# BIOSTAT III: Survival Analysis for Epidemiologists: Take-home examination

Therese Andersson

6–15 February, 2023

## 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 February 2023 and **the examination is due by 17:00 on Wednesday 22 February 2023**.
- The examination is in two parts. To pass the examination, you need to score at least 9/17 for Part 1 focused on rates and general regression modelling and 13/25 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.**

# 1 Description of the data

In this exam we use data on breast cancer patients. The exposure variable of interest is chemotherapy and we are interested on its effect on all-cause mortality. Start of follow-up is at date of surgery, and the time-scale of interest is time since surgery. Follow-up is restricted to 10 years after surgery, so everyone still at risk after 10 years is censored at that point. We also have information on age at surgery and the number of positive lymph nodes (i.e. metastases in lymph nodes). Below is a description of the variables used in this exam:

```
. codebook chemo agegrp enodes d risktime

-------------------------------------------------------------------------------
chemo                                                          Chemo therapy
-------------------------------------------------------------------------------

                  Type: Numeric (byte)
                 Label: adjchemo

                 Range: [0,1]                          Units: 1
         Unique values: 2                          Missing .: 0/2,982

            Tabulation: Freq.   Numeric  Label
                        2,402         0  no
                          580         1  yes

-------------------------------------------------------------------------------
agegrp                                        Age group in 4 categories
-------------------------------------------------------------------------------

                  Type: Numeric (float)
                 Label: agelabel

                 Range: [0,70]                         Units: 1
         Unique values: 4                          Missing .: 0/2,982

            Tabulation: Freq.   Numeric  Label
                          712         0  <45
                        1,119        45  45-59
                          690        60  60-70
                          461        70  70+

-------------------------------------------------------------------------------
enodes      Number of positive nodes (transformed as exp(-12 * nodes))
-------------------------------------------------------------------------------

                  Type: Numeric (float)

                 Range: [.01690747,1]                  Units: 1.000e-09
         Unique values: 28                         Missing .: 0/2,982

                  Mean: .795889
             Std. dev.: .263865

           Percentiles:     10%       25%       50%       75%       90%
                        .339596   .618783    .88692         1         1

-------------------------------------------------------------------------------
d                  Indicator for death due to any cause, 1=yes, 0=no
```

```
------------------------------------------------------------------------------

                  Type: Numeric (float)

                 Range: [0,1]                         Units: 1
         Unique values: 2                          Missing .: 0/2,982

            Tabulation: Freq.  Value
                        2,089  0
                          893  1

------------------------------------------------------------------------------
risktime                                Follow-up time in exact years
------------------------------------------------------------------------------

                  Type: Numeric (float)

                 Range: [.09856263,10]                Units: 1.000e-09
         Unique values: 1,663                      Missing .: 0/2,982

                  Mean: 6.70772
             Std. dev.: 2.92504

           Percentiles:     10%       25%       50%       75%       90%
                        2.25051   4.39973   7.22382   9.73306        10




. stset risktime, f(d==1) exit(time 10)

Survival-time data settings

         Failure event: d==1
Observed time interval: (0, risktime]
     Exit on or before: time 10

------------------------------------------------------------------------------
      2,982  total observations
          0  exclusions
------------------------------------------------------------------------------
      2,982  observations remaining, representing
      1,171  failures in single-record/single-failure data
 20,002.424  total analysis time at risk and under observation
                                                At risk from t =         0
                                     Earliest observed entry t =         0
                                          Last observed exit t =        10
```

# Part 1

## Q 1

Below is the output from a Poisson model with all-cause deaths as the outcome and chemotherapy, age group at surgery and number of positive nodes as explanatory variables.

```
. poisson d i.chemo i.agegrp enodes, exp(risktime) irr

Iteration 0:   log likelihood = -2783.5943
Iteration 1:   log likelihood = -2783.4413
Iteration 2:   log likelihood = -2783.4413

Poisson regression                                      Number of obs =  2,982
                                                        LR chi2(5)    = 477.51
                                                        Prob > chi2   = 0.0000
Log likelihood = -2783.4413                             Pseudo R2     = 0.0790

------------------------------------------------------------------------------
           d |        IRR   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       chemo |
         yes |   .8872142    .073595    -1.44   0.149     .7540859    1.043845
             |
      agegrp |
      45-59 |   .9009499   .0736836    -1.28   0.202     .7675129    1.057586
      60-70 |   .9331429   .0891465    -0.72   0.469      .773802    1.125295
        70+ |   1.490927   .1445439     4.12   0.000     1.232916    1.802933
             |
      enodes |   .1250902   .0121013   -21.49   0.000     .1034851     .151206
       _cons |   .2889345     .03085   -11.63   0.000     .2343771    .3561915
------------------------------------------------------------------------------

. est store A
```

a) Interpret the parameter for chemotherapy ('chemo') in the output above, including a statement about statistical significance. (2 p)

b) Interpret the parameter for age group '60-70' in the output above, including a statement about statistical significance. (2 p)

c) Write out the model formulation (linear predictor) for the model above, make sure to explain your notation. (2 p)

d) What is the hazard ratio comparing a patient who received chemotherapy and had surgery aged '60-70' to a patient who had no chemotherapy and had surgery aged '70+'? For this comparison assume that both patients had the same number of positive nodes. (2 p)

e) Based on the output given so far, is it possible to judge if age is a confounder? If yes, is age a confounder (motivate your answer)? If no, why is it not possible to judge if age is a confounder based on the output above? (2 p)

## Q 2

A second Poisson model is fitted below, including interaction terms between chemotherapy and age group. The model is also compared with the model fitted in Q1 using a likelihood-ratio test.

```
. poisson d i.chemo##i.agegrp enodes , exp(risktime) irr

Iteration 0:   log likelihood = -2783.0273
Iteration 1:   log likelihood = -2782.8722
Iteration 2:   log likelihood = -2782.8722

Poisson regression                                      Number of obs =  2,982
                                                        LR chi2(8)    = 478.65
                                                        Prob > chi2   = 0.0000
Log likelihood = -2782.8722                             Pseudo R2     = 0.0792

------------------------------------------------------------------------------
           d |        IRR   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       chemo |
         yes |    .880482   .1125227    -1.00   0.319     .6853932    1.131101
             |
      agegrp |
       45-59 |   .8892624   .0934244    -1.12   0.264     .7237759    1.092586
       60-70 |   .9329084   .0994644    -0.65   0.515     .7569828     1.14972
         70+ |   1.492669   .1601195     3.73   0.000     1.209636    1.841927
             |
chemo#agegrp |
  yes#45-59  |    1.03992   .1740583     0.23   0.815     .7490814    1.443681
  yes#60-70  |   .8491221    .397948    -0.35   0.727     .3388813    2.127613
    yes#70+  |   .4424782   .4468141    -0.81   0.419     .0611434    3.202096
             |
      enodes |   .1247113   .0120906   -21.47   0.000     .1031295    .1508096
       _cons |   .2905858     .03417   -10.51   0.000     .2307713     .365904
------------------------------------------------------------------------------

. est store B

. lrtest A B

Likelihood-ratio test
Assumption: A nested within B

 LR chi2(3) =   1.14
Prob > chi2 = 0.7678
```

a) Interpret the parameter for chemotherapy ('chemo') in the output above, including a statement about statistical significance. (2 p)

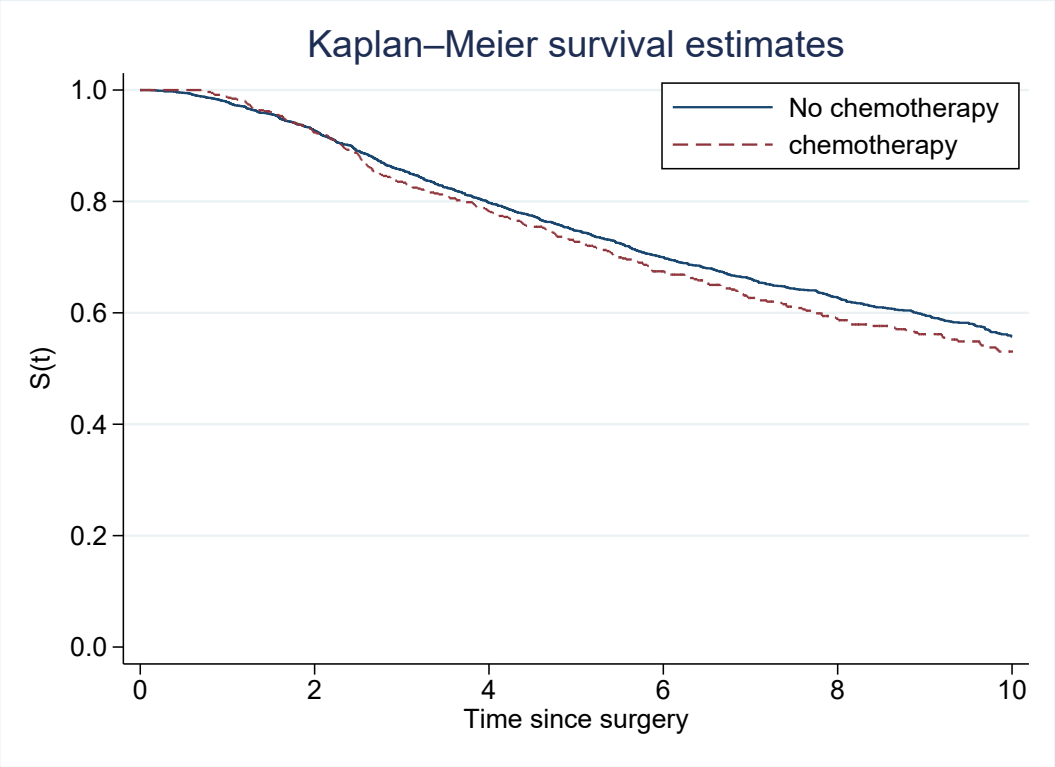
b) What is the hazard ratio comparing a patient who received chemotherapy and had surgery aged '60-70' to a patient who had no chemotherapy and had surgery aged '60-70'? For this comparison assume that both patients had the same number of positive nodes. (2 p)

c) Is there evidence of effect modification by age on the effect of chemotherapy? Motivate your answer. (3 p)

# Part 2

## Q 3

Below is a Kaplan-Meier graph of the survivor function for the 2 treatment groups, and the output from a log rank test.




```
. sts test chemo

         Failure _d: d==1
   Analysis time _t: risktime
  Exit on or before: time 10

Equality of survivor functions
Log-rank test

      |  Observed       Expected
chemo |    events         events
------+--------------------------
   no |       929         946.48
  yes |       242         224.52
------+--------------------------
Total |      1171        1171.00

                chi2(1) =   1.68
                Pr>chi2 = 0.1944
```

a) Based on the Kaplan-Meier graph, what is the 5-year survival for each of the 2 treatment groups (approximately)? (2 p)

b) Based on the Kaplan-Meier graph, which of the 2 treatment groups has a better survival? (2 p)

c) Based on the Kaplan-Meier graph, what can you conclude about the hazard rate of death for each treatment group? (4 p)

d) Would you say that the proportional hazards assumption is reasonable? Motivate your answer. (2 p)

e) Is there evidence of a difference in all-cause mortality between chemotherapy and no chemotherapy? (1 p)

## Q 4

Below is the output from a Cox model, and test of the proportional hazards assumption based on the Schoenfelds residuals from this model.

```
. stcox i.chemo i.agegrp enodes

        Failure _d: d==1
  Analysis time _t: risktime
  Exit on or before: time 10

Iteration 0:   log likelihood = -8957.8518
Iteration 1:   log likelihood = -8759.5628
Iteration 2:   log likelihood = -8701.5721
Iteration 3:   log likelihood = -8701.4741
Iteration 4:   log likelihood = -8701.4741
Refining estimates:
Iteration 0:   log likelihood = -8701.4741

Cox regression with Breslow method for ties

No. of subjects =       2,982                           Number of obs =  2,982
No. of failures =       1,171
Time at risk    = 20,002.4244
                                                        LR chi2(5)    = 512.76
Log likelihood = -8701.4741                             Prob > chi2   = 0.0000

------------------------------------------------------------------------------
          _t | Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       chemo |
         yes |   .8742699   .0725081    -1.62   0.105     .7431058    1.028585
             |
      agegrp |
      45-59  |   .8942929   .0731203    -1.37   0.172     .7618734    1.049728
      60-70  |    .921299   .0880301    -0.86   0.391      .763956    1.111048
        70+  |   1.507576   .1460935     4.24   0.000     1.246788    1.822913
             |
      enodes |   .1120086   .0109474   -22.40   0.000     .0924821    .1356579
------------------------------------------------------------------------------
```

```
. estat phtest, detail

Test of proportional-hazards assumption

Time function: Analysis time
-------------------------------------------------------
             |        rho     chi2       df    Prob>chi2
-------------+-----------------------------------------
    0b.chemo |          .        .        1           .
     1.chemo |    0.03882     1.76        1       0.1851
   0b.agegrp |          .        .        1           .
   45.agegrp |    0.04000     1.87        1       0.1715
   60.agegrp |    0.04358     2.22        1       0.1362
   70.agegrp |    0.09777    11.33        1       0.0008
      enodes |    0.10315    10.91        1       0.0010
-------------+-----------------------------------------
 Global test |               20.12        5       0.0012
-------------------------------------------------------
```

a) Is this model equivalent to the Poisson model in question 1 (Q1)? Motivate your answer. (2 p)

b) Write out the model formulation (linear predictor) of the Cox model. (2 p)

c) What is the hazard ratio comparing chemotherapy to no chemotherapy for patients within the same age category at surgery and the same number of positive nodes (enodes)? (2 p)

d) Is there evidence of non-proportional hazards for any of the covariates in the model? Motivate your answer. (2 p)

e) Would a stratified Cox model be suitable to deal with non-proportional hazards in the model above? Why/why not? (2 p)

## Q 5

a) Describe why it can be better to explore differences in survival outcomes using a regression model instead of a log-rank test. Motivate your answer. (2 p)

b) Describe a study where you would choose attained age as the time-scale in the survival analysis. Motivate your answer. (2 p)