Survival Analysis I (CHL5209H)

Olli Saarela

Dalla Lana School of Public Health
University of Toronto

olli.saarela@utoronto.ca

March 13, 2019

Prognostic modeling

Medical gnosis

- Medical research aims to advance the knowledge-base of gnosis in medicine (Miettinen 2011, p. 139).
- Miettinen (2011, p. 15):

  Gnosis - In medicine, a doctor's esoteric knowing about the health of a/the client

- The three subtypes of medical knowing are:
  1. Diagnosis
  2. Prognosis (descriptive or intervention-prognosis)
  3. Etiognosis

Prognosis

- Prognosis is defined by Miettinen (2011, p. 22) as

  Prognosis - A doctor's esoteric knowing about the future course and/or outcome of a/the client's health, specifically in respect to a particular illness (cf. 'Diagnosis' and 'Etiognosis')

- This could involve knowing about whether a currently absent illness will occur in the future, or the outcome of an already existing illness; Miettinen (2011, p. 23):

  Clinical prognosis - A doctor's (clinician's) esoteric knowing about whether a particular, currently absent illness (overt) will occur; also: regarding an already-existing illness, such knowing (probabilistic) about an adverse event/state (treatment induced perhaps) in its course and/or as its outcome

Prognostic factors

- However, the established terminology makes a distinction between predicting disease for presently healthy individuals, and predicting the outcome of presently diseased individuals (only latter would be 'prognostic').
- In the same meaning, we make a distinction between 'risk factors' and 'prognostic factors'.
- When referring to markers, rather than possible causal factors, terms 'risk indicators' and 'prognostic indicators' might be more appropriate. Miettinen (2011, p. 93):

  > *Prognostic indicators for adverse events/states are properly termed risk indicators; they need not be risk factors.*

- Further, 'predictive factor' is used to refer to something that predicts a response to a treatment (again, not necessarily causal, so different from 'effect modifier'). Same factors can be both prognostic and predictive in this sense.

Good prognosis?

- Miettinen (2011, p. 14):

  Good diagnosis/etiognosis/prognosis - One with probability close to that of correct diagnosis/etiognosis/prognosis.
  Note: 'Good prognosis' is commonly attributed to an illness, as a common misnomer for not-so-bad course, 'bad prognosis' being its corresponding misnomer for bad course. However, prognosis actually is a cognitive entity, possible only for a doctor to have; as the illness of a doctor's patient does not have a mind, it cannot have prognosis.

- Importantly, good prognosis in this sense does not require knowing the patient's outcome with certainty.

Prognostic probabilities

- Miettinen (2011, p. 22):

  Note 2: Clinical prognosis is knowing about the correct probability of the event's occurring or the state being present in/at a particular period/point of prognostic time. Correct prognosis is characterized by this probability, which represents the proportion of instances of the profile in general (in the abstract) such that, given the intervention, the event/state would occur in/at that period/point of prognostic time. (That proportion is implied by a suitable prognostic probability function.)

- We can obtain such prognostic probability functions by fitting a suitable survival model.

Prognostic probabilities (2)

- In the absence of competing causes (e.g. when the event of interest is death due to any cause), the prognostic probability is simply the s-year risk of the event occurring:

$$\pi_i(s) = 1 - \exp\{-\Lambda_i(s)\}.$$

- For example, using a Cox model, this could be estimated as

$$\hat{\pi}_i(s) = 1 - \exp\left\{-\hat{\Lambda}_0(s)\exp\{\hat{\beta}' x_i\}\right\},$$

where x_i are the predictors available at the time of prediction (remember, we cannot use future information here), and $\hat{\Lambda}_0(s)$ is given by the Breslow estimator.

- Note that we don't predict risks; we predict the outcome event using the risk as the measure.

Competing causes

- When we are not looking at all-cause mortality, but say, cause-specific mortality, in principle we have to take into account that a death due to a specific cause is preceded by survival from all causes (more about this later).
- In this case, the risk of event type j occurring first is obtained from the cause-specific cumulative incidence function:

$$\pi_{ij}(s) = \int_0^s \lambda_{ij}(t) S_i(t) \, \mathrm{d}t,$$

where

$$S_i(t) = \exp\left\{ -\sum_{j=1}^{J} \Lambda_{ij}(t) \right\}$$

is the overall survival function.

- Each one of the cause-specific cumulative hazard functions could be estimated through a Cox model as

$$\hat{\Lambda}_{ij}(t) = \hat{\Lambda}_{0j}(t) \exp\{\hat{\beta}_j' x_i\}.$$

Model validation

- How good are the model-based risks $\hat{\pi}_i(s)$ in predicting the outcome?
- This depends on the chosen criterion for 'good', but presumably this could be studied by comparing the risks to the actual observed outcomes.
- This would be referred to as model validation.
- Validation can be either internal (using the same dataset where the model was fitted), or external (using an independent dataset for validation).
- Two particular aspects of 'goodness' of the predictions would be how well they discriminate between those who will experience an event in the future and those who don't (discrimination), and how well the predictions match with the observed absolute risks in different subgroups (calibration).

Sensitivity and PPV

- Let π^* be a given threshold risk, maybe related to some clinical decision.
- In the absence of censoring, sensitivity could be defined as the probability that an individual who will experience the outcome event will have an estimated risk above the threshold (true positive), that is,

$$P(\hat{\pi}_i(s) \geq \pi^* \mid \tilde{N}_i(s) = 1).$$

- At this point we have fixed the risk model parameters to their estimates, so the probability here refers to the probability of individual i having predictor values that give a risk above the threshold.
- An alternative measure would be the positive predictive value

$$P(\tilde{N}_i(s) = 1 \mid \hat{\pi}_i(s) \geq \pi^*).$$

Specificity and NPV

- Sensitivity reflects how well the risk model identifies the individuals who will experience the event. On the other hand, specificity reflects how well the model identifies those who will not (true negative). This is the probability

$$P(\hat{\pi}_i(s) < \pi^* \mid \tilde{N}_i(s) = 0).$$

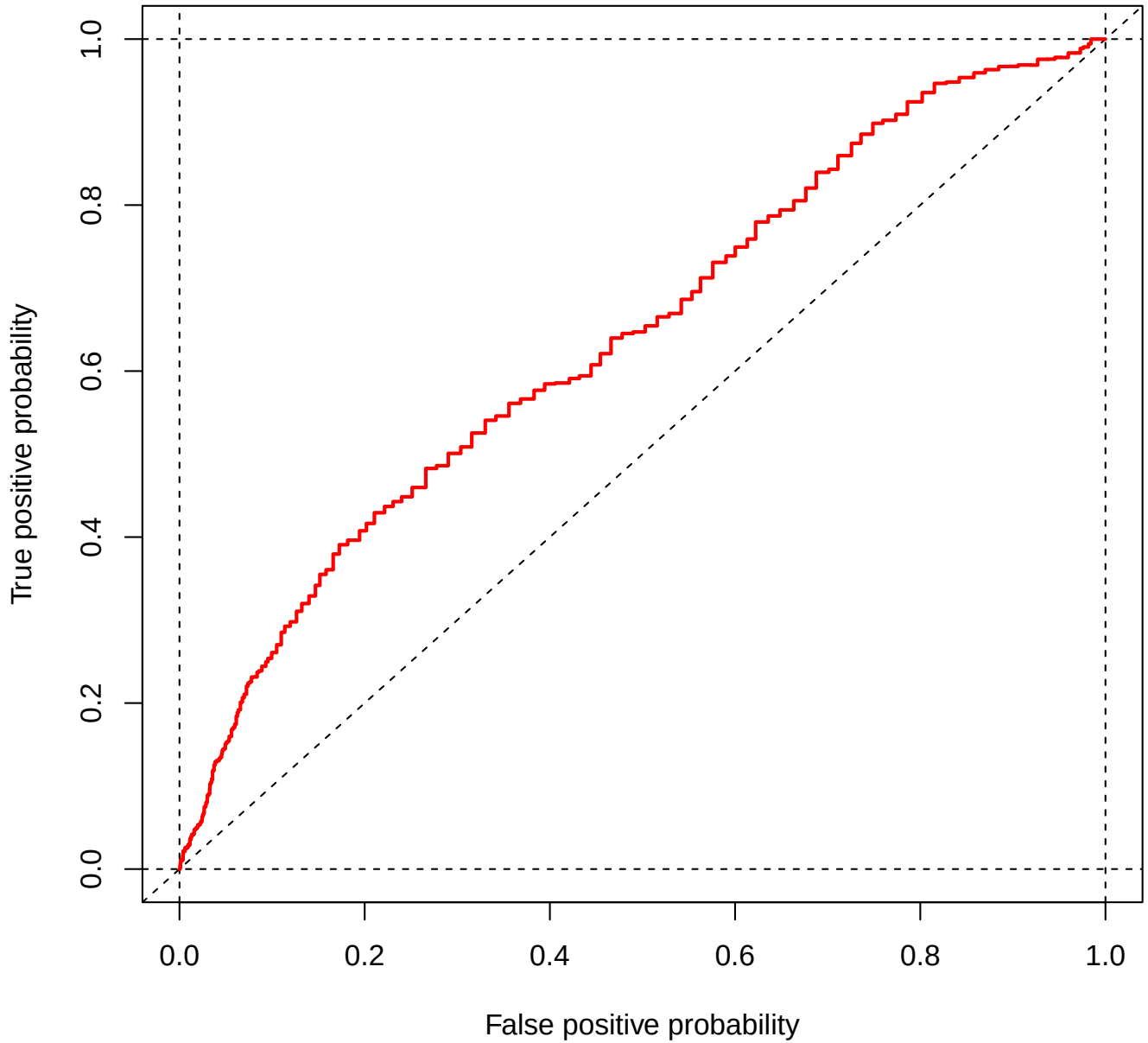
- An alternative measure would be the negative predictive value

$$P(\tilde{N}_i(s) = 0 \mid \hat{\pi}_i(s) < \pi^*).$$

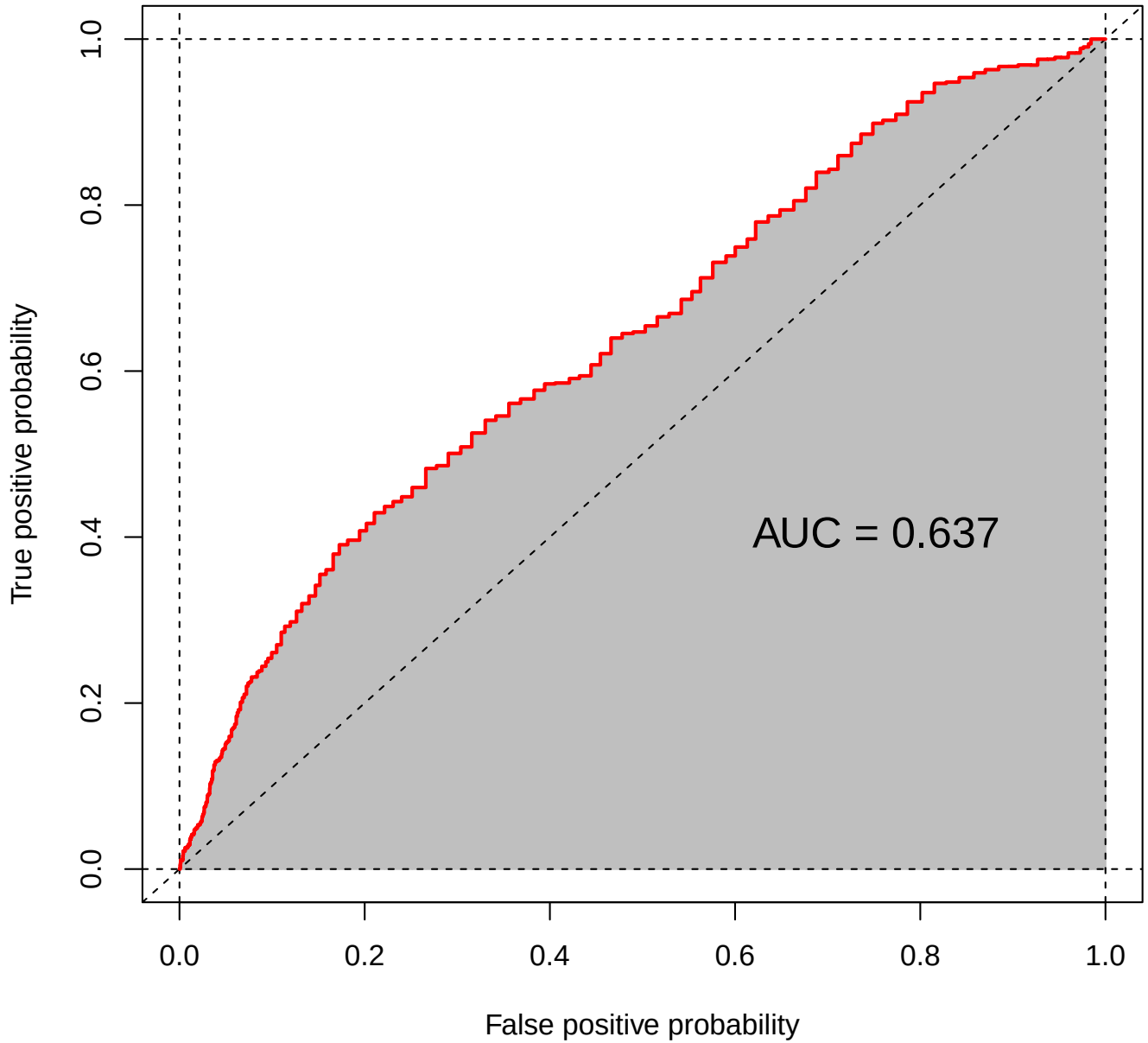
ROC curve

- There is a tradeoff between sensitivity and specificity; higher values of the threshold π^* give better specificity, but worse sensitivity, and vice versa. (Why?)
- Since we usually don't have a well-established threshold risk, we would usually calculate the sensitivity and 1-specificity (i.e. false positive probability) at all possible values of π^* and present these as a curve. The result is known as the receiver operating characteristics (ROC) curve.
- Note that when the predictors in the model have no prognostic value whatsoever, we have that $TPP = P(\hat{\pi}_i(s) \geq \pi^* \mid \tilde{N}_i(s) = 1) = P(\hat{\pi}_i(s) \geq \pi^*)$ and $FPP = P(\hat{\pi}_i(s) \geq \pi^* \mid \tilde{N}_i(s) = 0) = P(\hat{\pi}_i(s) \geq \pi^*)$, which means that the ROC curve is a diagonal line.

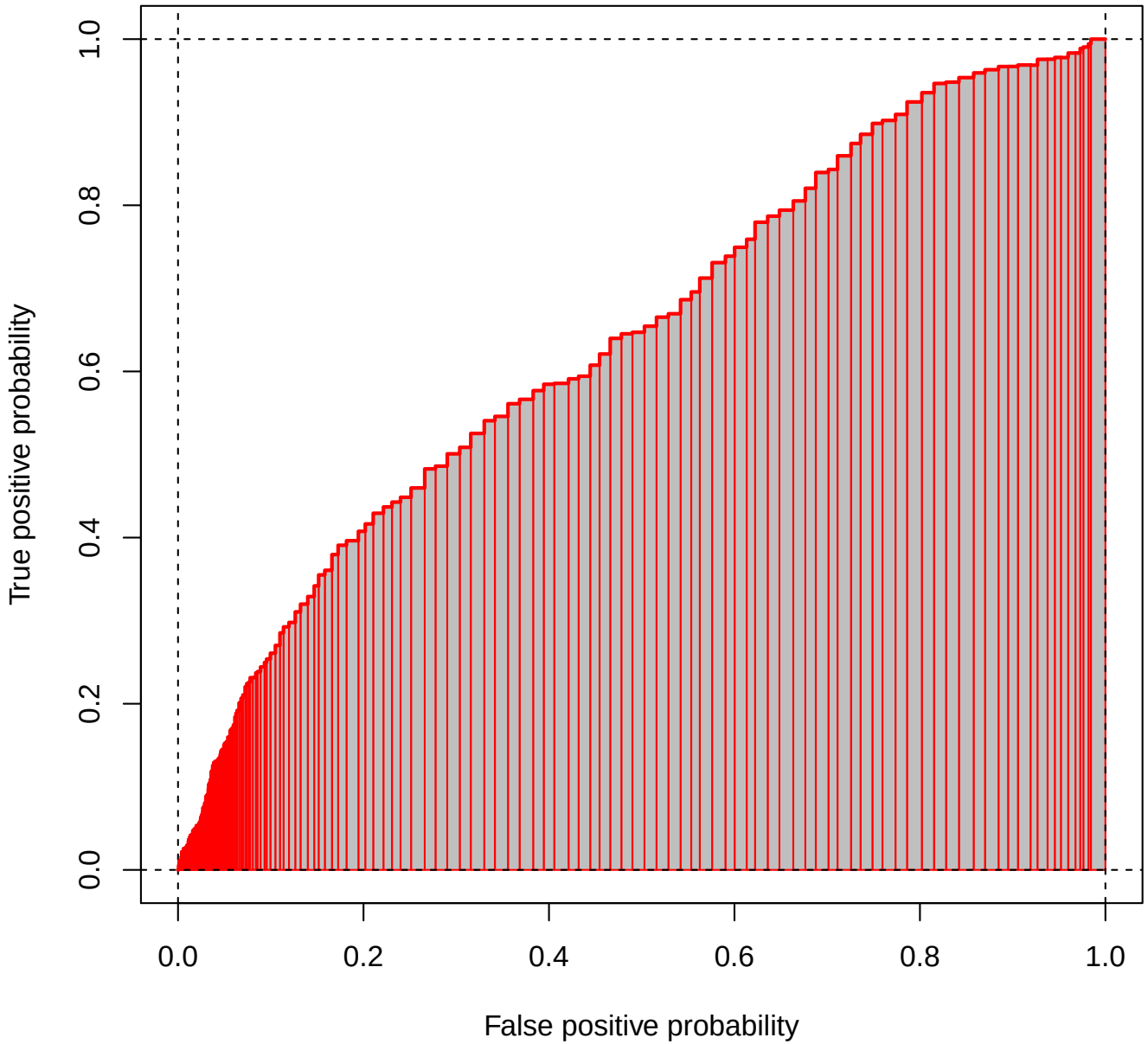
ROC curve


True positive probability
0.0
0.2
0.4
0.6
0.8
1.0
False positive probability
0.0
0.2
0.4
0.6
0.8
1.0

AUC



Calculating the AUC



AUC

- The area under the curve has a probabilistic interpretation, namely that the model correctly orders the risks of two individuals with and without an event, that is,

$$P(\hat{\pi}_i(s) > \hat{\pi}_j(s) \mid \tilde{N}_i(s) = 1, \tilde{N}_j(s) = 0).$$

- If $AUC = 1$, the model can always discriminate between the individuals with and without an event.
- In the absence of censoring, this could be estimated simply by calculating the proportion of concordant pairs.
- For non-censored event times, an analogous measure could be defined as

$$P(\hat{\pi}_i(s) > \hat{\pi}_j(s) \mid \tilde{T}_i < \tilde{T}_j).$$

- How to estimate this in the presence of censoring?

c-index

- One possible solution: compare only those censored/non-censored pairs where the observed time T_j of the censored individual j is longer than the observed time T_i of the non-censored individual i.
- Non-censored/non-censored pairs can be compared, with concordance meaning that $\hat{\pi}_i(s) > \hat{\pi}_j(s)$ for a pair with $T_i < T_j$.
- The resulting statistic is known as the concordance index, or c-index (Harrell et al. 1996), and is calculated automatically in the R `coxph` output.
- This can also be calculated using the `survConcordance` function of the survival package.

ROC curves and censoring

- How can we estimate sensitivity and specificity in the presence of censoring?
- Heagerty et al. (2000): use Bayes formula to get

$$P(\hat{\pi}_i(s) \geq \pi^* \mid \tilde{N}_i(s) = 1)$$
$$= \frac{[1 - P(\tilde{N}_i(s) = 0 \mid \hat{\pi}_i(s) \geq \pi^*)][1 - P(\hat{\pi}_i(s) < \pi^*)]}{1 - P(\tilde{N}_i(s) = 0)}$$

and

$$P(\hat{\pi}_i(s) < \pi^* \mid \tilde{N}_i(s) = 0)$$
$$= \frac{P(\tilde{N}_i(s) = 0 \mid \hat{\pi}_i(s) < \pi^*)P(\hat{\pi}_i(s) < \pi^*)}{P(\tilde{N}_i(s) = 0)}.$$

- Here the probabilities $P(\tilde{N}_i(s) = 0 \mid \cdot)$ can be estimated through the Kaplan-Meier method (see package `survivalROC`), and $P(\hat{\pi}_i(s) < \pi^*)$ through the ECDF of the risks.

Correcting for overfitting

- Validating the model in the same dataset where it was fitted will generally result in overoptimistic results.
- Ideally we would like to have a separate validation dataset, but in the absence of this, we can calculate the risks as

$$\hat{\pi}_i(s) = 1 - \exp\left\{-\hat{\Lambda}_{0(-i)}(s)\exp\{\hat{\beta}'_{(-i)}x_i\}\right\},$$

where $\hat{\Lambda}_{0(-i)}$ and $\hat{\beta}_{(-i)}$ are the baseline cumulative hazard and regression parameter estimates when observation i has been removed from the data. This is repeated for each $i = 1, \ldots, n$.

- This procedure is known as leave-one-out cross-validation; now the same observation is never used for both fitting the model, and for validating it.

Calibration measures

- A standard way to check for model calibration would be to divide the data into K groups based on quantiles of the risk estimates, and compare the expected and observed numbers of events in these groups.
- The comparison can be made using the Hosmer-Lemeshow test statistic:

$$\sum_{k=1}^{K} \frac{(O_k - E_k)^2}{N_k \bar{\pi}_k (1 - \bar{\pi}_k)} \sim \chi^2_{K-2}.$$

- Here $E_k = N_k \bar{\pi}_k$ and $\bar{\pi}_k$ is the average of the estimated risks in group k.
- For the χ^2 approximation to apply, K should be larger than # of predictors + 1, though $K = 10$ is common.
- In the presence of censoring, O_k could be estimated as $O_k \approx N_k[1 - \hat{S}_k(s)]$, where $\hat{S}_k(s)$ is the Kaplan-Meier survival probability in group k.
- The expected and observed counts E_k and O_k can also be compared visually in a calibration plot.

References

- Heagerty P, Lumley T, Pepe MS (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 56, 337–344.
- Harrell FE, Lee KL, Mark DB (1996). Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 15, 361–387.
- Miettinen OS (2011). Epidemiological Research: Terms and Concepts. Springer, Dordrecht.