

Medical Risk Prediction Models

With Ties To Machine Learning

In memory of my very good friend Michael W Kattan

Outline ¹

- Why should I care about statistical prediction models?
- I am going to make a prediction model. What do I need to know?
- How should I prepare for modeling?
- I am ready to build a prediction model
- Does my model predict accurately?
- How do I decide between rival models?
- What would make me an expert?
- Can't the computer just take care of all of this?

¹Medical risk prediction models: with ties to machine learning. Chapman and Hall/CRC, 2021.

Right on

The only useful function of a statistician is to make predictions, and thus to provide a basis for action – W. Edwards Deming

There are many ways to make a model, and every modeling expert has preferences regarding the general approach and tuning.

Prediction model timeline



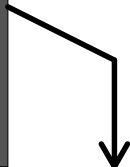
Censored means that patient was event free at the last contact but not followed until prediction time horizon t . Event can still happen but this is not observed.

Competing risk means that patient will never experience the event.

Learning
dataset

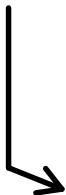


Regression
Smoothing
Tuning
Stacking
Penalization
Machine learning
Artificial intelligence



Risk
prediction
model

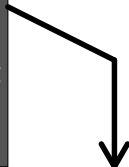
Patient
characteristics



Purpose dataset

Risk prediction
model

Prediction time horizon t



Predicted
risk
of event
until time t

Notation

Outcome

$$Y(t) = \begin{cases} 0 & \text{event-free or competing risk} \\ 1 & \text{event of interest before time } t \end{cases}$$

Predictors

$$X = (X^1, \dots, X^p)$$

Dataset

$$D_n = (Y_1(t), X_1, Y_2(t), X_2, \dots, Y_n(t), X_n)$$

Building the model

$$r : D_n \mapsto r(D_n) = \hat{M}_n$$

Using the model

$$\hat{M}_n : X_{\text{new}} \mapsto \hat{M}_n(X_{\text{new}}) \in [0, 1]$$

Example: logistic regression

$$\hat{M}_n(X) = \text{expit}(\hat{\alpha}_n + \hat{\beta}_n X)$$

Measuring prediction performance

Calibration:

$$p \mapsto P\{Y_i(t) = 1 | \hat{M}_n(X_i) = p\}$$

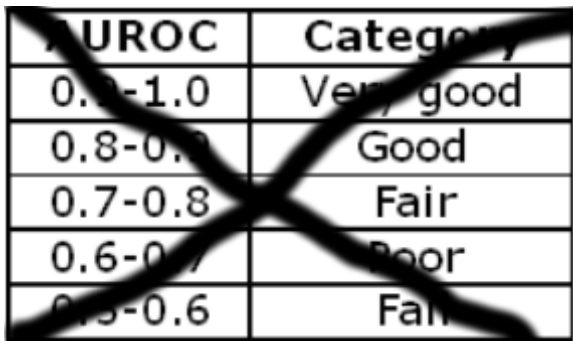
Discrimination:

$$\text{AUC}(t) = P(\hat{M}_n(X_i) \geq \hat{M}_n(X_j) | Y_i(t) = 1, Y_j(t) = 0)$$

Overall accuracy:

$$\text{Brier score}(t) = E \left\{ Y_i(t) - \hat{M}_n(X_i) \right\}^2$$

Interpretation of a prediction performance measure should always involve a benchmark, ideally that set by a rival prediction model.



AUROC	Category
0.9-1.0	Very good
0.8-0.9	Good
0.7-0.8	Fair
0.6-0.7	Poor
0.5-0.6	Fair

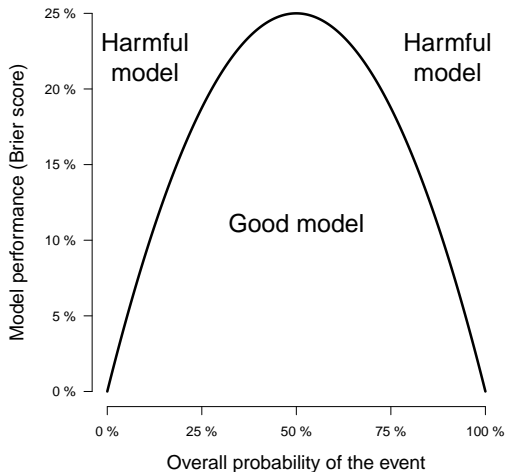
In a homogeneous population, even the perfect model can have low discrimination ability (AUC/AUCROC). In a heterogeneous population, even a bad model can have a high AUC.

Benchmark values for the AUC (concordance index) and Brier score at any fixed prediction time horizon t .

Benchmark prediction	AUC	Brier score	Interpretation
50% always	50%	25%	useless or harmful
Overall event probability always	50%	see Figure	useless
Coin toss	50%	50%	harmful
Uniform $[0,1]$	50%	33%	harmful

Being a rank statistic, the AUC is blind to miscalibration of the predicted risks. Hence, it cannot stand alone to assess models with respect to predictive accuracy.

Benchmark: the null model ignores the predictor variables



A good model outperforms the null model.

Example

```
library(riskRegression)
train <- readRDS("./practicals/data/type1-diabetes-train.rds")
test <- readRDS("./practicals/data/type1-diabetes-test.rds")
head(train)
```

Example

```
library(riskRegression)
train <- readRDS("./practicals/data/type1-diabetes-train.rds")
test <- readRDS("./practicals/data/type1-diabetes-test.rds")
head(train)
```

	pid	age	sex_male	diabetes_duration	smoking	motion	steno_prs
1	train-28abed0e	52.92496	0	40.74757	0	0	-0.13899942
2	train-f07f8cea	42.70053	0	36.11731	0	1	0.40351668
3	train-f343cabb	47.57017	0	29.81477	0	0	0.54502533
4	train-beb0e521	50.20713	0	40.73831	0	0	0.23982291
5	train-05d45f09	40.76193	0	22.58828	0	1	-0.55949839
6	train-82772fab	45.96690	0	28.94583	0	1	0.04637945
	eGFR_pre_trt	statin	HBA1C_post_trt	urine_albumin_post_trt	LDL_post_trt	SBP_po	
1	104.03674	0	49.51156	95.958599	8.816790	116	
2	90.13102	0	35.53323	44.175849	6.772076	97	
3	102.67096	0	75.10901	96.535995	6.853455	135	
4	113.77416	0	72.04421	39.454438	4.687593	130	
5	104.25661	0	28.79908	5.671578	1.392458	105	
6	98.40765	0	68.75695	230.856313	1.389660	124	

Conventional model and experimental model

```
# Logistic regression similar to Steno 1 risk engine
  formula
conventional_model <- glm(cvd_5year~sex_male + age +
  diabetes_duration + smoking + motion + HBA1C_post_
  trt + urine_albumin_post_trt + LDL_post_trt + SBP_
  post_trt + eGFR_post_trt,
                        data = train, family = "binomial")
# Logistic regression with interactions and reduced
  number of variables
experimental_model <- glm(cvd_5year~sex_male *SBP_post
  _trt + age + I(age>40) * eGFR_post_trt + diabetes_
  duration + smoking + motion,
                        data = train, family = "binomial")
```

Predicted risks for a single subject

```
data.frame("subject id" = c(24,24),  
           "model" = c("conventional","experimental"),  
           "risk prediction" = c(predictRisk(  
conventional_model,newdata = test[24,]),  
                                predictRisk(  
experimental_model,newdata = test[24,])))
```

	subject.id	model	risk.prediction
1	24	conventional	0.1826537
2	24	experimental	0.4429426

Evaluating model performance

```
x <- Score(list("Conventional model" = conventional_
  model, "Experimental model" = experimental_model),
  data = test,
  formula = cvd_5year~1,
  summary = "risks",
  plots = c("roc", "cal"))
summary(x, what = "score")
```

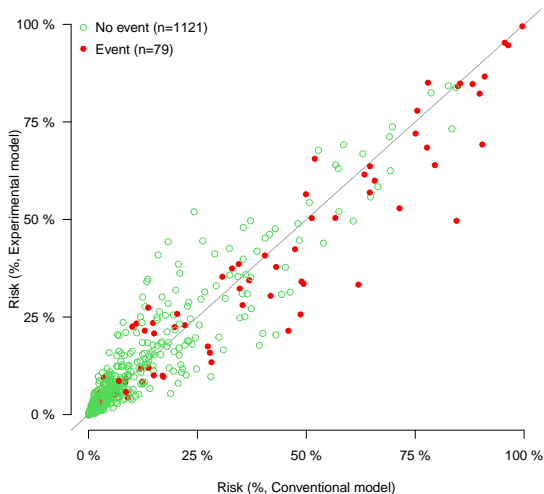
\$score

Key: <Model>

	Model	AUC (%)	Brier (%)
	<fctr>	<char>	<char>
1:	Null model	<NA>	6.1 [4.9;7.4]
2:	Conventional model	88.3 [84.6;92.0]	4.6 [3.7;5.5]
3:	Experimental model	87.8 [84.1;91.5]	4.9 [4.0;5.8]

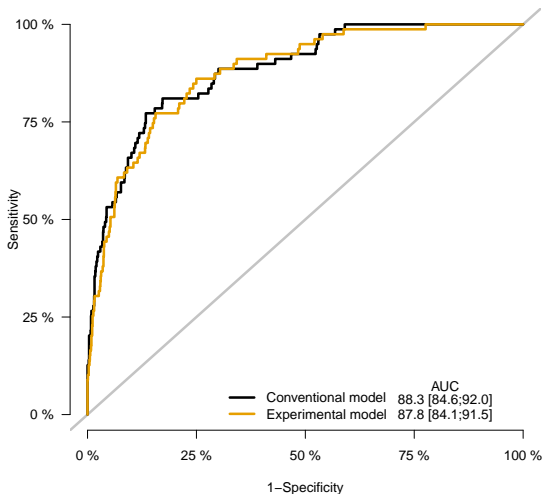
Predicted risks

```
plotRisk(x)
```



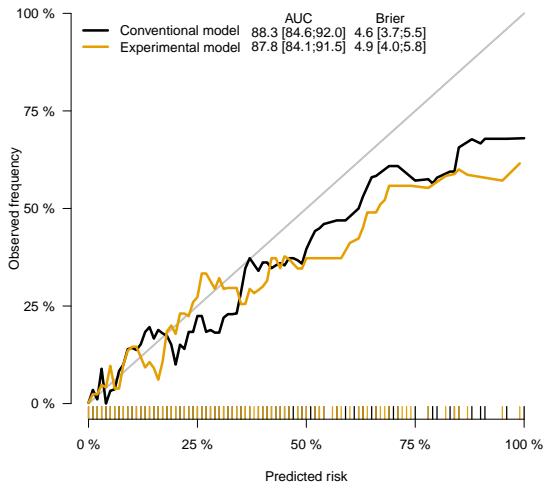
Discrimination

```
plotROC(x)
```



Calibration

```
plotCalibration(x)
```



Conditional versus expected performance

Conditional prediction performance is the performance of a risk prediction model conditional on a single-purpose dataset.

A researcher who provides a risk prediction model for clinical application is naturally interested in the conditional performance of the model.

The conditional prediction performance can be assessed by an external validation dataset or, with some limitations, using data splitting.

Conditional versus expected performance

Expected prediction performance is the performance of a modeling algorithm. It is the average performance across all the prediction models that a modeling algorithm can produce using all possible learning datasets of a fixed sample size.

A researcher who has invented a new algorithm for building prediction models is naturally interested in the average performance.

The expected prediction performance can either be assessed by computer simulation of many datasets or by using cross-validation and bootstrap methods.

Uncertainty

A probabilistic prediction has built-in uncertainty

Paradigm: A medical risk prediction model finds people in the data set who were alike the current person, i.e., with similar values of the risk predictors, and summarizes what happened to them.

Any useful model will provide more reliable predictions for people who are well represented in the data set than people at the border of the data set.

Summary and outlook

A medical risk prediction model predicts the probability with which an event occurs until a fixed prediction time horizon.

Prediction performance (metrics) can be used to decide between rival models.

The values of the prediction performance metrics do not have a direct clinical interpretation!

How useful a model is depends on what it is used for ...