BIOSTAT III: Survival Analysis for Epidemiologists in R: Take-home examination

Mark Clements

6–15 November, 2023

Contents

Instructions

- **The examination is individual-based: you are not allowed to cooperate with anyone**, although you are encouraged to consult the available literature. The examiner will use Urkund in order to assess potential plagiarism.
- The examination will be made available by noon on Wednesday 15 November 2023 and **the examination is due by 17:00 on Wednesday 22 November 2023**.
- The examination will be graded and results returned to you by Wednesday 29 November 2023.
- The examination is in two parts. To pass the examination, you need to score at least 7/12 for Part 1 focused on rates and general regression modelling and 11/19 for Part 2 on survival analysis.
- Do not write answers by hand: please use Word, LaTeX, Markdown or a similar format for your examination report and submit the report **as a PDF file**.
- Motivate all answers in your examination report. Define any notation that you use for equations. The examination report should be written in English.
- Email the examination report containing the answers **as a PDF file** to gunilla.nilsson.roos@ki.se. **Write your name in the email, but do NOT write your name or otherwise reveal your identity in the document containing the answers.**

Part 1

The **eha** package on CRAN includes data-frames `swepop` and `swedeaths`, which are the person-years and all deaths in Sweden by age, sex and calendar year. Below, we combine these two data-frames to create a `mort` data-frame with the following variables

- Sex, which is encoded using the variable `sex`, which is `"men"` for males and `"women"` for females
- Attained age, which is encoded in the variable `age` for single year of age for 0–99 and then 100 years and over.
- Calendar period, which is encoded in the variable `year` for single calendar year for 1969–2020.

- Person-years, encoded in the variable `pop`, which was calculated using the mid-year population.
- Number of deaths due to all causes, encoded in the variable `deaths`.

In the following, we restrict our analysis to adults aged 18 years and over (i) to simplify the modelling for age and (ii) to give a more appropriate labelling for "men" and "women". We have also removed the `id` variable by setting it to `NULL`, as it is not used.

```
library(eha)
mort = transform(swepop, deaths=swedeaths$deaths, id=NULL)

str(mort)

'data.frame': 10504 obs. of  5 variables:
 $ age   : int  0 0 1 1 2 2 3 3 4 4 ...
 $ sex   : Factor w/ 2 levels "women","men": 1 2 1 2 1 2 1 2 1 2 ...
 $ year  : int  1969 1969 1969 1969 1969 1969 1969 1969 1969 1969 ...
 $ pop   : num  52673 55728 56831 59924 58994 ...
 $ deaths: int  491 773 33 45 22 43 25 30 25 46 ...
```

Q1

(a) We fit a Poisson regression model for all cause mortality rates in Sweden for ages 18 years and over, with a continuous main effect for year minus 2000, an indicator for age>=65 and an indicator for being male. Write a formula for this regression model. As a reminder, please define your notation. (2 pts)

```
fit <- glm(deaths~I(year-2000)+I(age>=65)+I(sex=="men")+offset(log(pop)),
   data=mort, subset=(age>=18), family=poisson)
summary(fit)


Call:
glm(formula = deaths ~ I(year - 2000) + I(age >= 65) + I(sex ==
    "men") + offset(log(pop)), family = poisson, data = mort,
    subset = (age >= 18))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-52.729  -11.563   -2.643   17.708   62.510

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -6.113e+00  1.295e-03 -4720.5   <2e-16 ***
I(year - 2000)      -7.902e-03  3.071e-05  -257.3   <2e-16 ***
I(age >= 65)TRUE     2.961e+00  1.272e-03  2327.4   <2e-16 ***
I(sex == "men")TRUE  2.511e-01  9.243e-04   271.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 12311058  on 8631  degrees of freedom
Residual deviance:  3983999  on 8628  degrees of freedom
AIC: 4046824

Number of Fisher Scoring iterations: 5
```

(b) From the regression model output, show how to numerically calculate the mortality rate ratio for men compared with women, including the 95% confidence interval. (2 pts)

(c) From the regression model output, explain how to interpret the effects for `I(year - 2000)` and `I(age >= 65)TRUE`. (2 pts)

(d) From the regression model output, show how to numerically calculate the predicted rate for females aged 18–64 years in the year 2000, including the 95% confidence interval. (2 pts)

(e) We now extend the regression model output to include an interaction between `I(age>=65)` and `I(sex=="male")` (see R code and output below). Explain how to interpret the main effect for `I(sex=="male")` for this model. (2 pts)

```
fit <- glm(deaths~I(year-2000)+I(age>=65)*I(sex=="men")+offset(log(pop)),
   data=mort, subset=(age>=18), family=poisson)
summary(fit)


Call:
glm(formula = deaths ~ I(year - 2000) + I(age >= 65) * I(sex ==
    "men") + offset(log(pop)), family = poisson, data = mort,
    subset = (age >= 18))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-51.538  -11.382   -2.192   18.027   61.875

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -6.291e+00  1.953e-03 -3221.6   <2e-16 ***
I(year - 2000)                       -7.888e-03  3.071e-05  -256.8   <2e-16 ***
I(age >= 65)TRUE                      3.165e+00  2.066e-03  1531.8   <2e-16 ***
I(sex == "men")TRUE                   5.458e-01  2.430e-03   224.6   <2e-16 ***
I(age >= 65)TRUE:I(sex == "men")TRUE -3.468e-01  2.629e-03  -131.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 12311058  on 8631  degrees of freedom
Residual deviance:  3966234  on 8627  degrees of freedom
AIC: 4029061

Number of Fisher Scoring iterations: 5
```

(f) Explain how you would assess whether the mortality rate ratio for men compared with women is statistically different between (i) ages 18–64 years and (ii) 65 years and over. From the interaction regression model output, what is the mortality rate ratio for men aged 65 years and over compared with women aged 65 years and over? (2 pts)

Part 2

Q2

We now use data on chemotherapy for stage B/C colon cancer represented in the `survival::colon` dataset. The help page for the dataset is shown below:

```
library(survival)
help("colon", help_type="text", package="survival")
```

```
colon                 package:survival                 R Documentation

Chemotherapy for Stage B/C colon cancer

Description:

     These are data from one of the first successful trials of adjuvant
     chemotherapy for colon cancer. Levamisole is a low-toxicity
     compound previously used to treat worm infestations in animals;
     5-FU is a moderately toxic (as these things go) chemotherapy
     agent. There are two records per person, one for recurrence and
     one for death

Usage:

     colon
            data(cancer, package="survival")

Format:

       id:       id
       study:    1 for all patients
       rx:       Treatment - Obs(ervation), Lev(amisole), Lev(amisole)+5-FU
       sex:      1=male
       age:      in years
       obstruct: obstruction of colon by tumour
       perfor:   perforation of colon
       adhere:   adherence to nearby organs
       nodes:    number of lymph nodes with detectable cancer
       time:     days until event or censoring
       status:   censoring status
       differ:   differentiation of tumour (1=well, 2=moderate, 3=poor)
       extent:   Extent of local spread (1=submucosa, 2=muscle, 3=serosa,
                 4=contiguous structures)
       surg:     time from surgery to registration (0=short, 1=long)
       node4:    more than 4 positive lymph nodes
       etype:    event type: 1=recurrence,2=death

Note:

     The study is originally described in Laurie (1989).  The main
     report is found in Moertel (1990).  This data set is closest to
     that of the final report in Moertel (1991).

References:

     JA Laurie, CG Moertel, TR Fleming, HS Wieand, JE Leigh, J Rubin,
     GW McCormack, JB Gerstner, JE Krook and J Malliard.  Surgical
     adjuvant therapy of large-bowel carcinoma: An evaluation of
     levamisole and the combination of levamisole and fluorouracil: The
     North Central Cancer Treatment Group and the Mayo Clinic.  J
     Clinical Oncology, 7:1447-1456, 1989.

     CG Moertel, TR Fleming, JS MacDonald, DG Haller, JA Laurie, PJ
     Goodman, JS Ungerleider, WA Emerson, DC Tormey, JH Glick, MH
     Veeder and JA Maillard.  Levamisole and fluorouracil for adjuvant
     therapy of resected colon carcinoma. New England J of Medicine,
     332:352-358, 1990.

     CG Moertel, TR Fleming, JS MacDonald, DG Haller, JA Laurie, CM
     Tangen, JS Ungerleider, WA Emerson, DC Tormey, JH Glick, MH Veeder
```

```
and JA Maillard, Fluorouracil plus Levamisole as an effective
adjuvant therapy after resection of stage II colon carcinoma: a
final report.  Annals of Internal Med, 122:321-326, 1991.
```

(a) The Kaplan-Meier estimators for overall survival by tumour differentiation are shown in Figure 1. Carefully describe and interpret the three survival curves. (2 pts)

```
## Colour-blind palette of colours
cbPalette <- c("#999999", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2",
       "#D55E00", "#CC79A7")
sfit = survfit(Surv(time, status)~differ, data=survival::colon, subset=(etype==2))
plot(sfit, xlab="Time since randomisation (days)", ylab="Overall survival",
     col=cbPalette[1:3], lwd=1:3)
legend("topright", paste("differ =", 1:3), col=cbPalette[1:3], lwd=1:3, lty=1, bty="n")
```

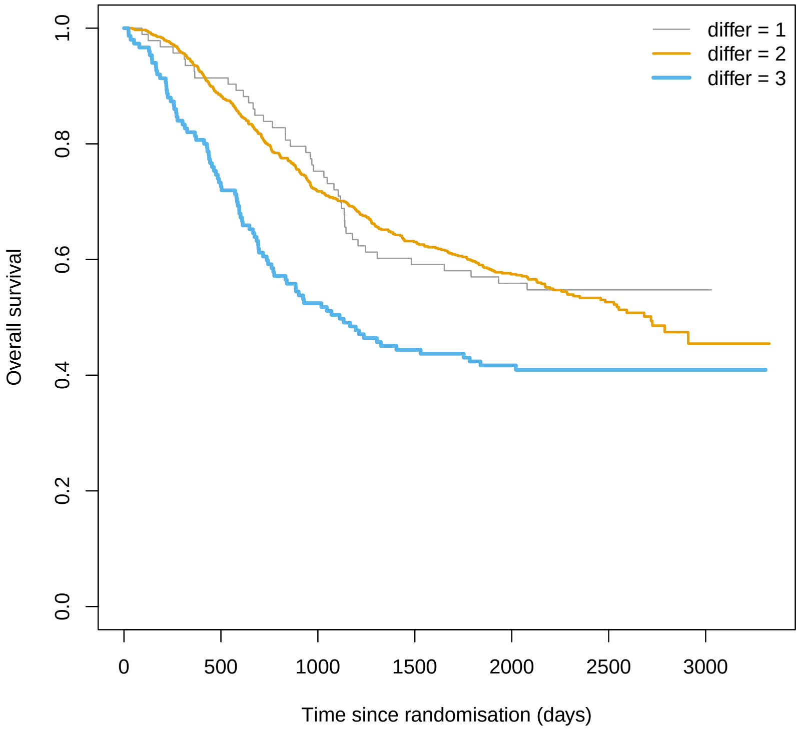

Figure 1: Kaplan-Meier overall survival curves by tumour differentiation

(b) Write out the regression equation for the Cox model specified in the following code. (2 pts)

```
colon2 = transform(subset(survival::colon, etype==2),
    Lev = ifelse(rx=="Lev",1,0),
    Lev_5FU = ifelse(rx=="Lev+5FU",1,0),
    differ2 = ifelse(differ==2,1,0),
    differ3 = ifelse(differ==3,1,0)
    )
fit = coxph(Surv(time,status)~Lev+Lev_5FU+differ2+differ3, data=colon2)
summary(fit)
```

```
Call:
coxph(formula = Surv(time, status) ~ Lev + Lev_5FU + differ2 +
    differ3, data = colon2)

  n= 906, number of events= 441
   (23 observations deleted due to missingness)

            coef exp(coef) se(coef)      z Pr(>|z|)
Lev     -0.02954   0.97089  0.11232 -0.263  0.79254
Lev_5FU -0.37086   0.69014  0.11961 -3.101  0.00193 **
differ2  0.04291   1.04385  0.16474  0.260  0.79448
differ3  0.54099   1.71770  0.18806  2.877  0.00402 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
Lev        0.9709     1.0300    0.7791    1.2100
Lev_5FU    0.6901     1.4490    0.5459    0.8725
differ2    1.0438     0.9580    0.7558    1.4417
differ3    1.7177     0.5822    1.1882    2.4833

Concordance= 0.568  (se = 0.014 )
Likelihood ratio test= 27.06  on 4 df,   p=2e-05
Wald test            = 28.17  on 4 df,   p=1e-05
Score (logrank) test = 28.6  on 4 df,   p=9e-06
```

(c) Based on the previous output, discuss whether there is any evidence that tumour differentiation is associated with overall survival after adjusting for treatment (Lev and $\text{Lev}_{5\text{FU}}$). Provide confidence intervals and p-values to support your argument. (2 pts)

(d) We are interested in whether the effect of treatment on overall survival varies by tumour differentiation. We fit a Cox model that includes main effects for treatment, a main effect for `I(differ==3)`, and interactions between treatment and `I(differ==3)`. Based on the regression model output, summarise and discuss the evidence for whether the treatment effect on overall survival varies by tumour differentiation. (2 pts)

```
fit = coxph(Surv(time,status)~Lev+Lev_5FU+differ3+Lev:differ3+Lev_5FU:differ3, data=colon2)
summary(fit)

Call:
coxph(formula = Surv(time, status) ~ Lev + Lev_5FU + differ3 +
    Lev:differ3 + Lev_5FU:differ3, data = colon2)

  n= 906, number of events= 441
   (23 observations deleted due to missingness)

                    coef exp(coef) se(coef)      z Pr(>|z|)
Lev             -0.02321   0.97705  0.12455 -0.186  0.85215
Lev_5FU         -0.36638   0.69324  0.13482 -2.718  0.00657 **
differ3          0.52500   1.69046  0.19273  2.724  0.00645 **
Lev:differ3     -0.04289   0.95802  0.28638 -0.150  0.88096
Lev_5FU:differ3 -0.02694   0.97342  0.29104 -0.093  0.92625
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                exp(coef) exp(-coef) lower .95 upper .95
Lev                0.9771     1.0235    0.7654    1.2472
Lev_5FU            0.6932     1.4425    0.5323    0.9029
differ3            1.6905     0.5916    1.1587    2.4663
Lev:differ3        0.9580     1.0438    0.5465    1.6794
Lev_5FU:differ3    0.9734     1.0273    0.5503    1.7220

Concordance= 0.57  (se = 0.014 )
Likelihood ratio test= 27.02  on 5 df,   p=6e-05
Wald test            = 28.2  on 5 df,   p=3e-05
Score (logrank) test = 29.13  on 5 df,   p=2e-05
```

(d) We want to assess the whether the hazard ratio comparing `differ==3` with `differ==1` varies with time. We initially fit a flexible parametric survival model without a time-varying effect and then fit a model that includes a time-varying effect. Based on the following code and output, what evidence is there for a time-varying hazard ratio? (2 pts)

```
library(rstpm2)
fit0 = stpm2(Surv(time,status)~Lev+Lev_5FU+differ2+differ3, data=colon2, df=4)
fit1 = stpm2(Surv(time,status)~Lev+Lev_5FU+differ2+differ3, data=colon2, df=4,
     tvc=list(differ3=3))
summary(fit1)
anova(fit0,fit1)

Maximum likelihood estimation

Call:
stpm2(formula = Surv(time, status) ~ Lev + Lev_5FU + differ2 +
    differ3, data = colon2, df = 4, tvc = list(differ3 = 3))

Coefficients:
                                            Estimate Std. Error z value
(Intercept)                                -8.966241   1.210950 -7.4043
Lev                                        -0.022403   0.112351 -0.1994
Lev_5FU                                    -0.355785   0.119577 -2.9754
differ2                                     0.042370   0.164733  0.2572
differ3                                     3.905793   1.392880  2.8041
nsx(log(time), df = 4)1                     7.587702   1.199364  6.3264
nsx(log(time), df = 4)2                     6.408176   0.768425  8.3394
nsx(log(time), df = 4)3                    13.364836   2.326346  5.7450
nsx(log(time), df = 4)4                     5.755228   0.535244 10.7525
as.numeric(differ3):nsx(log(time), df = 3)1 -2.217975   0.884171 -2.5085
as.numeric(differ3):nsx(log(time), df = 3)2 -5.912347   2.702843 -2.1875
as.numeric(differ3):nsx(log(time), df = 3)3 -2.085284   0.602981 -3.4583
                                               Pr(z)
(Intercept)                                1.318e-13 ***
Lev                                        0.8419511
Lev_5FU                                    0.0029265 **
differ2                                    0.7970229
differ3                                    0.0050455 **
```

```
nsx(log(time), df = 4)1                     2.509e-10 ***
nsx(log(time), df = 4)2                     < 2.2e-16 ***
nsx(log(time), df = 4)3                     9.193e-09 ***
nsx(log(time), df = 4)4                     < 2.2e-16 ***
as.numeric(differ3):nsx(log(time), df = 3)1 0.0121233 *
as.numeric(differ3):nsx(log(time), df = 3)2 0.0287093 *
as.numeric(differ3):nsx(log(time), df = 3)3 0.0005436 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

-2 log L: 7930.942
Likelihood Ratio Tests
Model 1: fit0, [negll]: (Intercept)+Lev+Lev_5FU+differ2+differ3+
          nsx(log(time), df = 4)1+nsx(log(time), df = 4)2+nsx(log(time), df =
          4)3+nsx(log(time), df = 4)4
Model 2: fit1, [negll]: (Intercept)+Lev+Lev_5FU+differ2+differ3+
          nsx(log(time), df = 4)1+nsx(log(time), df = 4)2+nsx(log(time), df =
          4)3+nsx(log(time), df = 4)4+as.numeric(differ3):nsx(log(time), df =
          3)1+as.numeric(differ3):nsx(log(time), df = 3)2+
          as.numeric(differ3):nsx(log(time), df = 3)3
  Tot Df Deviance  Chisq Df Pr(>Chisq)
1      9   7956.3
2     12   7930.9 25.363  3  1.296e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(e) Based on the plot of the time-varying hazard ratio comparing `differ==3` with `differ==1` (Figure 2), describe how the hazard ratio varies by time since randomisation? Explain your reasoning. (2 pts)

```
plot(fit1, newdata=data.frame(Lev=0,Lev_5FU=0,differ2=0,differ3=0), type="hr",
     var="differ3", ylim=c(0,3), xlab="Time since randomisation (days)")
box() # for a nice box around the plot
```

(f) An alternative approach to assess for non-proportionality would be to fit a Cox regression model and then use Schoenfeld residuals to (i) test for non-proportionality and (ii) plot for a smoothed log hazard ratio. Compare and contrast using Schoenfeld residuals with Cox regression with using flexible parametric survival models to investigate time-varying hazards ratios. (2 pts)

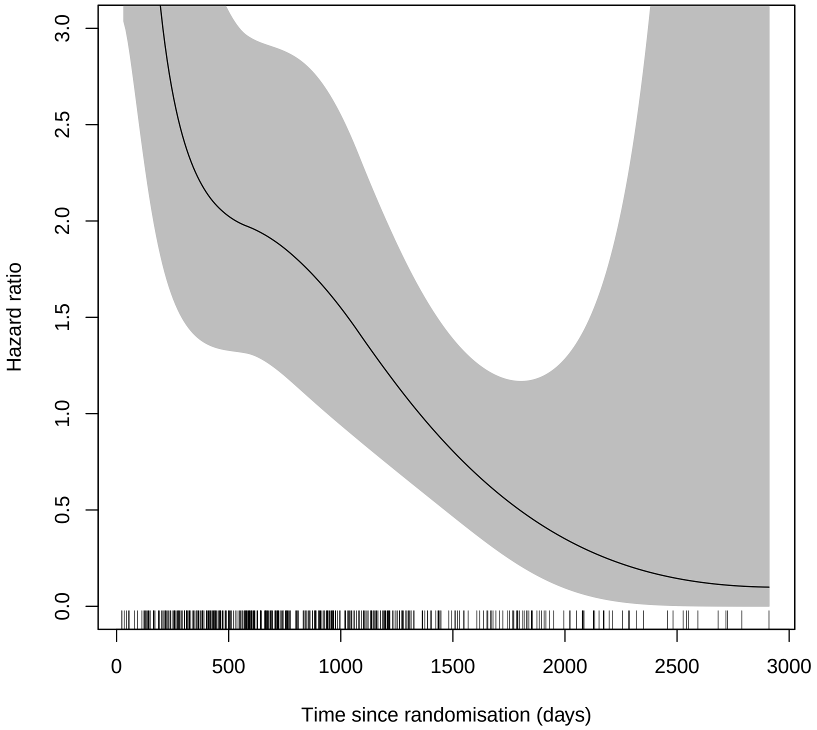

Figure 2: Time-varying hazard ratio for tumour differentiation 3 compared with tumour differentiation 1 for overall survival, colon cancer patients

Q3

We put the following question to ChatGPT: "In survival analysis, we can compare two groups using a variety of estimands. Discuss the different estimands and when we would prefer one estimand over another."

On 2023-11-02, ChatGPT responded:

> *In survival analysis, researchers often compare two or more groups to assess the impact of different factors on the time to an event (e.g., time to death, time to disease recurrence). The choice of estimand (the parameter of interest) is critical and should align with the research question and study design. Here are some common estimands used in survival analysis and when to prefer one over another:*
>
> 1. *Hazard Ratio (HR):*
>   - *Estimand: The HR quantifies the ratio of the hazard rates (instantaneous event rates) between two groups. It measures how the hazard of an event in one group compares to the hazard in another.*
>   - *When to Use: HR is often preferred when the research question focuses on comparing the relative instantaneous risk of an event between groups. It is commonly used in Cox proportional hazards models.*
> 2. *Risk Ratio (RR):*
>   - *Estimand: The RR measures the ratio of cumulative incidence or probability of an event at a specific time in one group relative to another group.*

  - *When to Use: RR is suitable when you want to compare the probability of experiencing an event by a certain time point between two groups. This may be preferred when the focus is on cumulative risk.*

3. *Median Survival Time:*
   - *Estimand: The median survival time is the time at which 50% of the individuals in a group have experienced the event.*
   - *When to Use: Median survival time is useful when you want to provide a simple and easily interpretable summary of the time to the event in different groups. It can be used for descriptive purposes but may not capture all relevant information.*
4. *Mean Survival Time:*
   - *Estimand: The mean survival time represents the average time to the event in a group.*
   - *When to Use: Mean survival time can be used when you are interested in the expected time to the event, but it may be sensitive to outliers. It is not commonly used because it does not account for censoring and may be affected by right-censored data.*
5. *Restricted Mean Survival Time (RMST):*
   - *Estimand: RMST represents the average survival time up to a specific time point (e.g., 5 years) in one group compared to another.*
   - *When to Use: RMST is valuable when you want to compare the overall survival experience over a specified time horizon. It provides a clinically interpretable measure and is less sensitive to censoring.*
6. *Area Under the Survival Curve (AUC):*
   - *Estimand: AUC quantifies the area under the survival curves for two or more groups, which represents the overall survival experience over the entire follow-up period.*
   - *When to Use: AUC is useful when you want to compare the overall survival experience without specifying a specific time point. It provides a comprehensive summary of survival differences over the entire follow-up period.*

The choice of estimand depends on the specific research question, the study design, and the objectives of the analysis. HR is often preferred for comparing instantaneous risks, while RR, median survival time, and RMST are suitable for comparing cumulative risks at specific time points. Mean survival time is less common due to its sensitivity to censoring, and AUC provides a comprehensive summary of survival differences over the entire follow-up period. Researchers should select the estimand that best addresses the research question and provides the most relevant insights for the study.

Within the context of the course material, critically appraise the answer given by ChatGPT. In particular, state which of the findings you agree with, which you think are wrong (any why), and whether there are any other estimands that could be of interest (and why). (5 pts)

(Part 1: 12 pts; Part 2: 19 pts)