(Evaluation of) Predicted Absolute Risk

PhD course: Statistical analysis of survival data

Thomas Alexander Gerds

Outline

Prediction = Time * Probability

Prediction performance

Data splitting

Right censored

Competing risks

Example

Absolute risk of cardiovascular disease is the probability that an individual with given risk factors and a given age will be diagnosed with cardiovascular disease within a given time period.

- the longer the time-period $\mapsto$ the higher the risk
- the higher the risk of non-cardiovascular death in the same time period $\mapsto$ the lower the risk.

The absolute risk has a direct interpretation for the single patient.

Hazard rates and hazard ratios do not have an intuitive interpretation for the patient.

The purpose of risk prediction (medical research)

Statistical models that estimate a personalized probability of developing cardiovascular disease and mortality will help

- clinicians identify individuals at higher risk
- allowing for earlier or more frequent screening
- counseling of behavioral changes to decrease risk

These types of models are also useful for the design of intervention trials, for personalized medicine, and in the analysis of observational data (propensity score).

Regression vs prediction

A regression model describes how the distribution of outcome depends on covariates.

- Regression parameters describe the magnitude of association.
- The parameters are estimated based on a data set.

Regression vs prediction

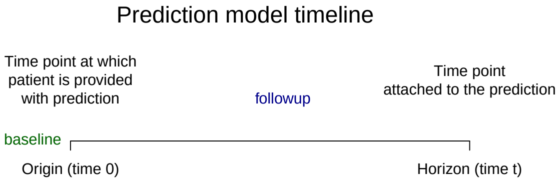
A regression model describes how the distribution of outcome depends on covariates.

- Regression parameters describe the magnitude of association.
- The parameters are estimated based on a data set.

A prediction model describes how the distribution of outcome depends on covariates.

- Predictions are expected outcomes for given covariate values.
- The model is the result of a statistical modelling strategy applied to a learning data set.
- The predictions are evaluated in independent validation data (to mimick the application in new patients).

The role of time



Prediction model timeline

Lost to followup, or (right) censored, means that patient was not followed until horizon time t.

Until time t, three things can happen:

- patient is event-free
- the event of interest has occurred
- (a competing event has occurred)

Prediction setup (no competing risks)

- Prediction time origin: time=0, $Y(0) = 0$
- Prediction time horizon: time=t, $Y(t) =?$
- Covariate vector $X \in R^p$ (age, sex, diabetes, stroke score,...)
- X must be measurable at time 0

Outcome at time horizon = t:

$$Y(t) = 1\{T < t\} = \begin{cases} 1 & \text{event between 0 and } t \\ 0 & \text{event-free at } t \end{cases}$$

Target

$$P(Y(t) = 1|X) = \text{Predicted risk of event between 0 and } t$$

Exercise 1: Predicting absolute risk using Cox regression

Consider a fitted multiple Cox regression model:

$$\hat{\lambda}_0(s) \exp(\hat{\beta}_1 X_1 + \cdots + \hat{\beta}_K X_p)$$

- $\hat{\beta}_k$ is the partial likelihood estimate of the log-hazard ratio
- $\hat{\lambda}_0$ is the Breslow estimate of the baseline hazard rate

---

1. Recall the general formula that relates the hazard rate to the survival probability.
2. Based on the Cox regression model derive the formula for the predicted t-year absolute risk for a new subject $X^{\texttt{new}}$.

Exercise 1 (continued)

```
library(riskRegression)
# help(predictRisk)
data(Melanoma)
# help(Melanoma)
coxfit <- survival::coxph(Surv(time,status!=0)∼age+sex
    +ulcer+epicel,data=Melanoma,x=TRUE)
Publish::publish(coxfit)
```

1. Evaluate the above R-code. Then interprete the hazard ratio of the variable `ulcer`
2. Use the function `predictRisk` to predict the 5-year and 10-year risks of all cause mortality for the following subjects:

age	sex	ulcer	epicel	thick
50	Male	present	present	5
50	Male	not present	present	5
50	Male	present	not present	5
50	Male	not present	not present	5

Then calculate the 5-year and 10-year absolute risk ratios separately for `epicel=present` and `epicel=not present`.

Predicting risks using other tools

Additive hazard regression

Neural Nets

Support Vector Machines

Bump hunting

Lars and his three cousins

Cart and RandomForests

Bayesian networks

Super learning

$X^{\texttt{new}} \rightarrow$ $\rightarrow \hat{R}_n(t|X^{\texttt{new}})$

Predicting risks using other tools

$X^{\texttt{new}} \rightarrow$

Additive hazard regression
~~Neural Nets~~ Deep learning
Support Vector Machines
Bump hunting
Lars and his three cousins
Cart and RandomForests
Bayesian networks
Super learning

$\rightarrow \hat{R}_n(t|X^{\texttt{new}})$

Would like to compare risk prediction models (modelling algorithms)

Prediction performance

Performance parameters: Accuracy

Brier score

$$\mathrm{Brier}(t) = E_{Y_i,X_i}\{Y_i(t) - \hat{R}_n(t|X_i)\}^2$$

Subject i was not used to train the risk prediction model $\hat{R}_n$.

Performance parameters: Accuracy

Brier score

$$\text{Brier}(t) = E_{Y_i, X_i}\{Y_i(t) - \hat{R}_n(t|X_i)\}^2$$

Subject i was not used to train the risk prediction model $\hat{R}_n$.

Interpretation

- The lower the better
- $\sqrt{\text{Brier}(t)}$ is the distance between the probabilistic prediction and the true event status at t
- The actual value of the estimated Brier score is difficult to interprete in isolation, i.e., is 0.11 good? Depends on population (distribution of X)
- The natural benchmark is the "the null model" which ignores the covariates X. It can be obtained with the Kaplan-Meier estimate of the average risk at t

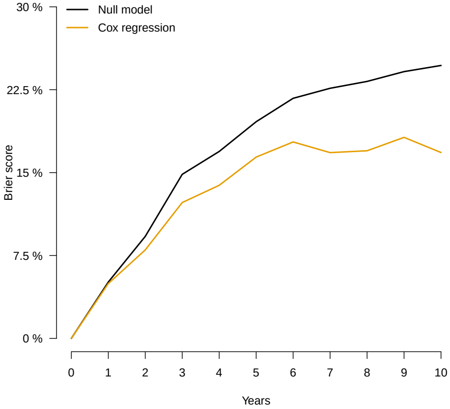
Example

```
coxfit <- survival::coxph(Surv(time,status!=0)~age+sex
    +ulcer+epicel,data=Melanoma,x=TRUE)
xx <- Score(list("Cox regression"=coxfit),
        metrics=c("auc","brier"),
        plots="calibration",
        null.model=1L,
        formula=Surv(time,status!=0)~1,
        data=Melanoma,
        se.fit=0L,
        times=seq(0,365.25*10,365.25))
xx$Brier$score[times==365.25*5]
```

model	times	Brier
Null model	1826.25	0.20
Cox regression	1826.25	0.16

Time-dependent performance

```
plotBrier(xx,axis1.at=seq(0,365.25*10,365.25),axis1.lab
    =0:10,xlab="Years")
```



Performance parameters: Discrimination

Area under the time-dependent ROC curve

$$AUC(t) = \mathrm{E}_i \mathrm{E}_j(\mathcal{I}\{\hat{R}_n(t, X_i) > \hat{R}_n(t, X_j)\} | T_i \leq t, T_j > t)).$$

Subjects i and j were not used to train the risk prediction model $\hat{R}_n$.

Performance parameters: Discrimination

Area under the time-dependent ROC curve

$$AUC(t) = \mathrm{E}_i \mathrm{E}_j(\mathcal{I}\{\hat{R}_n(t, X_i) > \hat{R}_n(t, X_j)\} | T_i \leq t, T_j > t)).$$

Subjects i and j were not used to train the risk prediction model $\hat{R}_n$.

Interpretation

- The higher the better
- Measures discrimination (invariant to monotone transformation of risks)
- The actual value of the estimated AUC is difficult to interprete in isolation, i.e., is 0.67 good? Depends on population (distribution of X)
- The natural benchmark is 0.5.

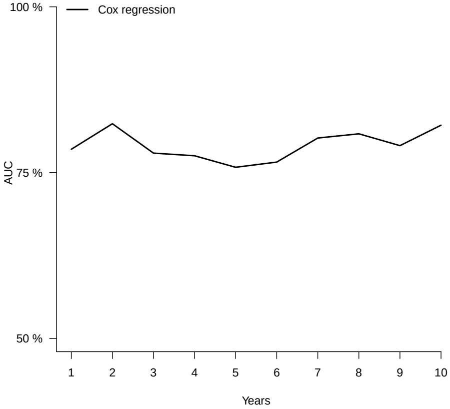
Example

```
coxfit <- survival::coxph(Surv(time,status!=0)~age+sex
    +ulcer+epicel,data=Melanoma,x=TRUE)
xx <- Score(list("Cox regression"=coxfit),
        metrics=c("auc","brier"),
        plots="calibration",
        null.model=TRUE,
        formula=Surv(time,status!=0)~1,
        data=Melanoma,
        se.fit=FALSE,
        times=seq(365.25,365.25*10,365.25))
xx$AUC$score[times==365.25*5]
```

model	times	AUC
Cox regression	1826.25	0.76

Time-dependent performance

```
plotAUC(xx,axis1.at=seq(0,365.25*10,365.25),axis1.lab
    =0:10,xlab="Years")
```



Performance parameters: Calibration

A predicted risk of 17% is reliable if it can be expected that the event will occur to about 17 out of 100 patients who all received a predicted risk of 17%.

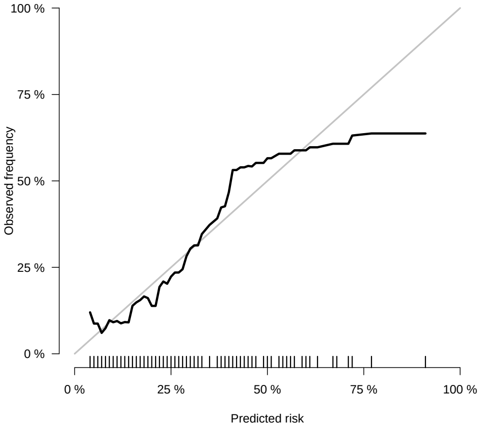
A statistical model $\hat{R}_n$ is calibrated if it provides reliable predictions for all subjects:

$$P(Y^{\mathtt{new}}(t) = 1 | \hat{R}_n(t|X^{\mathtt{new}}) = r) = r$$

A calibration plot shows predicted risks versus observed proportions. Smoothing is required to estimate the graph.

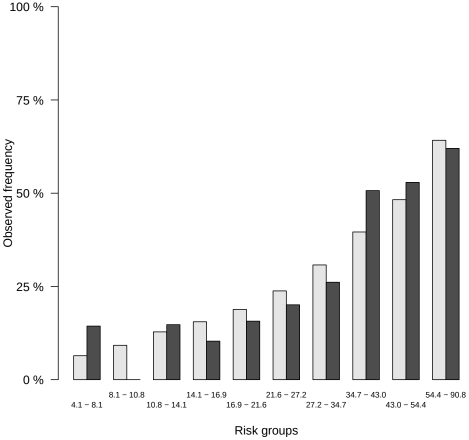
Calibration at t=5 years

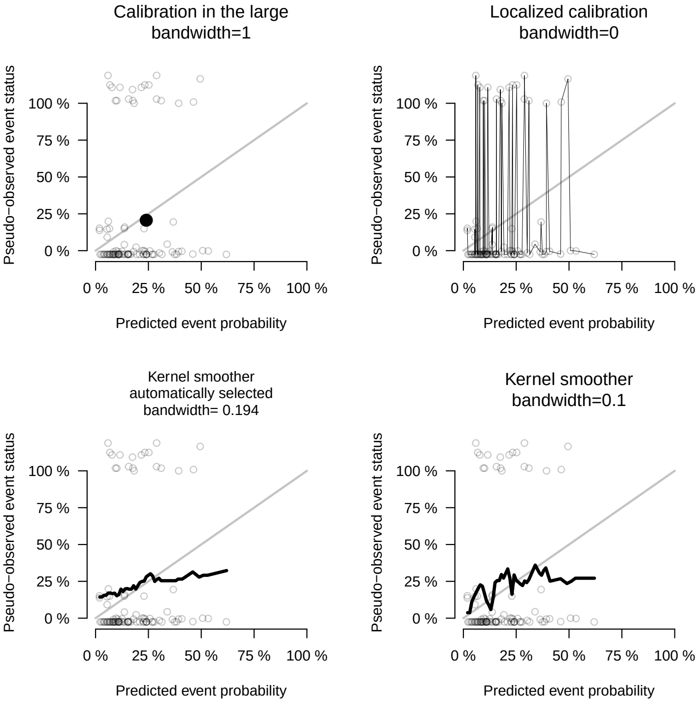
```
plotCalibration(xx,pseudo=FALSE,times=5*365.25,rug=TRUE)
```



Calibration at t=5 years

```
plotCalibration(xx,pseudo=FALSE,times=5*365.25,bars=TRUE)
```





Calibration in the large
bandwidth=1

Localized calibration
bandwidth=0

Kernel smoother
automatically selected
bandwidth= 0.194

Kernel smoother
bandwidth=0.1

Exercise 2

- Calculate the benchmark values for Brier score and AUC for the following “risk prediction models”:
  - coin toss (fair coin)
  - random number between 0 and 100%
  - always 50% risk
- Write down the interpretation of AUC(5years) in a sentence (i.e., translate the formula to english language that your grandpa could understand).
- Fit a second Cox regression model in the Melanoma data (formula below) and compare it with the model used in the previous slides regarding performance (Brier score and AUC).

```
coxfit2 <- survival::coxph(Surv(time,status!=0)~age+
    sex+ulcer+epicel+logthick,data=Melanoma,x=TRUE)
yy <- Score(list("Cox regression"=coxfit,
         "Cox (+logthick)"=coxfit2),...,se.fit=TRUE)
```

Data splitting

The estimation problem: a mission impossible

A statistical risk prediction model which only works in its own training data is practically useless.

Aim: to estimate how well the model generalizes to new data, i.e., how it will perform in

yet unseen patients

Dilemma: there are no new data!

An important distinction

Prediction modelling algorithm: R

Map a (learning) dataset to the set of risk prediction models

$$D_n = \{O_i\}_{i=1}^{n} \mapsto \mathcal{R} \qquad\qquad R(D_n) = \hat{R}_n$$

Prediction model: $\hat{R}_n$

Map patient characteristics to predicted risk for horizon t

$$(X^{\texttt{new}}, t) \mapsto [0,1] \qquad \hat{R}_n(t|X^{\texttt{new}}) \approx P(Y^{\texttt{new}}(t) = 1|X^{\texttt{new}})$$

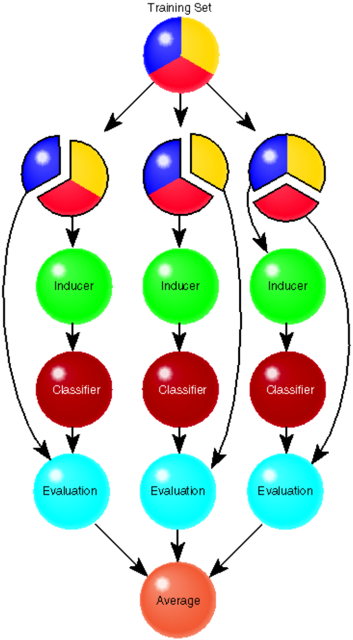
Fundamental idea

Data splitting is very intuitive:

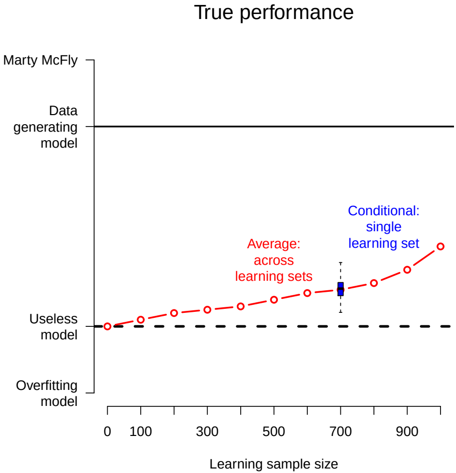
we hide one part of the data, learn on the rest, and then check our *knowledge* on what was hidden.

There is a hidden parameter here: how much we hide and how much we show.

Illustration: 3-fold CV



Learning curve



Dietterich (1998)

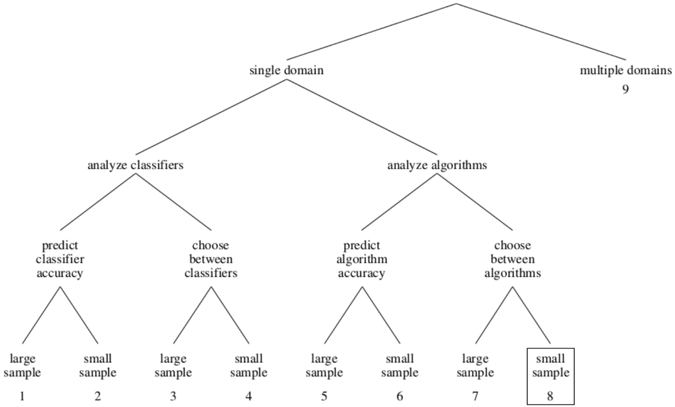
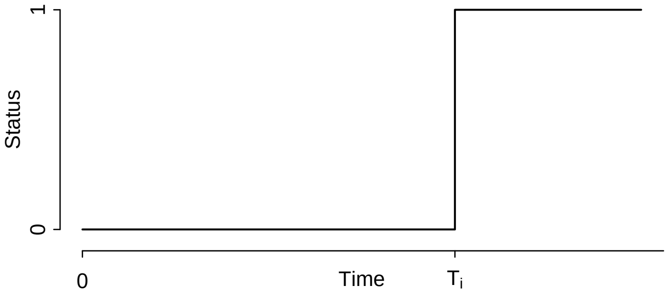


Figure 1: A taxonomy of statistical questions in machine learning. The boxed node (Question 8) is the subject of this paper.

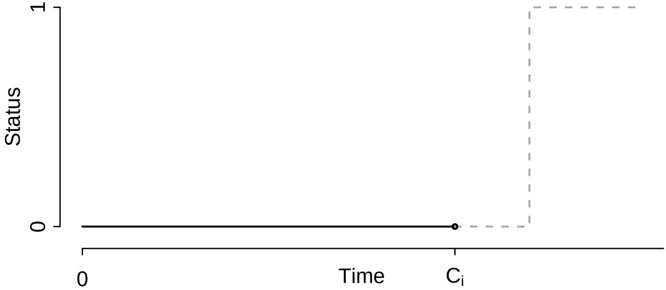
A frequently-applied strategy is to convert Question 2 into Question 6

Right censored

Uncensored observations



Right censored observations



Censored data problem

All performance parameters (expected Brier scores, concordance probabilities, etc.) depend on the unknown distribution of the future patients.

This data generating mechanism is best estimated by the empirical distribution of the validation data set.

However, when some patients are lost to follow-up (event-free) before the time horizon t, then their event status is unknown. Also the order of some pairs of patients is unknown (unusable).

In this case the performance parameters have to be estimated using some technique for right censored data to avoid bias.

Observations

Random variables

$$\begin{array}{ll} \mathrm{T} & \text{event time} \\ \mathrm{C} & \text{censoring time} \\ \tilde{T} = & \min(T, C) \\ \Delta = & 1\{T \leq C\} \\ \mathrm{X} & \text{predictors} \end{array}$$

Would like to observe: $(T, X) \sim \mathrm{Q}$

Observe: $(T, \Delta, X) \sim \mathrm{P}$

[1] Begun et al. The Annals of Statistics, 11(2):432–452, 1983.

Observations

Random variables

T	event time
C	censoring time
$\tilde{T} =$	$\min(T, C)$
$\Delta =$	$1\{T \leq C\}$
X	predictors

Would like to observe: $(T, X) \sim \mathrm{Q}$

Observe: $(T, \Delta, X) \sim \mathrm{P}$

Assume (at least) that C is conditionally independent or T given X. Then the density of the observations factorizes [1]

$$\mathrm{P}(\mathrm{d}s, \delta, \mathrm{d}x) = \{P(C > s | X = x)\mathrm{Q}(\mathrm{d}s, \mathrm{d}x)\}^{\delta}\{\dots\}^{1-\delta}$$

[1] Begun et al. The Annals of Statistics, 11(2):432–452, 1983.

Nuisance model

The estimate the distribution of the (validation) data set and to deal with censoring, we need a second model (a nuisance model) for the probability of being uncensored by time s:

$$G(s|X) = P(C > s|X)$$

Assume that at the prediction horizon t

$$G(t|x) > \epsilon > 0 \qquad \forall x$$

and suppose we have an estimate that satisfies:

$$\hat{G}_m \to G \qquad \text{when} \qquad m \to \infty$$

IPCW estimate of the expected Brier score

Weights are constructed based on the estimate $\hat{G}$ of G:

$$\omega_i(t|X_i) = \left\{ \frac{\mathcal{I}_{\{T_i \leq t, \Delta_i = 1\}}}{\hat{G}(T_i - |X_i)} + \frac{\mathcal{I}_{\{T_i > t, C_i > t\}}}{\hat{G}(t|X_i)} \right\}$$

IPCW estimate of the expected Brier score

Weights are constructed based on the estimate $\hat{G}$ of G:

$$\omega_i(t|X_i) = \left\{ \frac{\mathcal{I}_{\{T_i \le t, \Delta_i = 1\}}}{\hat{G}(T_i - |X_i)} + \frac{\mathcal{I}_{\{T_i > t, C_i > t\}}}{\hat{G}(t|X_i)} \right\}$$

The IPCW estimate of the expected Brier score in validation data

$$\frac{1}{m} \sum_{i \in V_m} \omega_i(t|X_i) \Big\{ Y_i(t) - \hat{R}_n(t|X_i) \Big\}^2 .$$

If the model for G is correctly specified the estimate is consistent.

Exercise 3

Convince yourself that the IPCW estimate of the Brier score is consistent (when $m \to \infty$).

Hints:

- Examine the two terms of the weights separately.
- For the first term use the decomposition of the density (Begun et al.)
- For the second term use the relation:

$$P(\min(T, C) > t|X) = P(C > t|X)\mathrm{P}(T > t|X)$$

Competing risks

Censored data & competing risks

A competing risk is an event after which it is clear that the patient will never experience the event of interest.

Data are called right-censored when the event time for a patient is not observed, and it is only known that the event time exceeds a certain value.

For a right censored observation it is also not observed if the subject experiences the event ever or not.

Competing risk (about the subject)

Speed = 0, arrives never!

Censored (about the observer)

Speed = ? arrival time?

Decision making in the presence of competing risks

Suppose a 40 year old and a 80 year old person need to decide for or against prophylactic coagulation therapy.

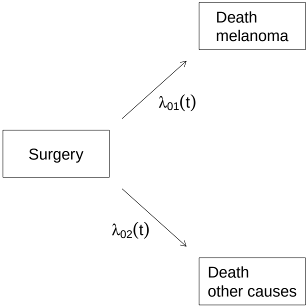
Suppose further the predicted risk of dying from cardiovascular disease within the next 10 years for both persons is 12%. How could this happen?

One plausible explanation is that the 40 year old person has other risk factors that the 80 year old person does not have.

Another plausible explanation is that the 80 year old person has a much higher risk to die due to non-cardiovascular disease within the next 10 years than the 40 year old person.

Competing risks

Any other event which changes the risk of the event being predicted may be considered a separate state in a *competing risk model*. Most commonly this would be dying from other causes.



Competing risks affect all stages of the process from the discovery of markers over modelling and assessment of risk predictions to medical decision making.

Setup

- Event time: T (time between 0 and an event)
- Cause of the event: $D \in \{1, .., K\}$
- Covariate vector X= X_1, . . . , X_p

Event status

$$N_k(t) = 1\{T \leq t, D = k\}$$

Assess performance of personalized risk predictions

$$\hat{R}_n(t|X_i) \approx \mathrm{P}\{T_i \leq t, D_i = k|X_i\} = F_k(t|X_i)$$

Parameters in the presence of competing risks

- The cause-k specific hazard function for a subject characterized by covariate vector X

$$\lambda_k(s|X) \approx \text{the probability of an event of type k tomorrow [s]}$$
$$\text{given no event until today [s-].}$$

- The cause-k specific cumulative incidence function

$$\begin{aligned} F_k(t|X) &= \mathrm{P}\{T \leq t, D = k|X\}. \\ &= \text{absolute risk of event k before time t} \end{aligned}$$

Relation between hazards and absolute risks

$$F_k(t|X) =$$

$$\int_0^t \underbrace{\exp\left(-\int_0^s \{\lambda_1(u|X) + \cdots + \lambda_K(u|X)\}\mathrm{d}u\right)}_{\text{No event of any cause until s}} \quad \underbrace{\lambda_k(s|X)}_{\text{Event type k at s}} \quad \mathrm{d}s.$$

- a covariate that reduces the cause-specific hazard of a competing risk indirectly increases the cumulative incidence of event j.
- covariates found to change F_k are those that change any of the cause-specific hazard functions.

Different tasks require different methods

We focus on the cause-k specific hazard to identify variables that affect the biology of cause-k events.

We focus on the cumulative incidence(s) to predict the risk(s), e.g., for patient counseling.

Prediction of absolute risk (formula I)

$F_1(t|X) =$ Cumulative incidence of event 1

$$\int_0^t \underbrace{\exp\left(-\int_0^s \{\hat\lambda_1(u|X) + \hat\lambda_2(u|X)\}\mathrm{d}u\right)}_{\text{No event of any cause until s}} \underbrace{\hat\lambda_1(s|X)}_{\text{Event type 1 at s}} \quad \mathrm{d}s.$$

- *Cox regression for events of type 1, e.g., stroke hazard*:

$$\hat\lambda_1(u|X) = \hat\lambda_{01}(u)\exp(\hat\beta X)$$

- *Cox regression for competing events, e.g., hazard of death other causes*:

$$\hat\lambda_2(u|X) = \hat\lambda_{02}(u)\exp(\hat\gamma X)$$

Melanoma example: hazard of cancer death

```r
library(riskRegression)
library(Publish)
data(Melanoma)
cscfit <- CSC(Hist(time,status)~age+sex+ulcer+epicel,
    data=Melanoma)
publish(cscfit$models[[1]])
```

Variable	Units	HazardRatio	CI.95	p-value
age		1.02	[1.00;1.04]	0.0359
sex	Female	Ref		
	Male	1.75	[1.03;2.97]	0.0398
ulcer	not present	Ref		
	present	3.43	[1.89;6.23]	<0.001
epicel	not present	Ref		
	present	0.48	[0.26;0.88]	0.0181

Melanoma example: hazard of death due to other causes

```r
library(riskRegression)
library(Publish)
data(Melanoma)
cscfit <- CSC(Hist(time,status)~age+sex+ulcer+epicel,
    data=Melanoma)
publish(cscfit$models[[2]])
```

Variable	Units	HazardRatio	CI.95	p-value
age		1.07	[1.03;1.12]	0.0011
sex	Female	Ref		
	Male	1.33	[0.45;3.93]	0.6082
ulcer	not present	Ref		
	present	1.35	[0.44;4.14]	0.5969
epicel	not present	Ref		
	present	1.49	[0.48;4.70]	0.4916

Right censored observations

Random variables

T	event time
D	type of event
C	censoring time
$\tilde{T} =$	$\min(T, C)$
$\Delta =$	$1\{T \leq C\}$
$\tilde{D} =$	D * Δ
X	predictors

Assume (at least) that C is conditionally independent or T given X.

Otherwise the joint distribution of (T, X) is not identifiable from the observations.

Brier score in the presence of competing risks

The formula and interpretation does not change

$$\text{Brier}(t,k) = \mathrm{E}_{T_i,D_i,X_i}\{N_{ik}(t) - \hat{R}_n(t|X_i)\}^2$$

- Null model is the Aalen-Johansen estimate

IPCW estimate of the expected Brier score (cause k)

$$\frac{1}{m}\sum_{i\in V_m}\left\{\frac{\mathcal{I}_{\{T_i\leq t,\Delta_i=1\}}}{\hat{G}(T_i-|X_i)}+\frac{\mathcal{I}_{\{T_i>t,C_i>t\}}}{\hat{G}(t|X_i)}\right\}\Big\{N_{ik}(t)-\hat{R}_n(t|X_i)\Big\}^2.$$

If the working model is correctly specified the estimate is consistent.

AUC in the presence of competing risks

$$AUC_k(t) =$$

$$\mathrm{E}_i\mathrm{E}_j(\mathcal{I}\{\hat{R}_n(t, X_i) > \hat{R}_n(t, X_j)\}|D_i = k, T_i \leq t, (T_j > t \text{ or } D_j \neq k))$$

AUC$_1$(t) is the probability that a random subject who experienced an event of type D=1 has received a higher predicted risk at baseline than another random subject who did not experience event of type D=1 within t years, i.e., is either alive and event free or died before time t without event type D=1.

IPCW estimate of AUC (cause k)

Weights:

$$\hat{W}_{ij,1} = \frac{\mathcal{I}_{\{T_i < T_j, T_i < C_i\}}}{\hat{G}(T_i - |X_i)\hat{G}(T_i|X_j)}$$

$$\hat{W}_{ij,2} = \frac{\mathcal{I}_{\{T_i \geq T_j, D_j \neq k, T_j \leq C_j\}}}{\hat{G}(T_i - |X_i)\hat{G}(T_j - |X_j)}$$

IPCW estimate:

$$\frac{\sum_{i,j \in V_m} \left(\hat{W}_{ij,1} + \hat{W}_{ij,2}\right) \mathcal{I}_{\{\hat{R}_n(t,X_i) > \hat{R}_n(t,X_j)\}} N_{ik}(t)}{\sum_{i,j \in V_m} \left(\hat{W}_{ij,1} + \hat{W}_{ij,2}\right) N_{ik}(t)}$$

Note: $\mathcal{I}_{\{\hat{R}_n(t,X_i) > \hat{R}_n(t,X_j)\}}$ may change over time.

Summary and conclusions

In survival analysis predictions and prediction performance are time-dependent. Performance is a parameter (of the distribution of the validation data) which needs to be estimated usually in the presence of right censored data.

With competing risks the way we do things need some slight adaptation, there are pitfalls, and some formula get more complex, but generally everything seems to be under control.

To interprete a prediction in most applications with competing risks we have to build several models, one for each competing risk.

Exercise 4

Compare the following two cause-specific Cox regression models with respect to ability to predict 5-year risks of cancer related death (Brier score and AUC):

```
cscfit1 <- CSC(Hist(time,status)~age+sex+ulcer+epicel,
    data=Melanoma)
cscfit2 <- CSC(Hist(time,status)~age+sex+ulcer+epicel+
    logthick,data=Melanoma)
zz <- Score(list(csc=cscfit1,"csc logthick"=cscfit2),
        formula=Hist(time,status)~1,
        data=Melanoma,
        cause=1,
        times=5*365.25)
```

Compare the results to that of Exercise 2.