossard, N. Cancer net survival on registry data: Use of the new unbiased Pohar-Perme estimator and magnitude of the bias with the classical methods. Int. J. Cancer 2013, 132, 2359-2369, doi:10.1002/ijc.27830.
4. UKIACR Standard Operating Procedure: Guidelines on Population Based Cancer Survival Analysis; 2016;
5. Pohar, M.; Stare, J. Relative survival analysis in R. Comput. Methods Programs Biomed. 2006, 81, 272-278, doi:10.1016/j.cmpb.2006.01.004.
6. Perme, M.P.; Pavlic, K. Nonparametric Relative Survival Analysis with the R Package relsurv. J. Stat. Softw. 2018, 87, doi:10.18637/jss.v087.i08.
7. Sarfati, D.; Blakely, T.; Pearce, N. Measuring cancer survival in populations: Relative survival vs cancer-specific survival. Int. J. Epidemiol. 2010, 39, 598-610, doi:10.1093/ije/dyp392.

Thanks!

Slides created via the R package xaringan