Survival Analysis

Pham Mai Tam
12/7/2022

Outline

- Introduction to survival analysis
- Common terms in survival analysis
- Methods of survival analysis

  Kaplan-Meier method, Log-rank test, Cox regression (Cox proportional hazards, CPH, model)
- Perform survival analysis in R using TCGA dataset

Introduction to survival analysis

- Survival analysis (time-to-event analysis) is a collection of statistical methods analyzing the duration of time (e.g. survival time) until the occurrence of event of interest (EOI) (e.g. death).
- Survival time is a follow-up time measured between the defined starting point and occurrence of an EOI (e.g. time from disease diagnosis to death)
- Some patients' survival time may not be known → censoring
- Standard statistical methods cannot be applied as data is censored, heavily skewed, and not normally distributed. → require survival analysis
- Main goal: to estimate the survival probability from survival time and assess the effect of the effects of factors on survival time.

Why survival analysis, not logistic regression?

Logistic regression	Survival analysis
Binary event outcome	Time-to-event outcome
If clinical outcome is mortality or not → logistic regression can be used	If time to mortality is an observed outcome, survival analysis is used (the time until death happens)

Common terms in survival analysis

- Survival function (survival probability), *S(t)*, refers to P of surviving of a patient from the starting time (e.g. start of diagnosis) to beyond a specific time *t*.
- Hazard function (hazard rate), *h(t)*, refers to instantaneous rate of occurrence of EOI given that the patient is survived until that time. → Higher value of hazard function, higher risk of EOI → a crucial part of Cox proportional hazards (CPH) model.
- Difference b/w *S(t)* and *h(t)*: *S(t)* probability of not having an EOI, *h(t)* probability of EOI occuring.
- Hazard ratio (HR) is described as the ratio of hazard rate or failure rate between two groups. → HR=1 indicates that there are no differences b/w two groups, HR>1 indicates that EOI is most likely to occur and vice versa.

Methods of survival analysis

- Kaplan-Meier method is a non-parametric method (It assumes no specific distribution of survival times and does not assume a relationship b/w survival times and independent variables)
- KM method estimates survival probability from the observed survival times (both censored and uncensored) that survival probability is plotted against time *t* in KM survival curve.
- KM method run only in a categorical variable. If you want to include many variables which is quantitative, you should use Cox regression model.

Kaplan-Meier (KM) survival function

$$S(t_i) = S(t_{i-1})(1 - \frac{d_i}{n_i})$$

- t_i = *a time when at least one event happened;*
- $S(t_i)$ = probability of survival at time t_i;
- $S(t_{i-1})$ = probability of survival at time t_{i-1};
- n_i = # of patients known to have survived (have not yet had an event or been censored) up to time t_i;
- d_i = # of patients having EOI (e.g. death) happened at time t_i

Example data from TCGA package

```
head(BRCA.mg)
#   bcr_patient_barcode     GATA3 new_tumor_days death_days followUp_days
# 1        TCGA-A1-A0SD  2.870500           <NA>       <NA>           437
# 2        TCGA-A1-A0SE  2.166250           <NA>       <NA>          1321
# 3        TCGA-A1-A0SH  1.323500           <NA>       <NA>          1437
# 4        TCGA-A1-A0SJ  1.841625           <NA>       <NA>           416
# 5        TCGA-A1-A0SK -6.025250           <NA>        967          <NA>
# 6        TCGA-A1-A0SM  1.804500           <NA>       <NA>           242
#   time_new_tumor time_death death_event
# 1            437        437           0
# 2           1321       1321           0
# 3           1437       1437           0
# 4            416        416           0
# 5             NA        967           1
# 6            242        242           0
```

Kaplan-Meier method

```r
## Run survival analysis
# First, create a survival object with survival time and outcome. In a survival object, the event parameter must be logical (T/F) where T=death,
or numeric (0/1) where 1=death, 0=alive or censored.
surv <- Surv(time = BRCA.mg$time_death, event = BRCA.mg$death_event)
head(surv)
# [1]  437+ 1321+ 1437+  416+  967   242+
# "+" means censored

fit <- survfit(formula = surv ~ BRCA.mg[, gene]>0, data=BRCA.mg) # gene expression > 0 --> up-regulated, gene expression < 0 --> down-regulated
fit
# Call: survfit(formula = surv ~ BRCA.mg[, gene] > 0, data = BRCA.mg)

#                             n events median 0.95LCL 0.95UCL
# BRCA.mg[, gene] > 0=FALSE 125     15     NA      NA      NA
# BRCA.mg[, gene] > 0=TRUE  465     66   3941    3126      NA

summary(fit)
# Call: survfit(formula = surv ~ BRCA.mg[, gene] > 0, data = BRCA.mg)

#                 BRCA.mg[, gene] > 0=FALSE
#  time n.risk n.event survival std.err lower 95% CI upper 95% CI
#   524     88       1    0.989  0.0113        0.967        1.000
#   548     86       1    0.977  0.0160        0.946        1.000
#   571     83       1    0.965  0.0197        0.928        1.000
#   612     79       1    0.953  0.0229        0.909        0.999
#   639     74       1    0.940  0.0260        0.891        0.993
```

At t_1=524 days → 0.989 = 1 x (1 – 1/88)

At t_2=548 days → 0.977 = 0.989 x (1 – 1/86)

At t_3=571 days → 0.965 = 0.977 x (1 – 1/83)

At t_4=612 days → 0.953 = 0.965 x (1 - 1/79)

Log-rank test

- Log-rank test is a **non-parametric hypothesis test**, which compares estimates of the hazard functions of the two groups at each observed event time (e.g. compare 2 survival curves)
- H_o: there are no differences in the survival curves b/w G1 and G2 ($h_1(t) = h_2(t)$) → 2 groups has identical hazard function
- H_1: there are differences in the survival curves b/w G1 and G2

Log-rank test

$$\chi^2 = \sum_{i}^{n} \frac{(O_i - E_i)^2}{E_i}$$

- X^2 = *log-rank statistic*
- O_i = # of observed events in group i;
- E_i = # of expected events in group i;
- n = # of groups
- Log-rank test value is compared against critical value from X^2 distribution with n-1 degree of freedoms

Log-rank test

```
log_rank_test <- survdiff(formula = surv ~ BRCA.mg[, gene]>0, data=BRCA.mg)
log_rank_test
# Call:
# survdiff(formula = surv ~ BRCA.mg[, gene] > 0, data = BRCA.mg)

#                             N Observed Expected (O-E)^2/E (O-E)^2/V
# BRCA.mg[, gene] > 0=FALSE 125       15     19.2     0.936      1.24
# BRCA.mg[, gene] > 0=TRUE  465       66     61.8     0.292      1.24

#  Chisq= 1.2  on 1 degrees of freedom, p= 0.3
# p = 0.3 --> there are no significant differences between the survival curves (high and low gene GATA3 expression)
```

Cox proportional hazards (CPH) model (Cox regression)

- Unlike KM curves and log-rank test are useful only when predictor variable is categorical (e.g. treatment A vs B, male vs female), Cox model works with quantitative predictors (e.g. gene expression, weight, age ...)
- CPH model uses hazard function instead of survival probability or survival time → hazard function is a measure of effect in CPH model.

Cox model

$$h(t) = h_0(t) \times exp(\sum_{i}^{n} b_i \times X_i)$$

- h(t) = expected hazard at time t
- $h_o(t)$ = baseline hazard → an intercept
- X = independent variables
- When there is no effect of independent variables (X=0) $h(t) = h_0(t)$
- The quantities exp(b_i) are called hazard ratios (HR). A value of b_i greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the i^{th} covariate increases, the event hazard increases and thus the length of survival decreases.

Cox regression model

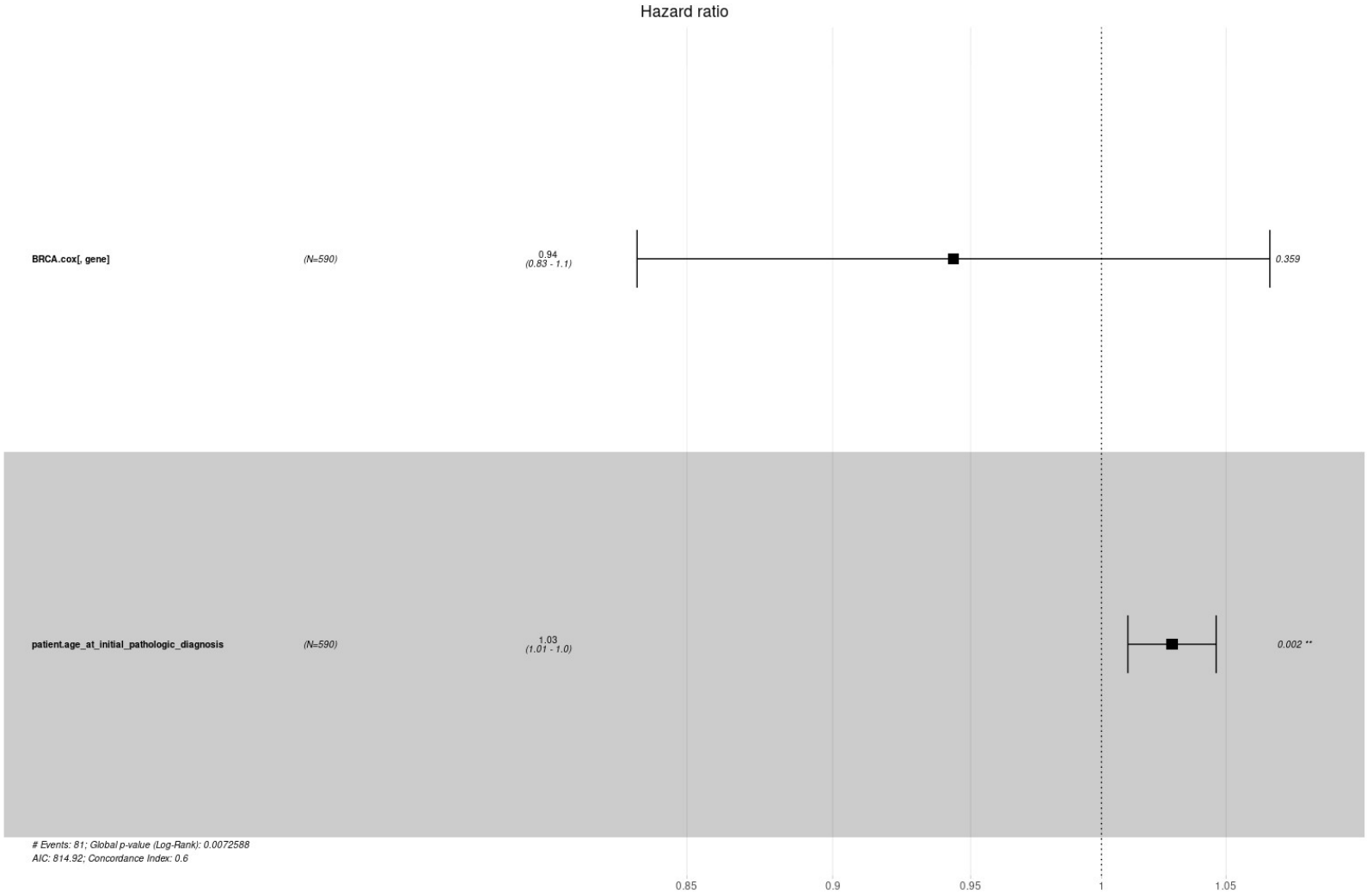
In summary,

- HR = 1: no effect
- HR < 1: reduction in hazard
- HR > 1: Increase in hazard

Notice in cancer studies,

- A covariate with HR > 1 (b > 0) is called bad prognostic factor
- A covariate with HR < 1 (b < 0) is called good prognostic factor

```r
cox_reg_model <- coxph(formula = surv ~  BRCA.cox[, gene] + patient.age_at_initial_pathologic_diagnosis, data = BRCA.cox)
summary(cox_reg_model)
# Call:
# coxph(formula = surv ~ BRCA.cox[, gene] + as.numeric(patient.age_at_initial_pathologic_diagnosis),
#     data = BRCA.cox)

#   n= 590, number of events= 81

#                                                       coef exp(coef)
# BRCA.cox[, gene]                                    -0.058018  0.943633
# as.numeric(patient.age_at_initial_pathologic_diagnosis)  0.027640  1.028026
#                                                     se(coef)      z
# BRCA.cox[, gene]                                    0.063261 -0.917
# as.numeric(patient.age_at_initial_pathologic_diagnosis)  0.008829  3.131
#                                                     Pr(>|z|)
# BRCA.cox[, gene]                                    0.35908
# as.numeric(patient.age_at_initial_pathologic_diagnosis)  0.00174 **
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#                                                     exp(coef) exp(-coef)
# BRCA.cox[, gene]                                       0.9436     1.0597
# as.numeric(patient.age_at_initial_pathologic_diagnosis)    1.0280     0.9727
#                                                     lower .95 upper .95
# BRCA.cox[, gene]                                       0.8336     1.068
# as.numeric(patient.age_at_initial_pathologic_diagnosis)    1.0104     1.046

# Concordance= 0.602  (se = 0.036 )
# Likelihood ratio test= 9.85  on 2 df,   p=0.007
# Wald test            = 9.85  on 2 df,   p=0.007
# Score (logrank) test = 9.93  on 2 df,   p=0.007

ggforest(cox_reg_model, data = BRCA.cox)
```


Hazard ratio
BRCA.cox[, gene]
(N=590)
0.94
(0.83 - 1.1)
0.359
patient.age_at_initial_pathologic_diagnosis
(N=590)
1.03
(1.01 - 1.0)
0.002 **
Events: 81; Global p-value (Log-Rank): 0.0072588
AIC: 814.92; Concordance Index: 0.6
0.85
0.9
0.95
1
1.05