PH 716 Applied Survival Analysis

Part 9: Competing Risks

Zhiyang Zhou (zhou67@uwm.edu, zhiyanggeezhou.github.io)

2026/05/06 16:58:02

What are competing risks?

- In many real-world studies, a subject can experience one of **several different types of events**. In practice, we only observe the first event, and follow-up stops afterward. As a result, once one event is observed, the other event types are no longer observable for that subject.

Examples of competing risks in medical studies

- A person may die from: natural causes, accidental death, homicide, suicide, etc.
- A patient may develop: heart disease,diabetes, kidney diseases, etc.
- If one event is observed, the follow-up will be terminated.

Data structure for competing risks

- i: subject index, $i = 1, \ldots, n$
- $\widetilde{T}_i = \min(T_i, C_i)$: observed time for subject i
  - T_i: authentic time to first event for subject i
  - C_i: censoring time for subject i
- Δ_i: the type of the first observed event for subject i
  - $\Delta_i = k$, $k = 1, \ldots, K$: $T_i = \widetilde{T}_i$ and the first event is of type k
  - $\Delta_i = 0$: $T_i = C_i$, i.e., the subject is censored before any event occurs
- $x_{i1}, \ldots, x_{ip}$: values of covariates for subject i

Recall the hazard function

- Limited to continuous T_i
- Hazard function

$$\lambda_i(t) = \lim_{\delta \to 0^+} \frac{\Pr(t \leq T_i < t + \delta \mid T_i \geq t)}{\delta}$$

  - The instantaneous risk of experiencing the first event at time t, assuming the subject has survived up to t
  - NOT accounting for competing risks
- Why event type matters? If we ignore event types and treat all events as the same, we may miss important patterns.
  - Example: A drug effect might strongly extend the cancer-free survival but not survival from other causes.

Cause-specific hazard

- Instead of one hazard shared by all event types, we now have one per type:

$$\lambda_i^{(k)}(t) = \lim_{\delta \to 0^+} \frac{\Pr(t \le T_i < t+\delta, \Delta_i = k \mid T_i \ge t)}{\delta}, \quad k = 1, \ldots, K$$

  - Interpretation: at time t, the instantaneous risk of event type k, given the subject has not had any event yet
  - Overall instantaneous risk = sum of instantaneous risks for each event type: $\lambda_i(t) = \sum_{k=1}^K \lambda_i^{(k)}(t)$
- Cumulative cause-specific hazard $\Lambda_i^{(k)}(t) = \int_0^t \lambda_i^{(k)}(u)\mathrm{d}u$
  - Overall cumulative risk = sum of cumulative risks for each event type: $\Lambda_i(t) = \sum_{k=1}^K \Lambda_i^{(k)}(t)$

Cumulative incidence function (CIF, aka sub-distribution function)

- CIF: $F_i^{(k)}(t) = \Pr(T_i \le t, \Delta_i = k)$
  - Interpretation: the probability that the first observed event is of type k and occurs by time t
  - CDF = sum of CIFs for each event type: $F_i(t) = \sum_{k=1}^K F_i^{(k)}(t) \Rightarrow S_i(t) = 1 - \sum_{k=1}^K F_i^{(k)}(t)$
  - NOT a cumulative distribution function (CDF) because $F_i^{(k)}(\infty) < F_i(\infty) = 1$
- Theoretical notes
  - $\lambda_i^{(k)}(t) = \frac{\mathrm{d}F_i^{(k)}(t)/\mathrm{d}t}{S_i(t)}$
    - Proof: $\lambda_i^{(k)}(t) = \lim_{\delta\to 0^+} \frac{\Pr(t\le T_i<t+\delta, \Delta_i=k, T_i\ge t)}{\delta \Pr(T_i \ge t)} = \lim_{\delta\to 0^+} \frac{\Pr(t\le T_i<t+\delta, \Delta_i=k)}{\delta S_i(t)} = \frac{\mathrm{d}F_i^{(k)}(t)/\mathrm{d}t}{S_i(t)}$
  - $F_i^{(k)}(t) = \int_0^t \lambda_i^{(k)}(u) S_i(u)\mathrm{d}u$ (backing up KM estimators of CIFs and overall survival)

When covariates are not considered: naive KM estimator

- Assuming that
  - T_i iid across i, i.e., $T_i \overset{\text{iid}}{\sim} T$
  - T_i independent of C_i
  - Times to different events are independent (typically violated in medical cases)
    - Implying that at each time point the hazard of each event is the same for subjects at risk as for subjects that have experienced other competing events by that time
- Implementation
  - Given k, take the event type k as the event of interest with other types considered as censored
  - Apply KM estimator to the resulting binary setting
- Theoretical notes
  - Naive KM estimator gives cause-specific survival by $\prod_{j:t_j\le t}\{1-\hat\lambda^{(k)}(t_j)\}$
    - $0 = t_0 < t_1 < \cdots < t_J$: unique failure times
    - $\hat\lambda^{(k)}(t_j) = d_{kj}/r_j$: an estimate of the cause-specific hazard function
      - d_{kj}: # of event k that happened exactly at time t_j
      - r_j: # of individuals at risk up to time t_j
- Underestimating the survival probability (i.e., overestimating the failure probability)
  - Taking other types as censoring mistakenly assumes people who had other events could still experience the event of interest later.
  - The bias inflated when the hazards of competing events are larger.

Ex. 9.1 High risk population in `asaur::prostateSurvival`

- Dataset `asaur::prostateSurvival` involves covariates as below.
  - `grade`: a factor with levels `mode` (moderately differentiated) and `poor` (poorly differentiated)
  - `stage`: a factor with levels `T1ab` (Stage T1, clinically diagnosed), `T1c` (Stage T1, diagnosed via a PSA test), and `T2` (Stage T2)
  - `ageGroup`: a factor with levels `66-69`, `70-74`, `75-79`, & `80+`

  - `survTime`: the survival time from diagnosis to death (from prostate cancer or other causes) or last date known alive
  - `status`: a censoring variable, `0` (censored), `1` (death from prostate cancer), and `2` (death from other causes)
- Use naive KM estimators to give the survival curve of high risk population (i.e., `grade="poor"`, `stage="T2"` & `ageGroup="80+"`) for each event type.

```
options(digits=4)
library(asaur)
library(survival)
sapply(asaur::prostateSurvival, class)
data.ex91 = asaur::prostateSurvival[
  asaur::prostateSurvival$grade == "poor" &
  asaur::prostateSurvival$stage == "T2" &
  asaur::prostateSurvival$ageGroup == "80+"
  ,
]
km.prost.naive = survfit(
  Surv(survTime, event=(data.ex91$status==1)) ~ 1,
  data=data.ex91
)
km.other.naive = survfit(
  Surv(survTime, event=(data.ex91$status==2)) ~ 1,
  data=data.ex91
)
plot(
  km.prost.naive$surv ~ km.prost.naive$time, type="s", ylim=c(0,1), lwd=2, col="blue",
  xlab="Months from prostate cancer diagnosis",
  ylab='Estimated survival probability',
)
lines(km.other.naive$surv ~ km.other.naive$time, type="s", col="green", lwd=2)
legend(
  "topright",
  c(
    "Prostate",
    "Other"
  ),
  col=c('blue','green'), lwd=2
)
```

When covariates are not considered: KM estimator of CIF

- Assuming that
  - T_i iid across i, i.e., $T_i \overset{\text{iid}}{\sim} T$
  - T_i independent of C_i
- Implementation
  - Instead of naive KM: estimate overall survival and then combine it with cause-specific hazards
  - Estimate overall survival $S(t)$ by $\widehat{S}_{\text{KM}}(t) = \prod_{j:t_j \le t}\{1 - \sum_{k=1}^{K} \hat{\lambda}^{(k)}(t_j)\}$
    - $0 = t_0 < t_1 < \cdots < t_J$: unique failure times
    - $\hat{\lambda}^{(k)}(t_j) = d_{kj}/r_j$: an estimate of the cause-specific hazard function
      - d_{kj}: # of event k that happened exactly at time t_j
      - r_j: # of individuals at risk up to time t_j
  - Estimate CIF $F^{(k)}(t)$ by $\widehat{F}^{(k)}_{\text{KM}}(t) = \sum_{j:t_j \le t} \hat{\lambda}^{(k)}(t_j)\widehat{S}_{\text{KM}}(t_j - 1)$

Revisit Ex. 9.1

Use the KM estimator of CIF of high risk population

```
options(digits=4)
library(asaur)
library(survival)
library(mstate)
sapply(asaur::prostateSurvival, class)
data.ex91 = asaur::prostateSurvival[
  asaur::prostateSurvival$grade == "poor" &
  asaur::prostateSurvival$stage == "T2" &
  asaur::prostateSurvival$ageGroup == "80+"
  ,
]
km.cif = Cuminc(
  time = data.ex91$survTime,
  status = data.ex91$status
)
km.cif

# Plot of CIFs and the overall survival function
plot(
  km.cif$CI.1 ~ km.cif$time, type="s", ylim=c(0,1), lwd=2, col="blue",
  xlab="Months from prostate cancer diagnosis",
  ylab="Probability"
)
lines(km.cif$CI.2 ~ km.cif$time, type="s", lwd=2, col="green")
lines(km.cif$Surv ~ km.cif$time, type="s", lwd=2, col="red")
legend(
  "topright",
  c(
    "CIF (prostate)",
    "CIF (other)",
    'Overall survival'
  ),
  col=c('blue','green','red'), lwd=2
)

# Stacked plot
library(ggplot2)
cuminc_data = as.data.frame(km.cif[, c('time','Surv','CI.1','CI.2')])
cuminc_data = tidyr::pivot_longer(
  cuminc_data, cols = -time, names_to = "Types", values_to = "estimate")
ggplot(data = cuminc_data, aes(x = as.numeric(time), y = estimate, fill = Types)) +
  geom_area(alpha = 0.6) +
  labs(x = "Months from prostate cancer diagnosis", y = "Probability") +
  theme_minimal()
```

Modeling effects of covariates: cause-specific Cox PH model

- Assuming that
  - T_i independent across i given covariates
  - The independent and non-informative censoring
  - Cause-specific proportional hazards

        * $\lambda_i^{(k)}(t) = \lambda_0^{(k)}(t)\exp(\sum_{j=1}^p x_{ij}\beta_j^{(k)})$
            · $\lambda_0^{(k)}(t)$: baseline cause-specific hazard of event k
            · $\beta_1^{(k)}, \ldots, \beta_p^{(k)}$: covariate effects varying from one event to another
        * OR $\lambda_i^{(k)}(t) = \lambda_0^{(k)}(t)\exp(\sum_{j=1}^p x_{ij}\beta_j)$, i.e., β_j shared by all the K events
- Implementation
    - For $\lambda_i^{(k)}(t) = \lambda_0^{(k)}(t)\exp(\sum_{j=1}^p x_{ij}\beta_j^{(k)})$
        * Specify one event of interest and fit a Cox PH model with the remaining $K-1$ events treated as censoring
        * Repeat the above step and obtain K Cox PH models
    - For $\lambda_i^{(k)}(t) = \lambda_0^{(k)}(t)\exp(\sum_{j=1}^p x_{ij}\beta_j)$
        * First reshape the data frame in the "long format"
        * Then fit a Cox PH model stratified by the encoded event label in the long format
- When $\hat{\lambda}_i^{(k)}(t)$ is ready
    - $\widehat{S}_i(t) = \exp\{-\sum_{k=1}^K \int_0^t \hat{\lambda}_i^{(k)}(u)\mathrm{d}u\}$
    - $\widehat{F}_i^{(k)}(t) = \int_0^t \hat{\lambda}_i^{(k)}(u)\widehat{S}_i(u)\mathrm{d}u$
- Pros and cons
    - Easy to implement
    - Straightforward interpretation of regression coefficients in terms of hazard ratio BUT inconvenient to interpret coefficients in terms of contributions to CIFs
    - Bias induced since we treat other events as censoring

Ex. 9.2 Patients at "T2"-stage in `asaur::prostateSurvival`

- Consider patients with `stage="T2"`.

```
options(digits=4)
library(asaur)
library(survival)
sapply(asaur::prostateSurvival, class)
data.ex92 = asaur::prostateSurvival[
  asaur::prostateSurvival$stage == "T2"
  ,
]

# Cause-specific Cox PH model w/o shared coefficients
data.ex92$status.1 = (data.ex92$status==1)
data.ex92$status.2 = (data.ex92$status==2)
cph.prost = coxph(
  Surv(survTime, status.1)~grade + ageGroup,
  data = data.ex92
)
summary(cph.prost)
cph.other = coxph(
  Surv(survTime, status.2)~grade + ageGroup,
  data = data.ex92
)
summary(cph.other)

# Cause-specific Cox PH model w shared coefficients
## Reshape the data into the long format
data.ex92.long = NULL
K = length(unique(data.ex92$status))-1
for (i in 1:nrow(data.ex92)){
```

```
  data.curr = data.ex92[rep(i, times=K),]
  data.curr$event = c('prostate', 'other')
  data.curr$status.long=rep(0,K-1)
  if(data.ex92$status[i]>=1) {
    data.curr$status.long[which(data.curr$event==c('prostate', 'other')[data.ex92$status[i]])]=1
  }
  data.ex92.long = rbind(data.ex92.long, data.curr)
}
data.ex92.long = subset(data.ex92.long, select=-c(status)) # remove columns to avoid confusion

## Cox PH model stratified by event
cph.strat = coxph(
  Surv(survTime, status.long)~grade + ageGroup+strata(event),
  data = data.ex92.long
)
summary(cph.strat)
```

Sub-distribution hazard function

$$\bar{\lambda}_i^{(k)}(t) = -\frac{\mathrm{d}\ln\{1-F_i^{(k)}(t)\}}{\mathrm{d}t} = \frac{\mathrm{d}F_i^{(k)}(t)/\mathrm{d}t}{1-F_i^{(k)}(t)}, \quad k = 1, \ldots, K$$

- Interpretation: the instantaneous rate at which cumulative incidence for cause k accumulates over time.
- NOT the cause-specific hazard function: $\bar{\lambda}_i^{(k)}(t) \leq \lambda_i^{(k)}(t)$
- Theoretical note: $F_i^{(k)}(t) = 1 - \exp\{-\int_0^t \bar{\lambda}_i^{(k)}(u)\mathrm{d}u\}$ (backing up the Fine-Gray model)

Modeling effects of covariates: Fine-Gray model

- Assuming that
  - T_i independent across i given covariates
  - The independent and non-informative censoring
  - $\bar{\lambda}_i^{(k)}(t) = \bar{\lambda}_0^{(k)}(t)\exp(\sum_{j=1}^p x_{ij}\beta_j^{(k)})$
    - $\bar{\lambda}_0^{(k)}(t)$: baseline sub-distribution hazard of event k
    - $\beta_1^{(k)}, \ldots, \beta_p^{(k)}$: covariate effects potentially varying from one event to another
- When $\hat{\bar{\lambda}}_i^{(k)}(t)$ is ready

$$\widehat{F}_i^{(k)}(t) = 1-\exp\left\{-\int_0^t \hat{\bar{\lambda}}_i^{(k)}(u)\mathrm{d}u\right\} = 1-\exp\left\{-\int_0^t \hat{\bar{\lambda}}_0^{(k)}(u)\mathrm{d}u\exp\left(\sum_{j=1}^p x_{ij}\beta_j^{(k)}\right)\right\} = 1-\{1-\widehat{F}_0^{(k)}(t)\}^{\exp(\sum_{j=1}^p x_{ij}\beta_j^{(k)})}$$

  - $\widehat{F}_0^{(k)}(t) = 1 - \exp\{-\int_0^t \hat{\bar{\lambda}}_0^{(k)}(u)\mathrm{d}u\}$: baseline CIF of event k
- Interpretation:
  - If $\beta_j^{(k)} > 0$, then a larger value of covariate j tend to be associated with larger CIF over time for event type k;
  - If $\beta_j^{(k)} < 0$, then a larger value of covariate j tend to be associated with smaller CIF over time for event type k.
- Pros and cons
  - Direct modeling of CIFs
  - Difficult to interpret regression coefficients in terms of hazard ratio, BUT may see effects of covariates to CIFs directly
  - Might not satisfy that $\sum_{k=1}^K \widehat{F}_i^{(k)}(t) \leq 1$ because $\widehat{F}_i^{(k)}$ are fitted independently.

Revisit Ex. 9.2

- Poorly differentiated patients (`grade=poor`) have higher risk for death from both `prostate` and `other`.
- Elder patients also have higher risk for the death from both conditions.

```
options(digits=4)
library(asaur)
data.ex92 = asaur::prostateSurvival[
  asaur::prostateSurvival$stage == "T2"
  ,
]
# Model fitting
cov1 = model.matrix(~ grade + ageGroup, data = data.ex92)[,-1]
cph.subdisthz.prost = cmprsk::crr(
  ftime = data.ex92$survTime,
  fstatus = data.ex92$status,
  cov1 = cov1,
  failcode=1
)
summary(cph.subdisthz.prost)
cph.subdisthz.other = cmprsk::crr(
  ftime = data.ex92$survTime,
  fstatus = data.ex92$status,
  cov1 = cov1,
  failcode=2
)
summary(cph.subdisthz.other)

# Predicted CIFs
cov1_new = matrix( # Specifying subjects with grade=poor and age= 70-74
  c(1, 1, 0, 0),
  byrow = T,
  ncol = 4
)
colnames(cov1_new) = colnames(cov1); cov1_new
predict.prost = predict(
  cph.subdisthz.prost,
  cov1 = cov1_new
)
head(predict.prost)

predict.other = predict(
  cph.subdisthz.other,
  cov1 = cov1_new
)
head(predict.other)

# Create overall survival
pool.time = sort(unique(c(cph.subdisthz.prost$uftime, cph.subdisthz.other$uftime)))
cifs = data.frame(
  time = pool.time,
  surv = numeric(length(pool.time)),
  cif.prost = numeric(length(pool.time)),
  cif.other = numeric(length(pool.time))
```

```
)
for (i in 1:nrow(cifs)){
  if (!(i %in% predict.prost[,1])){
    if (cifs$time[i] < predict.prost[1,1])
      cifs$cif.prost[i] = 0
    else cifs$cif.prost[i] = cifs$cif.prost[i-1]
  }else cifs$cif.prost[i] = predict.prost[,2][predict.prost[,1] == cifs$time[i]]

  if (!(i %in% predict.other[,1])){
    if (cifs$time[i] < predict.other[1,1])
      cifs$cif.other[i] = 0
    else cifs$cif.other[i] = cifs$cif.other[i-1]
  }else cifs$cif.other[i] = predict.other[,2][predict.other[,1] == cifs$time[i]]
}
cifs$surv = 1-cifs$cif.prost-cifs$cif.other
head(cifs)

# Trajectory plot of CIFs and overall survival
plot(
  cifs$cif.prost ~ cifs$time,
  type="s", ylim=c(0,1), lwd=2, col="blue",
  xlab="Months from prostate cancer diagnosis",
  ylab='Probability',
)
lines(cifs$cif.other ~ cifs$time, type="s", col="green", lwd=2)
lines(cifs$surv ~ cifs$time, type="s", col="red", lwd=2)
legend(
  "topright",
  c(
    "Prostate",
    "Other",
    'Overall Survival'
  ),
  col=c('blue','green','red'), lwd=2
)

# Stacked plot of CIFs
cifs_long = tidyr::pivot_longer(
  cifs, cols = -time, names_to = "Types", values_to = "estimate")
library(ggplot2)
ggplot(data = cifs_long, aes(x = as.numeric(time), y = estimate, fill = Types)) +
  geom_area(alpha = 0.6) +
  labs(x = "Months from prostate cancer diagnosis", y = "Probability") +
  theme_minimal()
```