COURSE STK4080/9080

Mandatory assignment Autumn 2019

Solution

Problem 1. The data set contains a random subsample of 747 patients from the SIR 3 (Spread of nosocomial Infections and Resistant pathogens) cohort study at the Charité university hospital in Berlin, with prospective assessment of data to examine the effect of hospital-acquired infections in intensive care.

The variables are as follows:

pneu is 1 for patients with pneumonia present on admission, and 0 for no pneumonia

status is the patient's status at the end of the observation period; 1 for discharge (alive); 2 for death; and 0 for patients still in the unit when the data base was closed (i.e., right censoring)

time is a patient's length of hospital stay (days)

age is a patient's age at admission

sex is 0 for a male patient and 1 for a female patient

There were 97 patients with pneumonia on admission. Overall, 657 patients were discharged alive, 76 patients died, and 14 patients were still in the unit at the end of the study. 21 of the patients who died had pneumonia on admission.

We read the data into R by the command

```
pneudat=read.table("https://folk.ntnu.no/bo/STK4080/obligdat.txt",header=T)
```

We then make separate data sets for `pneu`=0 and 1 by

```
pneudat.pneu = pneudat[which(pneudat$pneu==1),]
pneudat.nopneu = pneudat[which(pneudat$pneu==0),]
```

a) *Describe the present case as a competing risks problem with the two competing causes discharge and death. (For simplicity you may label the causes by 1=discharge; 2=death.)*

Define the cause-specific hazard functions for the two causes 1 and 2, where separate sets of functions should be defined for patients with and without pneumonia at admission.

What are the intuitive interpretations of these functions in the current case?

Solution: One measures the time from a patient is admitted to the hospital to the first occurrence of either discharge or death. Thus we measure a pair (T, H) where T is the time to the event and H equal to 1 or 2 is the type of event.

Cause-specific hazards for cause h are defined by

$$\begin{aligned} \alpha_{h,pneu}(t)dt &= P(t \leq T < t + dt | T > t, \mathsf{pneu} = 1) \\ \alpha_{h,nopneu}(t)dt &= P(t \leq T < t + dt | T > t, \mathsf{pneu} = 0) \end{aligned}$$

The cause-specific hazards give the probability of event of discharge (respectively death) in a small time interval from time t, conditional on still being in hospital at time t,

b) *Give a brief argument that the cumulative cause-specific hazard functions can be estimated nonparametrically by the Nelson-Aalen estimator. [Hint: Verify that multiplicative models are obtained].*

Then estimate and plot the cumulative cause-specific hazard functions for each of the two causes 1 (discharge) and 2 (death), separately for pneu*=0 and* pneu*=1.*

You may, for easy comparison, display the plots as a 2×2 matrix, using the same scales on the axes.

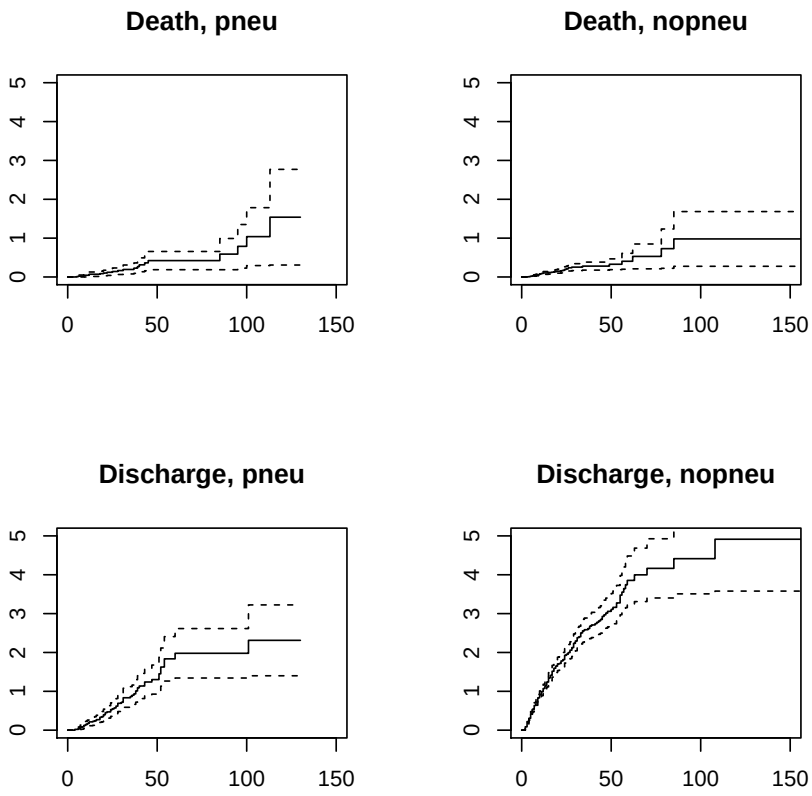
Solution: See Slides 7, p. 11. Consider $\alpha_{h,pneu}(t)$. Let $N_h(t)$ count the number of observed events of type h in $[0, t]$, and let $Y(t)$ be the number at risk (i.e. alive in hospital) just prior to time t. The intensity process of $N_h(t)$ takes the multiplicative form

$$\lambda_h(t) = \alpha_{h,pneu}(t)Y(t)$$

which shows that we have a multiplicative model. Hence the Nelson-Aalen estimator can be used to estimate the cumulative cause-specific hazards.

Similar comment can be given for $\alpha_{h,nopneu}(t)$

```
cause.death.pneu =survfit(Surv(time,status==2)~1,type="fh2",
data=pneudat.pneu)
cause.death.nopneu =survfit(Surv(time,status==2)~1,type="fh2",
data=pneudat.nopneu)
cause.discharge.pneu =survfit(Surv(time,status==1)~1,type="fh2",
data=pneudat.pneu)
cause.discharge.nopneu =survfit(Surv(time,status==1)~1,type="fh2",
data=pneudat.nopneu)
par(mfrow=c(2,2))
plot(cause.death.pneu,fun="cumhaz",xlim=c(0,150),ylim=c(0,5),
main="Death, pneu")
plot(cause.death.nopneu,fun="cumhaz",xlim=c(0,150),ylim=c(0,5),
main="Death, nopneu")
plot(cause.discharge.pneu,fun="cumhaz",xlim=c(0,150),ylim=c(0,5),
main="Discharge, pneu")
plot(cause.discharge.nopneu,fun="cumhaz",xlim=c(0,150),ylim=c(0,5),
main="Discharge, nopneu")
```



c) *The plots obtained in subproblem (b) can seemingly be interpreted as if pneumonia at admission has no effect on the death hazard. It turns out, on the other hand, that more patients with pneumonia die than patients withouot pneumonia. How can this fact be explained from the combined information of the four plots?*

Solution: (Partly copied from the book "Competing Risks and Multistate Models with R" by Jan Beyersmann, Arthur Allignol and Martin Schumacher, Springer Science+Business Media, 2012 .)

Consider the four plots of the Nelson-Aalen estimates. Apparently, pneumonia appears to have no effect on the death hazard. However, this does not imply that pneumonia has no effect on mortality. The reason is that pneumonia appears to reduce the discharge hazard. This implies:

1. Pneumonia appears to reduce the all-cause hazard for end of intensive care unit stay.
2. Patients with pneumonia on admission stay longer on the unit. During this prolonged stay, they are exposed to an essentially unchanged death hazard.
3. As a consequence, more patients with pneumonia die than patients without pneumonia.

This is a typical competing risks phenomenon. Because there is more than one hazard acting on an individual, we cannot tell from one hazard alone what an individual's future course will be. The 'force of death' is not influenced by pneumonia status, but the 'force of discharge' is substantially reduced by pneumonia on admission.

d) *As an alternative to the above analysis involving four cause-specific hazards, one might define the cause-specific hazards for cause 1 and 2 by using (two) Cox-models with* `pneu` *acting as a covariate.*

Write down the appropriate cause-specific hazards for the two causes.

Solution: For $h = 1, 2$, let

$$P(t \leq T < t + dt, H = h | T > t, \mathsf{pneu}) = \alpha_h(t|\mathsf{pneu}) = \alpha_{0h}(t)e^{\beta \cdot \mathsf{pneu}}$$

where α_{0h} is the baseline hazard for cause h.

e) *Do the analyses in R using the cause-specific hazards from subproblem (d).*

What can you conclude from the results? Compare with the conclusions that were drawn from the Nelson-Aalen plots.

Solution: First, consider the cause 1=discharge.

```
> fit.discharge = coxph(Surv(time, status == 1) ~ pneu, data = pneudat)
> summary(fit.discharge)
Call:
coxph(formula = Surv(time, status == 1) ~ pneu, data = pneudat)

  n= 747, number of events= 657

        coef exp(coef) se(coef)      z Pr(>|z|)
pneu -1.0901    0.3362   0.1299 -8.391   <2e-16 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

     exp(coef) exp(-coef) lower .95 upper .95
pneu    0.3362      2.974    0.2606    0.4337

Concordance= 0.577  (se = 0.008 )
Likelihood ratio test= 91.7  on 1 df,   p=<2e-16
Wald test            = 70.4  on 1 df,   p=<2e-16
Score (logrank) test = 77.09  on 1 df,   p=<2e-16
```

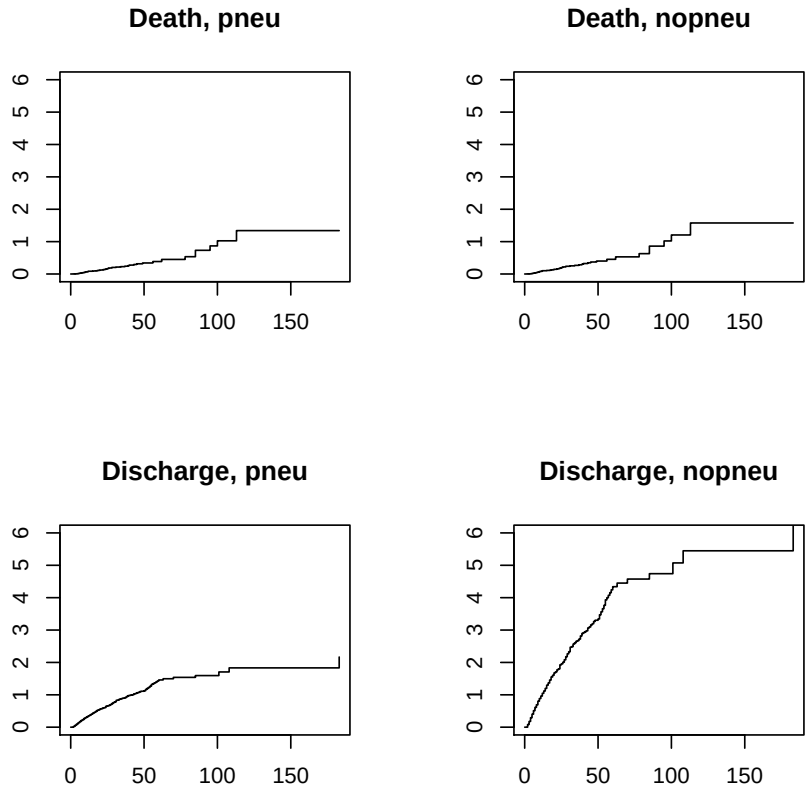
Then, consider cause 2=death:

```
> fit.death = coxph(Surv(time, status == 2) ~ pneu, data = pneudat)
> summary(fit.death)
Call:
coxph(formula = Surv(time, status == 2) ~ pneu, data = pneudat)

  n= 747, number of events= 76

        coef exp(coef) se(coef)      z Pr(>|z|)
pneu -0.1622    0.8503   0.2678 -0.606    0.545

     exp(coef) exp(-coef) lower .95 upper .95
pneu    0.8503      1.176     0.503     1.437

Concordance= 0.531  (se = 0.023 )
Likelihood ratio test= 0.37  on 1 df,   p=0.5
Wald test            = 0.37  on 1 df,   p=0.5
Score (logrank) test = 0.37  on 1 df,   p=0.5
```

The conclusions are similar to the ones obtained by using the four Nelson-Aalen plots. The variable `pneu` is not significant for the risk of death, but is highly significant for the discharge risk.

f) *Plot the estimated cumulative cause-specific hazards for* `pneu` *= 0 and 1 from the above analysis, for each of the two causes 1 and 2. Argue that these four plots are supposed to estimate the same functions as the four Nelson-Aalen plots obtained in subproblem (b). Compare the two sets of plots and comment.*

```
par(mfrow=c(2,2))
plot(survfit(fit.death,data.frame(pneu=1)),fun="cumhaz",ylim=c(0,6), main="Dea
plot(survfit(fit.death,data.frame(pneu=0)),fun="cumhaz",ylim=c(0,6),main="Deat
plot(survfit(fit.discharge,data.frame(pneu=1)),fun="cumhaz",ylim=c(0,6),main="
plot(survfit(fit.discharge,data.frame(pneu=0)),fun="cumhaz",ylim=c(0,6),main="
```

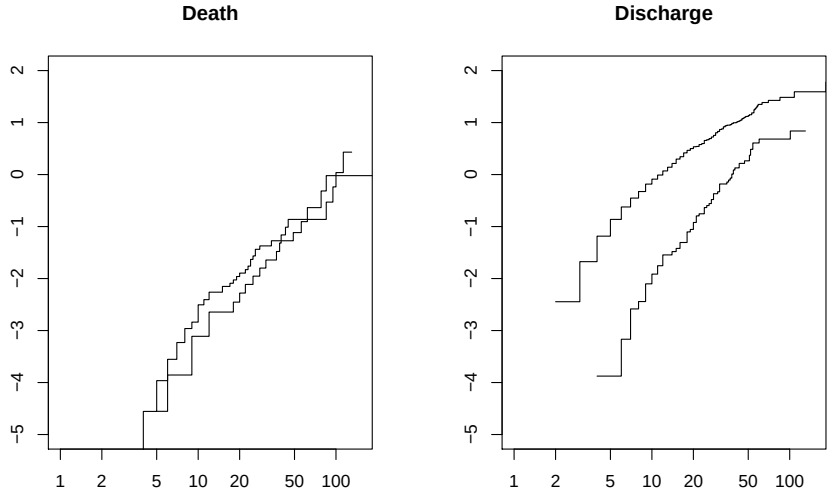


Solution: The four curves are for exactly the same combinations as the ones obtained earlier. They are, however, now based on a proportional hazards model, so the two plots for discharge are proportional, as are the ones for death.

In order to check the proportionality of hazards in the Cox models, you may also plot the log of cumulative hazard, which is the same as $\log(-\log)$ *of the*

survival function, and can be plotted in R by replacing `fun = "cumhaz"` *by* `fun = "cloglog"` *in the plots of subproblem (b). What do you see?*

```
par(mfrow=c(1,2))
plot(cause.death.pneu,fun="cloglog",ylim=c(-5,2),xlim=c(1,150),
conf.int=F, main="Death")
lines(cause.death.nopneu,fun="cloglog",conf.int=F,main="Death")
plot(cause.discharge.pneu,fun="cloglog",ylim=c(-5,2),xlim=c(1,150),
conf.int=F,main="Discharge")
lines(cause.discharge.nopneu,fun="cloglog",conf.int=F,main="Discharge")
```



Solution:

The curves in these plots are supposed to be parallell if the proportional hazards hold. This might seem to be OK for Death, but not necessarily for Discharge.

g) *Extend the Cox regression models fitted in subproblem (e) to include also the covariates* `age` *and* `sex` *(in addition to* `pneu`*). Write down the resulting hazard functions. Do the covariates* `age` *and* `sex` *contribute in a significant manner? Do their contributions differ for the two causes?*

Solution: Consider the cause Death first:

```
> fit.death.full = coxph(Surv(time, status == 2) ~ pneu+sex+age,data=pneudat)
> summary(fit.death.full)
Call:
coxph(formula = Surv(time, status == 2) ~ pneu + sex + age, data = pneudat)
```

```
  n= 747, number of events= 76

          coef exp(coef)  se(coef)      z Pr(>|z|)
pneu -0.084246  0.919205  0.270987 -0.311   0.7559
sex   0.409390  1.505899  0.235895  1.735   0.0827 .
age   0.014215  1.014317  0.007417  1.917   0.0553 .
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

     exp(coef) exp(-coef) lower .95 upper .95
pneu    0.9192     1.0879    0.5404     1.563
sex     1.5059     0.6641    0.9484     2.391
age     1.0143     0.9859    0.9997     1.029

Concordance= 0.596  (se = 0.041 )
Likelihood ratio test= 7.24  on 3 df,   p=0.06
Wald test            = 7.16  on 3 df,   p=0.07
Score (logrank) test = 7.21  on 3 df,   p=0.07
```

The covariate `pneu` is still not significant, while `age` is close to being significant at 5% level, while `sex` shows a tendency for higher risk for females, but this is not significant.

Now consider the cause Discharge:

```
> fit.discharge.full = coxph(Surv(time, status == 1) ~ pneu+sex+age, data = pn
> summary(fit.discharge.full)
Call:
coxph(formula = Surv(time, status == 1) ~ pneu + sex + age, data = pneudat)

  n= 747, number of events= 657

          coef exp(coef)  se(coef)      z Pr(>|z|)
pneu -1.105551  0.331028  0.130402 -8.478   <2e-16 ***
sex   0.122453  1.130266  0.079999  1.531    0.126
age  -0.003239  0.996766  0.002170 -1.493    0.135
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

     exp(coef) exp(-coef) lower .95 upper .95
pneu    0.3310     3.0209    0.2564    0.4274
sex     1.1303     0.8847    0.9662    1.3221
age     0.9968     1.0032    0.9925    1.0010
```

```
Concordance= 0.585  (se = 0.013 )
Likelihood ratio test= 96.04  on 3 df,   p=<2e-16
Wald test            = 74.64  on 3 df,   p=4e-16
Score (logrank) test = 81.24  on 3 df,   p=<2e-16
```

While pneu is here very significant, as we have seen before, the covariates sex and age do not have a significant influence.

h) *Define the cumulative incidence functions (sometimes called the subdistribution functions) for the two causes 1 and 2, one pair of functions for each value of* pneu*. Describe in words what these functions mean in the present case.*

Describe briefly how the functions can be estimated from data. (The estimator presented in the lectures and in the book is a special case of the so-called Aalen-Johansen estimator).

Solution: The cumulative incidence functions are defined by

$$P(T \leq t, H = h | \text{pneu} = 1); h = 1, 2$$
$$P(T \leq t, H = h | \text{pneu} = 0); h = 1, 2$$

The first set give the probability of being discharged alive ($h = 1$) or to die while hospitalized ($h = 2$) if being infected ($\text{pneu} = 1$) before admission. The second set is similar, for the case when $\text{pneu} = 0$.

i) *Estimation routines for the cumulative incidence functions can be found in several R-packages, for example* `cmprsk`*,* `mstate` *or* `timereg`*. You may consider these, but for the purpose of this exercise, it suffices to use the* `survival` *package and the command, for example, (see Slides 9)*

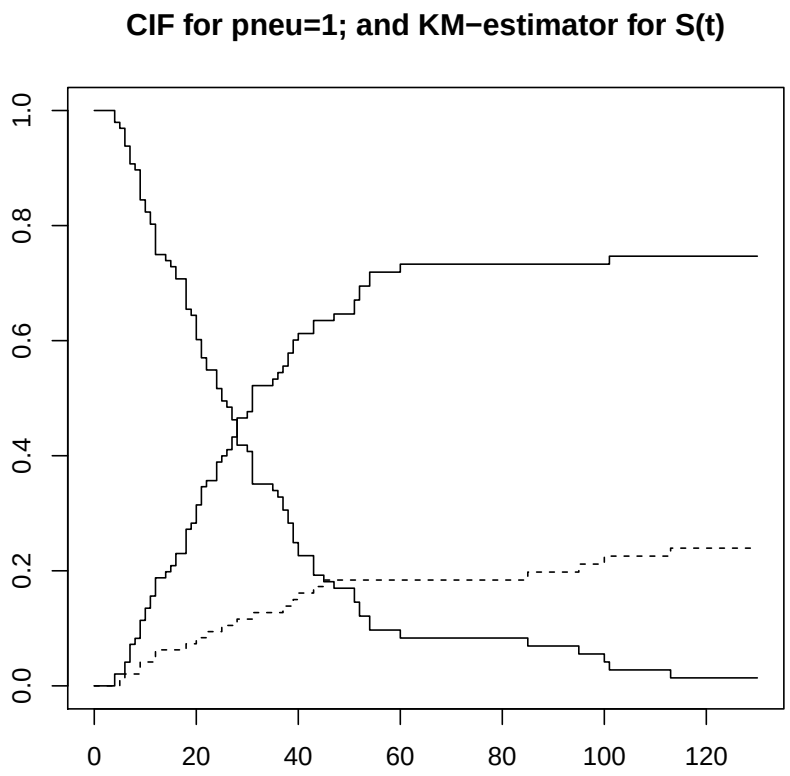
Explore the given command (see below) by using 'summary'. What does it calculate? You may then plot the two cumulative incidence functions for cause 1 and 2.

```
>cumin.pneu=survfit(Surv(time,status,type="mstate")~1,data=pneudat.pneu)
> summary(cumin.pneu)
Call: survfit(formula = Surv(time, status, type = "mstate") ~ 1, data = pneuda

 time n.risk n.event   P(1)   P(2)    P()
    4     97       2 0.0206 0.0000 0.9794
    5     95       1 0.0206 0.0103 0.9691
    6     94       3 0.0412 0.0206 0.9381
    7     91       3 0.0722 0.0206 0.9072
    8     87       1 0.0826 0.0206 0.8968
```

```
    9     86       5 0.1139 0.0415 0.8446
   10     80       2 0.1350 0.0415 0.8235
   11     78       2 0.1561 0.0415 0.8024
   12     76       5 0.1878 0.0626 0.7496
   14     71       1 0.1983 0.0626 0.7391
   15     70       1 0.2089 0.0626 0.7285
....
```

The columns P(1) and P(2) calculate the estimated CIF for cause 1 and 2, respectively. The column P() estimates the survival function of T. The three functions are plotted in the same plot by the command

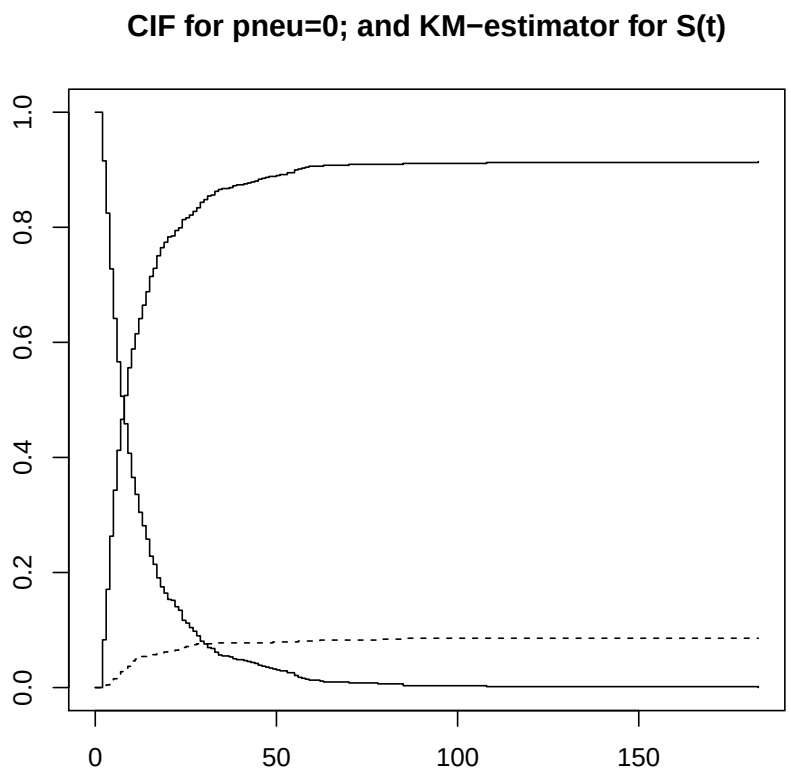
```
plot(cumin.pneu, conf.int=F,lty=c(1,2))
```



We then do the same for the data `pneudat.nopneu`.

```
> cumin.nopneu=survfit(Surv(time,status,type="mstate")~1,data=pneudat.nopneu)
```

```
> summary(cumin.nopneu)
Call: survfit(formula = Surv(time, status, type = "mstate") ~ 1, data = pneuda

 time n.risk n.event   P(1)    P(2)     P()
    2    650      55 0.0831 0.00154 0.91538
    3    595      59 0.1708 0.00462 0.82462
    4    536      63 0.2631 0.00923 0.72769
    5    473      56 0.3431 0.01538 0.64154
    6    417      49 0.4123 0.02154 0.56615
    7    368      39 0.4662 0.02769 0.50615
    8    329      31 0.5077 0.03385 0.45846
    9    295      33 0.5559 0.03695 0.40718
   10    262      27 0.5885 0.04628 0.36522
.....

plot(cumin.nopneu, conf.int=F,lty=c(1,2),main="CIF for pneu=0; and KM-estimato
```



Give an explanation of the findings, and discuss them in view of what you found by considering the estimated cumulative cause-specific hazards. Do the the two sets of plots express the same information? In what sense do they complement each other?

Solution: (Copied and edited from the book by Beyersmann et al.) Finally, we check whether our interpretation of the cumulative hazards analysis has been correct by looking at the Aalen-Johansen estimators of the cumulative incidence functions, again stratified for pneumonia status. The cumulative incidence function for death, say, displays the expected proportion of individuals dying on the unit over the course of time. If our interpretation of the cumulative hazards analysis has been correct, the estimated cumulative incidence function for death, $\hat{P}(T \leq t, H = 2)$, within patients with pneumonia should run above those patients without pneumonia. Plots of cumulative hazards as in subproblem (b) and plots of cumulative incidence functions as in the current subproblem both have their relative merits: obviously, it is easier to tell from the *cif*s whether pneumonia increases individual mortality. However, we need to look at the cumulative cause-specific hazards to see whether increased mortality is due to an increase of the death hazard, say, or Ů as in the present case Ů due to a decrease of the discharge hazard.

Solution: (Copied and edited from the book by Beyersmann et al.) The cumulative incidence function for Death displays the expected proportion of individuals dying on the unit over the course of time. If our interpretation of the cumulative hazards analysis has been correct, the estimated cumulative incidence function for death, $\hat{P}(T \leq t, H = 2)$, within patients with pneumonia should run above those patients without pneumonia. Plots of cumulative hazards as in subproblem (b) and plots of cumulative incidence functions as in the current subproblem both have their relative merits: obviously, it is easier to tell from the *cif*s whether pneumonia increases individual mortality. However, we need to look at the cumulative cause-specific hazards to see whether increased mortality is due to an increase of the death hazard, say, or Ů as in the present case Ů due to a decrease of the discharge hazard.

Problem 2. Let the situation be as in Problem 1. It might be tempting to use Kaplan-Meier estimators as alternatives to the cumulative incidence functions (*cif*) plotted in Problem 1.

Consider for example cause 2 (death) to be the interesting endpoint of the study. Then, as we did for the estimation of cause specific hazards in Problem 1, let observations with `status` = 0 (censored) or 1 (discharge) be

considered as censored observations. The so-called ***naive Kaplan-Meier estimator*** is then given by

$$1 - \hat{S}_2(t)$$

where $\hat{S}_2(t)$ is the ordinary Kaplan-Meier estimator under the above conditions. (Here the index 2 stands for cause 2).

a) *Do the appropriate estimation in R and plot the naive Kaplan-Meier estimate* $1 - \hat{S}_2(t)$ *(you may alternatively plot* $\hat{S}_2(t)$ *itself).*

Compare the naive Kaplan-Meier estimate and the estimated cif from Problem 1(i).

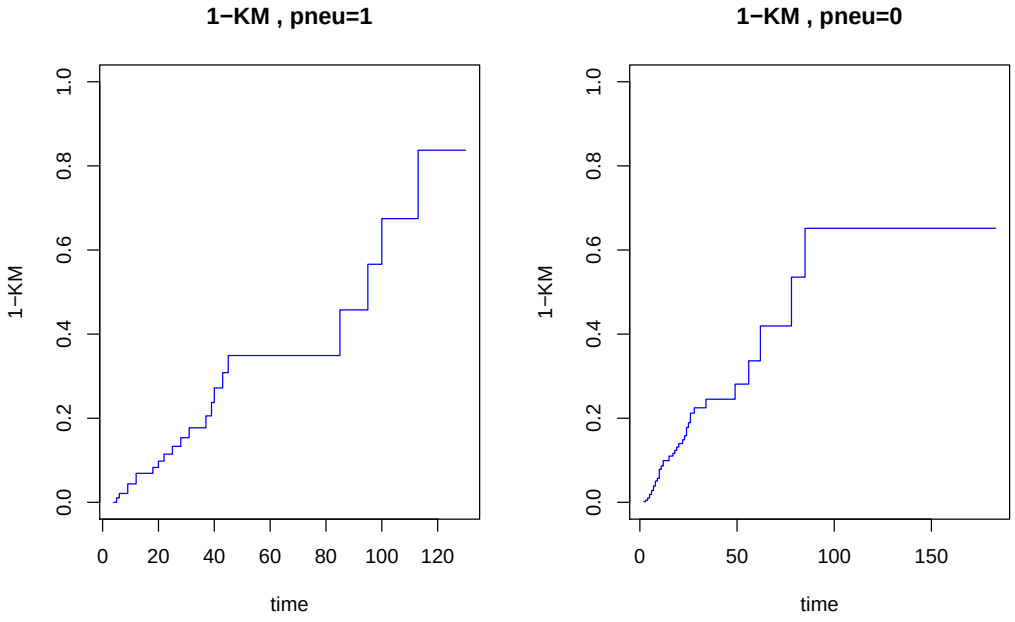
It is a well known fact, and will follow from subproblem (b) below, that the naive Kaplan-Meier estimator overestimates the true cif. You should check that this is the case for the estimated curves that you have obtained with R.

Solution:

```
fit.km=survfit(Surv(time ,status ==2)~1,data=pneudat.pneu
conf.type="plain", type="kaplan-meier"),
summary(fit.km)
Call: survfit(formula = Surv(time, status == 2) ~ 1,
data = pneudat.pneu, conf.type = "plain", type = "kaplan-meier")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5     95       1    0.989  0.0105       0.9690        1.000
    6     94       1    0.979  0.0147       0.9501        1.000
    9     86       2    0.956  0.0214       0.9141        0.998
   12     76       2    0.931  0.0273       0.8775        0.984
   18     67       1    0.917  0.0302       0.8579        0.976
   20     61       1    0.902  0.0332       0.8369        0.967
  .....

fit.km_nopneu=survfit(Surv(time ,status ==2)~1,data=pneudat.nopneu,
conf.type="plain")
summary(fit.km_nopneu)
Call: survfit(formula = Surv(time, status == 2) ~ 1, data =
pneudat.nopneu, conf.type = "plain")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2    650       1    0.998 0.00154       0.9954        1.000
```

```
  3    595       2    0.995 0.00282       0.9896        1.000
  4    536       3    0.990 0.00426       0.9812        0.998
  5    473       4    0.981 0.00593       0.9695        0.993
  6    417       4    0.972 0.00751       0.9570        0.986
  7    368       4    0.961 0.00910       0.9434        0.979
.....
```

```
par(mfrow=c(1,2))
plot(fit.km$time ,1-fit.km$surv , type="s", col="blue", main = "1-KM,
pneu=1",xlab="time", ylab="1-KM", ylim=c(0 ,1.0))
plot(fit.km_nopneu$time , 1-fit.km_nopneu$surv , col="blue",
type="s",main = "1-KM ,pneu=0", xlab="time",ylab="1-KM",ylim=c(0 ,1.0))
```



When comparing to the corresponding *cif*s for Death, it is seen that these curves are much higher. The reason will follow from the text and solution of subproblem (b) below.

b) *Assume now that the involved distributions are absolutely continuous, so that for example survival functions $S(t)$ are differentiable.*

It follows from p. 7 on Slides 9 that the cif for a specific cause h can be written

$$\int_0^t S(u)\alpha_{0h}(u)du. \tag{1}$$

Here $S(t) = P(T > t)$, where T is the failure time, irrespective of cause of failure. (This formula is the basis for the Aalen-Johansen estimator, see Problem 1(h)).

Argue that the naive Kaplan-Meier estimator can be considered as an estimator of the survival function of a lifetime "T_h" which has hazard function $\alpha_{0h}(t)$ and corresponding survival function

$$S_h(t) = e^{-\int_0^t \alpha_{0h}(u)du}.$$

Solution: This KM-estimator is generated using exactly the same data that were used above for the Nelson-Aalen estimators for the cause-specific hazards, say, $\alpha_h(u)$. Hence the KM-estimator can be considered to be estimators of $S_h(t) = \exp\{-\int_0^t \alpha_h(u)du\}$.

Show that

$$1 - S_h(t) = \int_0^t S_h(u)\alpha_{0h}(u)du \tag{2}$$

is a general identity for a survival function and its hazard rate.

Solution: The left hand side is the cumulative distribution of the corresponding lifetime, while on the right hand side, $S_h(u)\alpha_{0h}(u)$ is the density of the lifetime. The integral of this from 0 to t is the cumulative distribution function.

Compare (1) to (2) and argue that the naive Kaplan-Meier estimator overestimates the cumulative incidence function.

The *cif* equals

$$\int_0^t S(u)\alpha_h(u)du$$

where

$$S(u) = \exp\{-\int_0^t (\alpha_1(u) + \alpha_2(u))du\} \leq \exp\{-\int_0^t \alpha_h(u)du\} = S_h(u)$$