Risk Prediction

Advanced Statistical Analysis

Albert Henry

UCL Institute of Health Informatics

https://ihi-risk-teaching.netlify.app/

Learning objectives

This lecture will cover:

- **Deterministic** and **probabilistic** methods for prediction
- **2x2 contigency table**
- Measures to evaluate **model performance**
- Model validation

in the context of **binary risk prediction**

Acknowledgements

Materials and data presented are based on previous slides prepared by L. Palla, D. Prieto, and E. Williamson

Scenario

Consider a dataset consisting some variables collected from a group of individuals. We want to predict which individuals develop a certain **binary event** Y in future.

Examples

Population	Available data	Outcome to predict
Surgical patients	age, sex, comorbidities, medications, severity of disease	complications after surgery
Covid-19 patients	age, sex, vaccination status, virus strain, chest X-ray	ICU admission within 10 days
Healthy individuals	age, sex, lifestyle, socioeconomic status, genetics, lipid profile	myocardial infarction in 10-year follow-up

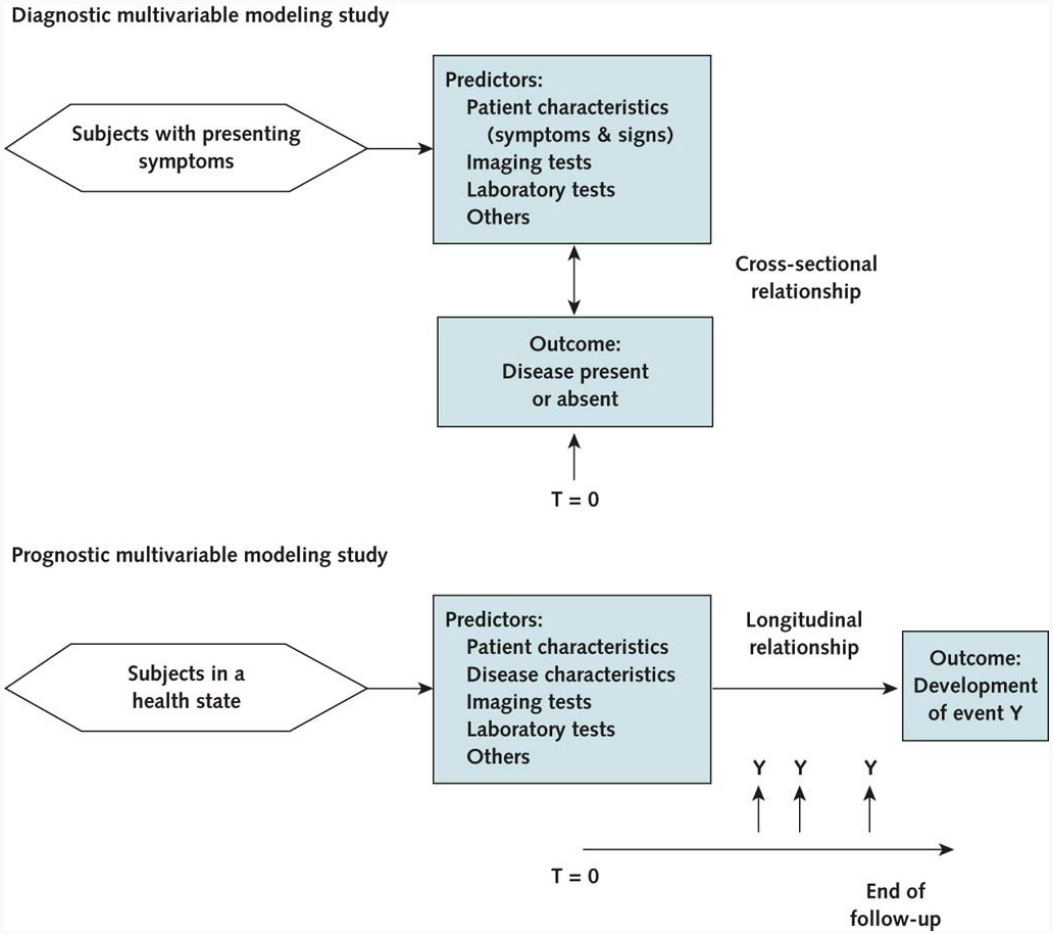
Deterministic vs probabilistic prediction

- **Deterministic:**
  - **Classify** each individual into one of the two possible outcomes.
  - Often used in **(Supervised) Machine Learning**
- **Probabilistic:**
  - Assign each individual a **probability** of developing the outcome.
  - Often used in **Biostatistics** and is known as **Risk Prediction***

Both methods use individual-level data on a set of variables (or **predictors**) to develop a prediction model

*Other names: risk prediction model, predictive model, prognostic (or prediction) index or rule, and risk score

Diagnostic vs Prognostic prediction



Moons KGM, Altman DG, Reitsma JB, et al. Ann Intern Med. 2015 Jan 6;162(1):W1–73.

Example

Who will develop myocardial infarction in the next 5 years?

id	age	sex	diabet	SBP	Classify	Predict	Observed
1	35	F	Yes	145			
2	35	M	No	130			
3	55	F	No	115			
4	55	M	Yes	170			
5	65	F	No	135			
6	65	M	Yes	140			
7	75	M	Yes	160			
8	75	F	No	130			
9	85	F	Yes	130			
10	85	M	No	160			

Deterministic classification

Suppose we have a pre-calculated deterministic classification model `C`:

```
C(age, sex, diabet, SBP) = 0 or 1
```

id	age	sex	diabet	SBP	Classify	Predict	Observed
1	35	F	Yes	145	No		
2	35	M	No	130	No		
3	55	F	No	115	No		
4	55	M	Yes	170	Yes		
5	65	F	No	135	No		
6	65	M	Yes	140	Yes		
7	75	M	Yes	160	Yes		
8	75	F	No	130	No		
9	85	F	Yes	130	Yes		
10	85	M	No	160	Yes		

Probabilistic prediction

Suppose we have a pre-calculated probabilistic prediction model `P`:

```
P(age, sex, diabet, SBP) = [0, 1]
```

id	age	sex	diabet	SBP	Classify	Predict	Observed
1	35	F	Yes	145	No	0.15	
2	35	M	No	130	No	0.05	
3	55	F	No	115	No	0.10	
4	55	M	Yes	170	Yes	0.55	
5	65	F	No	135	No	0.30	
6	65	M	Yes	140	Yes	0.52	
7	75	M	Yes	160	Yes	0.60	
8	75	F	No	130	No	0.40	
9	85	F	Yes	130	Yes	0.55	
10	85	M	No	160	Yes	0.60	

Observation

... after 5 years follow-up

id	age	sex	diabet	SBP	Classify	Predict	Observed
1	35	F	Yes	145	No	0.15	No
2	35	M	No	130	No	0.05	No
3	55	F	No	115	No	0.10	No
4	55	M	Yes	170	Yes	0.55	No
5	65	F	No	135	No	0.30	Yes
6	65	M	Yes	140	Yes	0.52	No
7	75	M	Yes	160	Yes	0.60	Yes
8	75	F	No	130	No	0.40	No
9	85	F	Yes	130	Yes	0.55	Yes
10	85	M	No	160	Yes	0.60	Yes

Model validation

- Both `C` and `P` models are not always correct in predicting the outcome
- The goal of model validation is to evaluate model performance by **comparing predictions against observed values**.
- For binary prediction, model validation usually starts with creating a 2x2 contigency table / confusion matrix consisting all four possible **pairs of predicted-observed values**

2x2 Contigency table / Confusion matrix

id	Predicted	Observed
1	No	No
2	No	No
3	No	No
4	Yes	No
5	No	Yes
6	Yes	No
7	Yes	Yes
8	No	No
9	Yes	Yes
10	Yes	Yes

Predicted	Observed	Count
Yes	Yes	3
Yes	No	2
No	Yes	1
No	No	4

Predicted	Observed	Term
+	+	True Positive
+	-	False Positive
-	+	False Negative
-	-	True Negative

2x2 Contigency table / Confusion matrix

	Observed	
Predicted	**+**	**-**
+	3	2
-	1	4

	Observed	
Predicted	**+**	**-**
+	True Positive	False Negative
-	False Positive	Ture Negative

With 2 x 2 contigency table, we can calculate several useful metrices* to evaluate model performance, including:

- sensitivity, recall, hit rate, or true positive rate (TPR)
- specificity, selectivity or true negative rate (TNR)
- precision or positive predictive value (PPV)
- negative predictive value (NPV)

*for a full list, refer to Wikipedia entry for Confusion matrix

Sensitivity

- a.k.a **True Positive Rate**, **Recall**
- Probability of **correctly predicting the event**
- Given you are someone who will **have the event**, what is the probability of a **positive classification**?

	Observed	
Predicted	**+**	**-**
+	TP	FP
-	FN	TN

$$Sensitivity = \frac{\sum True\ Positive}{\sum Observed\ Positive}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity

- a.k.a **Selectivity**, **True Negative Rate**
- Probability of **correctly predicting the *non-* event**
- Given you are someone who will ***not* have the event**, what is the probability of a **negative classification**?

	Observed	
Predicted	**+**	**-**
+	TP	FP
-	FN	TN

$$Specificity = \frac{\sum True\ Negative}{\sum Observed\ Negative}$$

$$Specificity = \frac{TN}{TN + FP}$$

Positive predictive value (PPV)

- Probability of **having the event** given a **positive classification**
- **Positively correlated** with **disease prevalence**

	Observed	
Predicted	**+**	**-**
+	TP	FP
-	FN	TN

$$PPV = \frac{\sum True\ Positive}{\sum Predicted\ Positive}$$

$$PPV = \frac{TP}{TP + FP}$$

Negative predictive value (NPV)

- Probability of ***not*** **having the event** given a **negative classification**
- **Negatively correlated** with **disease prevalence**

	Observed	
Predicted	**+**	**-**
+	TP	FP
-	FN	TN

$$NPV = \frac{\sum True\ Negative}{\sum Predicted\ Negative}$$

$$NPV = \frac{TN}{TN + FN}$$

Comparing *classifications* with observations

	Observed		
Predicted	**+**	**-**	**Total**
+	3	2	**5**
-	1	4	**5**
Total	**4**	**6**	**10**

Sensitivity **= ?**

PPV **= ?**

Specificity **= ?**

NPV **= ?**

Comparing *classifications* with observations

	Observed		
Predicted	**+**	**-**	**Total**
+	3	2	5
-	1	4	5
Total	4	6	10

Sensitivity = 3/4 = 0.75

PPV = 3/5 = 0.6

Specificity = 4/6 = 0.67

NPV = 4/5 = 0.8

Comparing *predictions* with observations

Order by risk and choose a **cut-off point** to classify as "Yes", e.g.

"Yes" if probability (P) > 0.1

id	Predict	Observed	Prob >0.1
2	0.05	No	No
3	0.10	No	No
1	0.15	No	Yes
5	0.30	Yes	Yes
8	0.40	No	Yes
6	0.52	No	Yes
4	0.55	No	Yes
9	0.55	Yes	Yes
7	0.60	Yes	Yes
10	0.60	Yes	Yes

Cut-off point: *Yes* if P > 0.1

	Observed		
Predicted	**+**	**-**	**Total**
+	4	4	8
-	0	2	2
Total	4	6	10

Sensitivity = 4/4 = 1

PPV = 4/8 = 0.5

Specificity = 2/6 = 0.33

NPV = 2/2 = 1

Higher sensitivity, lower specificity than the classification algorithm

Cut-off point: *Yes* if P > 0.4

id	Predict	Observed	Prob >0.4
2	0.05	No	No
3	0.10	No	No
1	0.15	No	No
5	0.30	Yes	No
8	0.40	No	No
6	0.52	No	Yes
4	0.55	No	Yes
9	0.55	Yes	Yes
7	0.60	Yes	Yes
10	0.60	Yes	Yes

Cut-off point: *Yes* if P > 0.4

	Observed		
Predicted	**+**	**-**	**Total**
+	3	2	5
-	1	4	5
Total	4	6	10

Sensitivity = 3/4 = 0.75

PPV = 3/5 = 0.6

Specificity = 4/6 = 0.67

NPV = 4/5 = 0.8

Same contigency table as the classification algorithm

Cut-off point: *Yes* if P > 0.55

id	Predict	Observed	Prob >0.55
2	0.05	No	No
3	0.10	No	No
1	0.15	No	No
5	0.30	Yes	No
8	0.40	No	No
6	0.52	No	No
4	0.55	No	No
9	0.55	Yes	No
7	0.60	Yes	Yes
10	0.60	Yes	Yes

Cut-off point: *Yes* if P > 0.55

	Observed		
Predicted	**+**	**-**	**Total**
+	2	0	2
-	2	6	8
Total	4	6	10

Sensitivity = 2/4 = 0.5

PPV = 2/2 = 1

Specificity = 6/6 = 1

NPV = 6/8 = 0.75

Lower sensitivity, higher specificity than the classification algorithm

All cut-off points

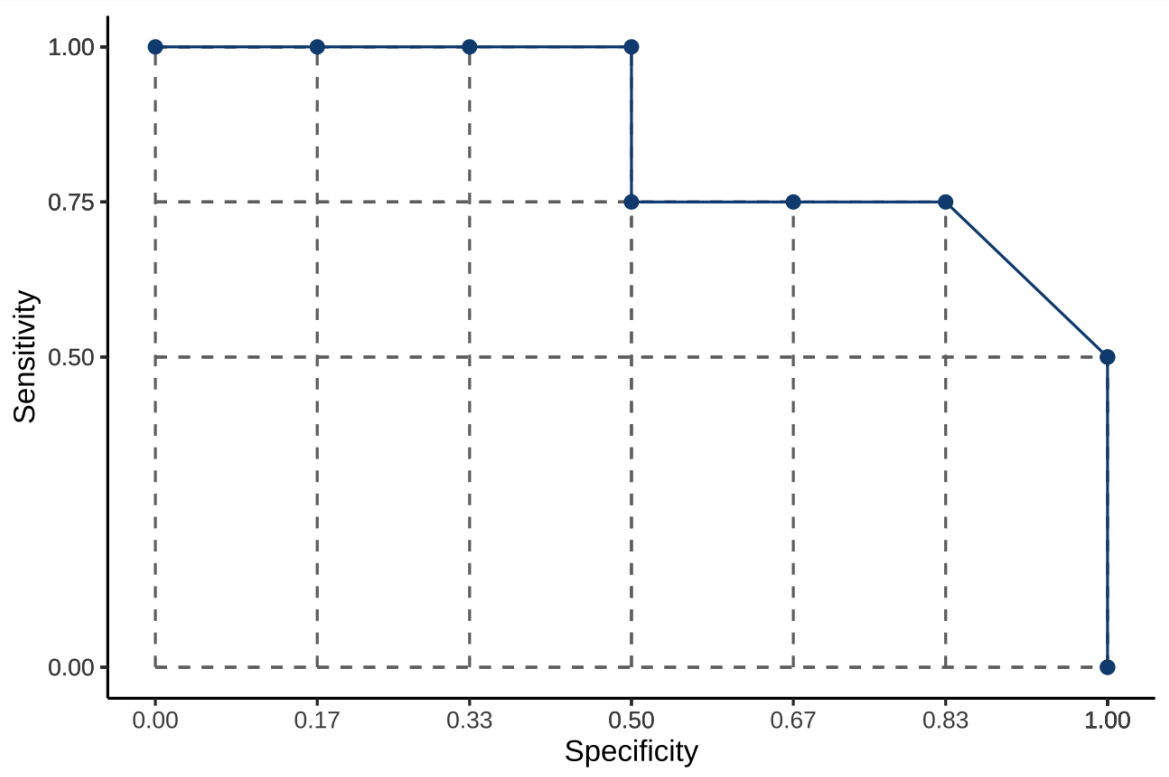
If we repeat this process for each probability value, we can obtain a list of sensitivity and specificity values

id	Predict	Observed	Cut-off	Sensitivity	Specificity
2	0.05	No	P >0.05	1.00	0.17
3	0.10	No	P >0.1	1.00	0.33
1	0.15	No	P >0.15	1.00	0.50
5	0.30	Yes	P >0.3	0.75	0.50
8	0.40	No	P >0.4	0.75	0.67
6	0.52	No	P >0.52	0.75	0.83
4	0.55	No	P >0.55	0.50	1.00
9	0.55	Yes	P >0.55	0.50	1.00
7	0.60	Yes	P >0.6	0.00	1.00
10	0.60	Yes	P >0.6	0.00	1.00

Receiver Operating Characterictic (ROC) Curve

A curve linking all the sensitivity against the specificity values

Sens	Spec
1.00	0.17
1.00	0.33
1.00	0.50
0.75	0.50
0.75	0.67
0.75	0.83
0.50	1.00
0.50	1.00
0.00	1.00
0.00	1.00



Area Under the [ROC] Curve (AUC / AUROC)

What is the AUC?

Area Under the [ROC] Curve (AUC / AUROC)

What is the AUC?

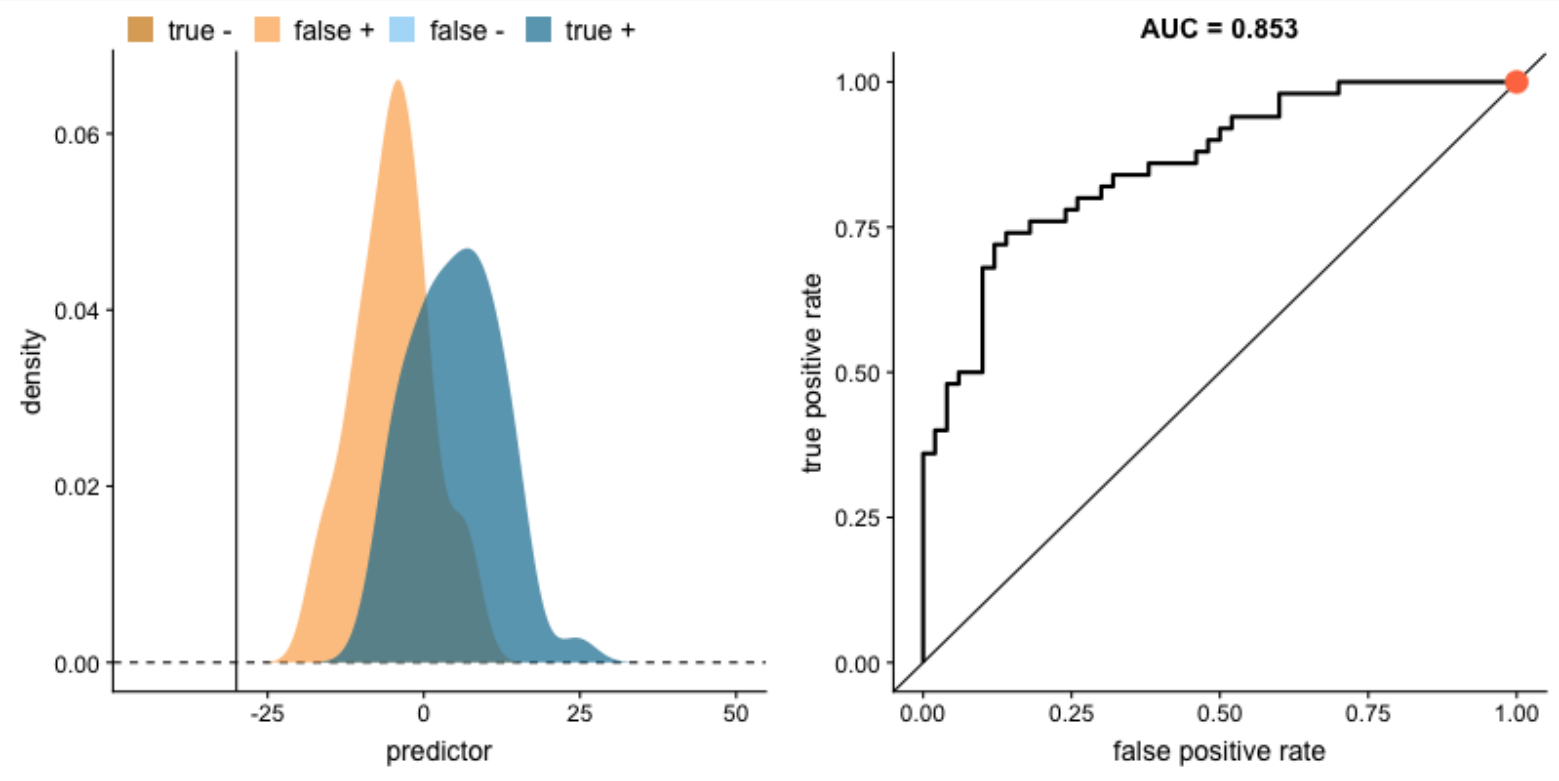
Sample calculation with yardstick package in R

```
# df = the dataset shown in previous slides
roc <- yardstick::roc_auc(df, truth = Observed, Predict)
roc
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.854
```

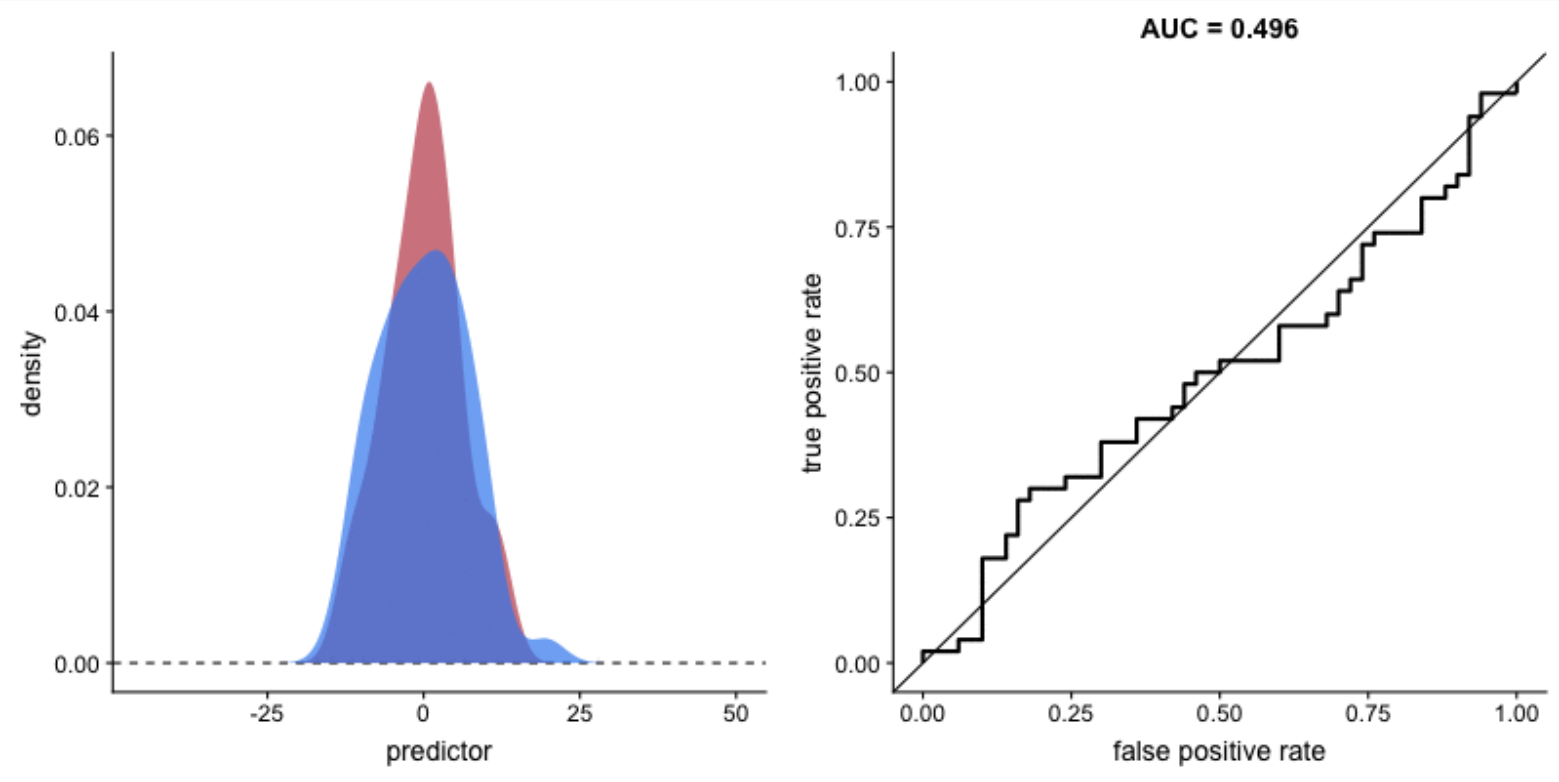
AUC = 0.8541667

AUC and the distributions of predictors in the two outcome groups at different cut-off values



by Dariya Sydykova (follow link for more info)

AUC as a measure of model discrimination



by Dariya Sydykova (follow link for more info)

What does AUC tell us?

- AUC estimates the **probability that a randomly chosen observed "yes" was assigned a higher probability than a randomly observed "no"** by the model
- Real world models will have **AUC from 0.5 to 1**. A value closer to 1 indicates better performance in separating "yes" and "no".
- For binary classification, AUC is equal to **concordance (C) statistic**

How do we come up with predictions?

- We propose a statistical model for the **probability of the event happening** $P\left(Y_i = 1\right)$ depending on the other variables
- For example a **logistic model**:

$$\log\left(\frac{P\left(Y_i = 1\right)}{1 - P\left(Y_i = 1\right)}\right) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \cdots \tag{1}$$

- We need a **training set** where we can observe all the variables $Y_i, X_i, Z_i, \ldots$ to estimate the **coefficients** $\beta_0, \beta_1, \beta_2, \ldots$
- Once we have the coefficients that best fit the data we can **calculate the predicted risk for each individual** i

$$\widehat{P}\left(Y_i = 1\right) = \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \cdots}}{1 + e^{\widehat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i + \cdots}} \tag{2}$$

Dataset for model development & model validation

Internal validation

- The validity of claims for the underlying **population that the data originated from (reproducibility)**
- **Split sample validation**: split dataset randomly into **training** (for model development) and **test** set (for validation)
- Other methods: **cross validation** and **bootstrap** resampling

External validation

- Generalizability of claims to **'plausibly related' populations** not included in the initial study population **(transportability)**
- e.g. **temporal** or **geographical** validation

Steyerberg EW, Vergouwe Y (2014)

A larger example with 2000 individuals

We will use the variables `Age, Sex, SBP, and BMI` to predict if the person will be dead (`Death = 1`) or alive (`Death= 0`) in 5 years time

id	Age	SBP	BMI	Sex	Death
1	47	116.8	25.6	Female	Alive
2	71	113.8	21.1	Male	Dead
3	41	130.7	25.6	Male	Alive
4	71	118.2	25.9	Male	Alive
5	54	120.3	20.7	Male	Alive
6	67	126.4	22.8	Male	Alive
7	71	129.3	32.9	Female	Dead
8	73	117.4	27.5	Male	Alive
...	..	.....	.....	.....	.....
2000	46	111.1	24	Male	Alive

Model *M1*: Logistic regression

Stata command:

```
logistic dead age sex sbp bmi
```

```
Logistic regression                                     Number of obs = 2,000
LR chi2(4) = 377.64
Prob > chi2 = 0.0000
Log likelihood = -945.62964                             Pseudo R2 = 0.1664
------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.09391   .0062246    15.77   0.000     1.081778    1.106178
         sex |   1.301071   .1484571     2.31   0.021     1.040341    1.627145
         sbp |   1.050522   .0061981     8.35   0.000     1.038444    1.062741
         bmi |   1.025531   .0195045     1.33   0.185     .9880064    1.06448
       _cons |   1.39e-06   1.38e-06   -13.52   0.000     1.96e-07    9.81e-06
------------------------------------------------------------------------------
```

Note that BMI is not statistically significant (`P = 0.185`) but let's stay with this model for now

Make predictions from Model *M1*

Create variable: **logit of the probability of death - equation (1)**

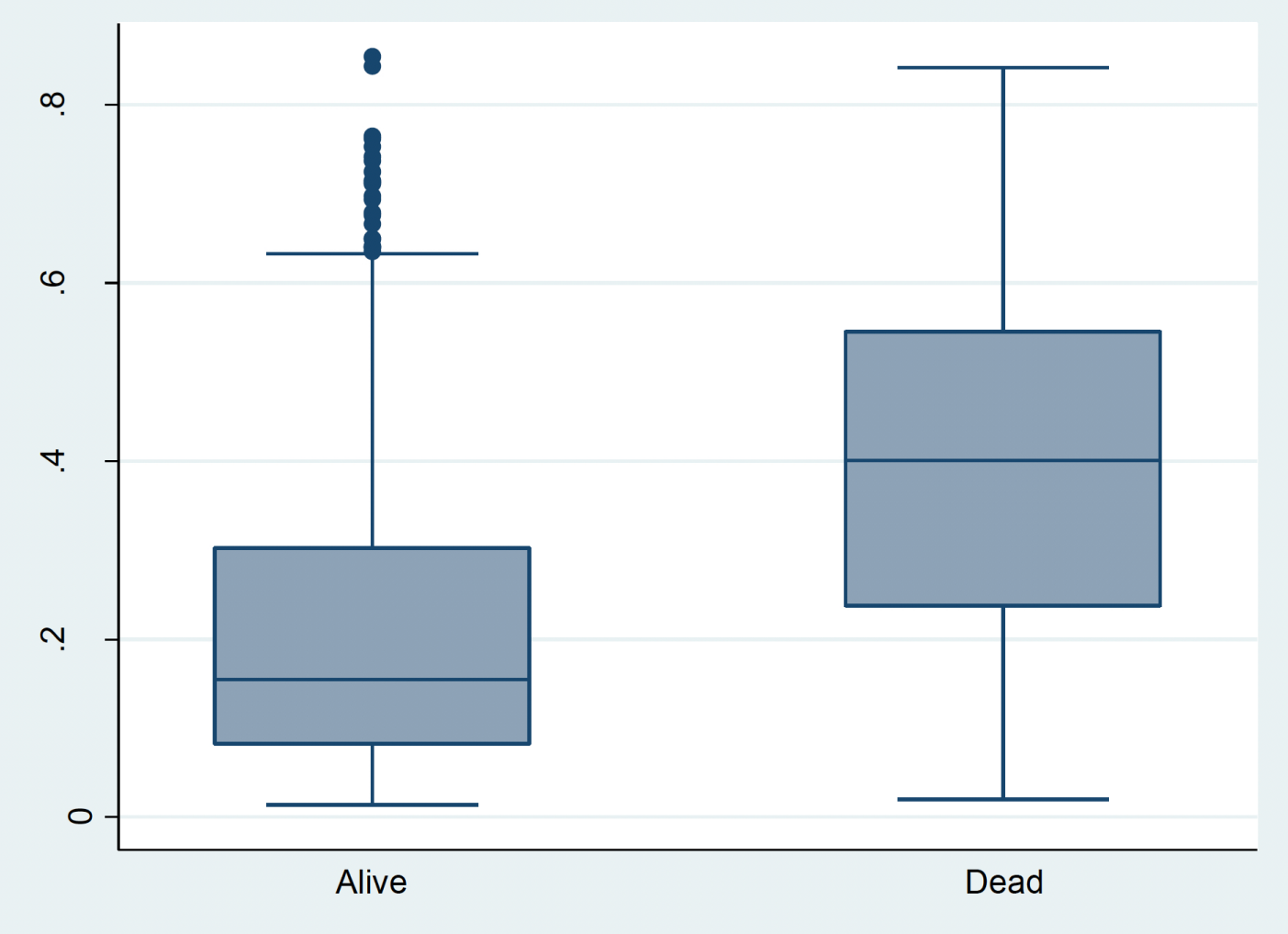
```
predict m1lp, xb
```

Create variable: **predicted probability of death - equation (2)**

```
predict m1pr
```

id	age	sbp	bmi	sex	dead	m1lp	m1pr
10	46	111.1	24	Male	Alive	-3.014752	.0467639
11	79	117.5	28.7	Male	Dead	.381197	.594162
12	72	113.2	24.2	Male	Alive	-.572492	.3606619
13	54	123.4	26.1	Female	Alive	-1.900696	.1300298
14	45	105.7	24.7	Female	Alive	-3.616203	.0261807

Predicted probability of death in `dead` and `alive` group



Model calibration

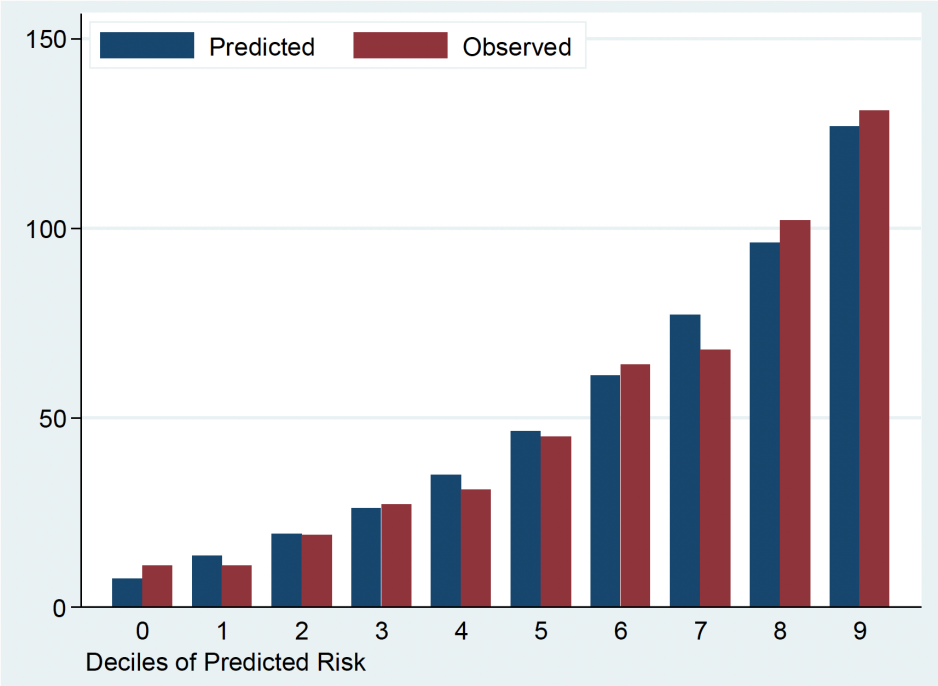
Evaluate **goodness of fit** with **Hosmer-Lemeshow test**

```
estat gof, group(10) table
```

```
Logistic model for dead, goodness-of-fit test
(Table collapsed on quantiles of estimated probabilities)
+--------------------------------------------------------+
| Group |   Prob   |  Obs_1 | Exp_1  | Obs_0 | Exp_0 | Total |
|-------+----------+--------+--------+-------+-------+-------|
|    1  | 0.0520 |    11   |   7.4  |  189  | 192.6 |  200  |
|    2  | 0.0831 |    11   |  13.6  |  189  | 186.4 |  200  |
|    3  | 0.1133 |    19   |  19.3  |  181  | 180.7 |  200  |
|    4  | 0.1504 |    27   |  26.1  |  173  | 173.9 |  200  |
|    5  | 0.2011 |    31   |  35.0  |  169  | 165.0 |  200  |
|    6  | 0.2670 |    45   |  46.4  |  155  | 153.6 |  200  |
|    7  | 0.3460 |    64   |  61.1  |  136  | 138.9 |  200  |
|    8  | 0.4277 |    68   |  77.0  |  132  | 123.0 |  200  |
|    9  | 0.5384 |   102   |  96.2  |   98  | 103.8 |  200  |
|   10  | 0.8544 |   131   | 126.7  |   69  |  73.3 |  200  |
+--------------------------------------------------------+
Hosmer-Lemeshow chi2(8) = 5.98 Prob > chi2 = 0.6499
```

Model calibration

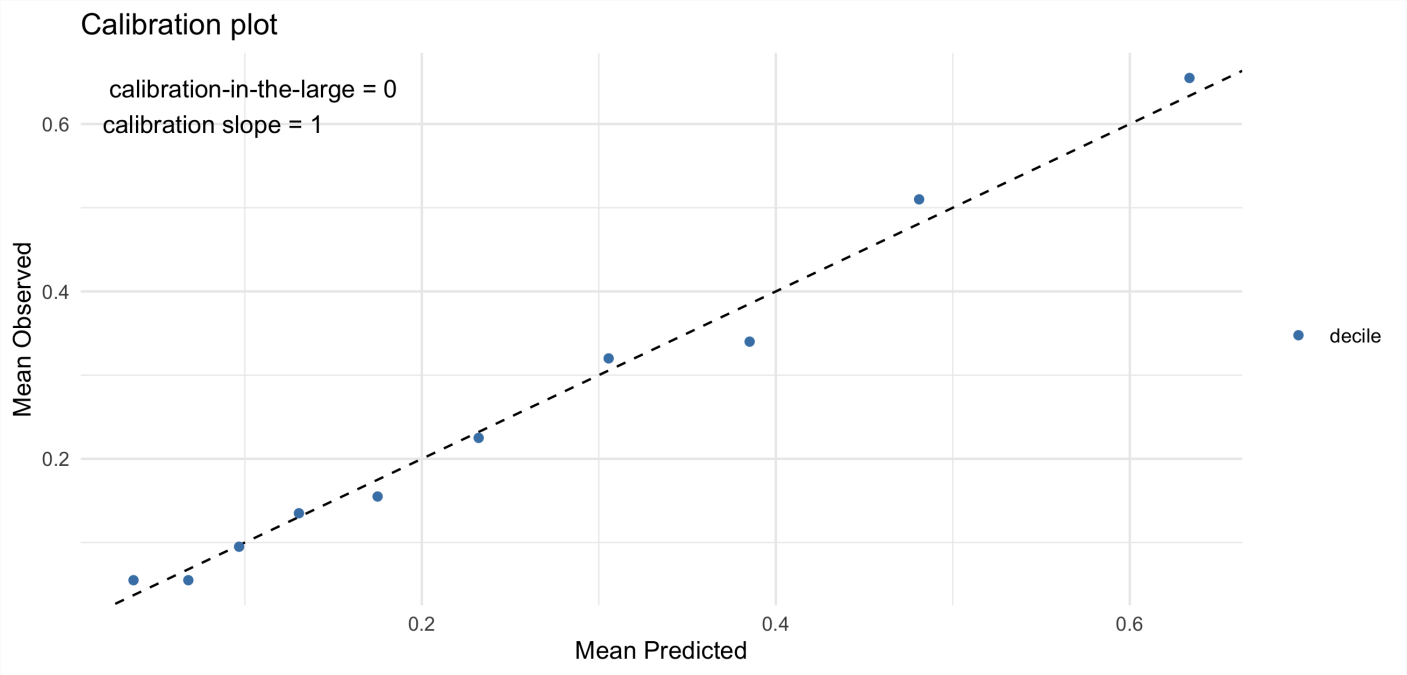
Goodness of fit: Observed and expected events by deciles of risk



Limitation: Hosmer-Lemeshow test can not tell the direction of miscalibration and relies on arbitrary grouping

Model calibration with calibration plot

- **Calibration in the large**: compares average predicted risk with observed risk (**0**: ideal, **<0**: underestimation, **>0**: overestimation)
- **Calibration slope**: evaluate spread of estimated risk (**1**: ideal, **<1**: too extreme, **>1**: too moderate)



Further discussion: Steyerberg EW, Vergouwe Y (2014) and Calster BV *et al.* (2019)

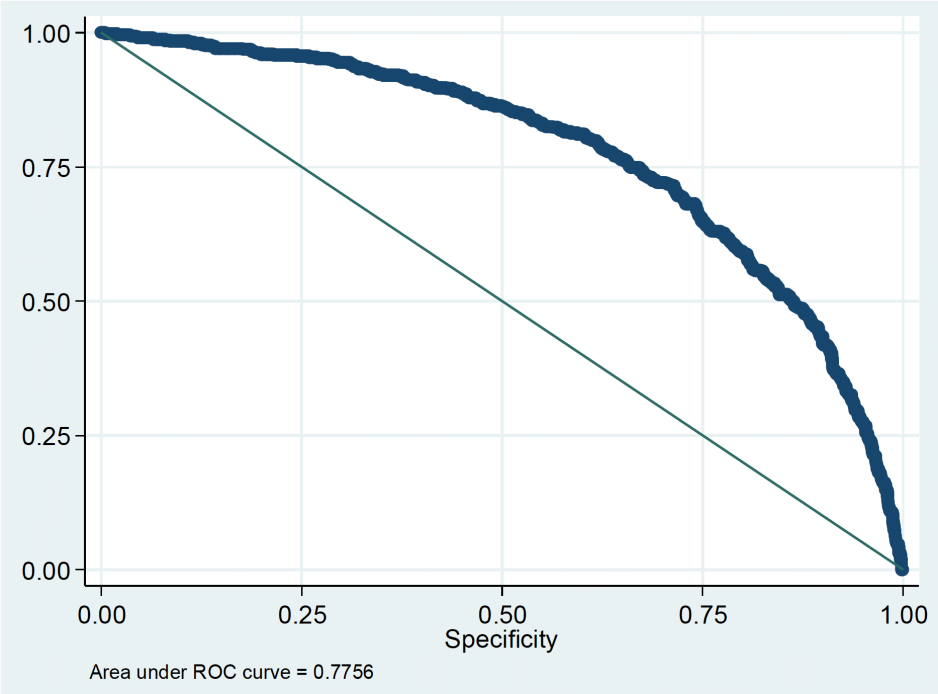
Contingency table: cut-off $P(Y = 1) \geq 0.3$

```
estat classification, cutoff(0.3)
```

```
------------------------ True ------------
   Classified |         D          ~D      |     Total
--------------+-------------------------------+-----------
            + |       333          375      |       708
            - |       176         1116      |      1292
--------------+-------------------------------+-----------
        Total |       509         1491      |      2000

Classified + if predicted Pr(D) >= .3
-------------------------------------------------------
Sensitivity                     Pr( +| D)   65.42%
Specificity                     Pr( -|~D)   74.85%
Positive predictive value       Pr( D| +)   47.03%
Negative predictive value       Pr(~D| -)   86.38%
False + rate for classified +   Pr(~D| +)   52.97%
False - rate for classified -   Pr( D| -)   13.62%
Correctly classified                        72.45%
-------------------------------------------------------
```

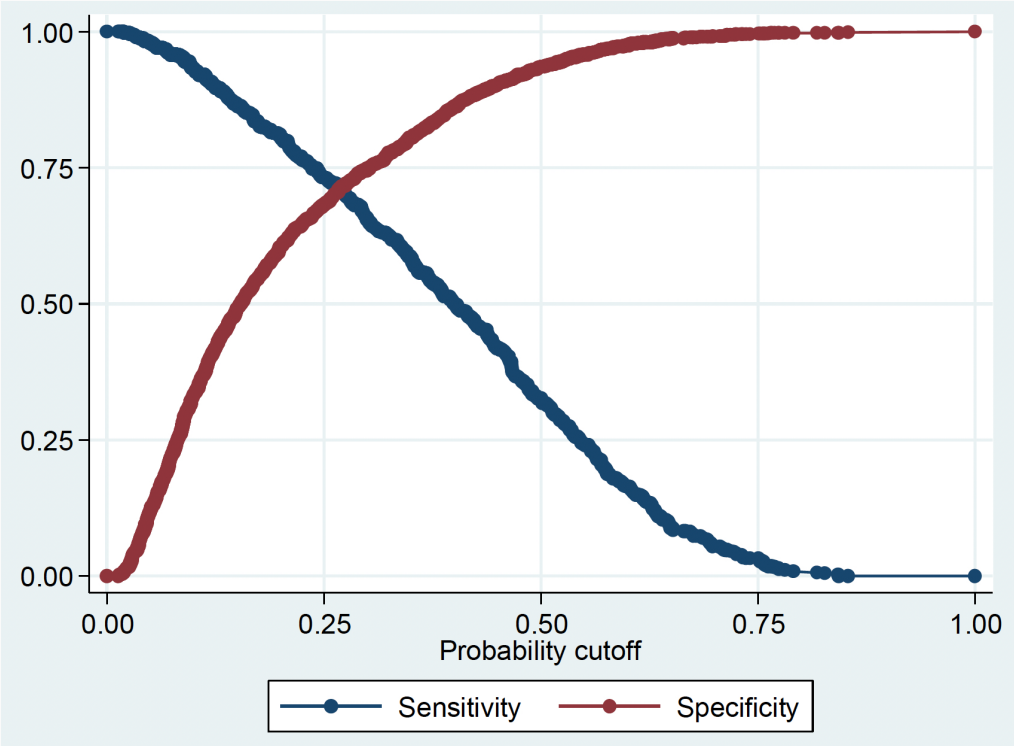
ROC curve from model *M1*



AUC = 0.78

There is a 78% probability that a person who died is assigned a higher predicted risk by the model than a person was alive by the end of the follow up

Sensitivity and Specificity for model *M1* by cut-off value



Plotting **sensitivity and specificity against cut-off value** can help to select the most appropriate cut-off

In practice, this trade-off often needs to be decided on a **case-by-case** basis

Key measures to evaluate model performance

Calibration

- The agreement between the predicted & observed outcomes
- For a group of patients with 10% predicted risk, do 10% experience the event?
- e.g. Goodness of fit test

Discrimination

- The ability of the model to distinguish between "event" and "non-event"
- e.g. AUC / C statistic

Clinical usefulness

- Does the model provide accurate predictions at the patient level that can be used to guide clinical decision making?
- e.g. Decision curve analysis

Summary

- Prediction modelling can be broadly categorised into **deterministic** and **probabilistic** methods
- **2 x 2 contigency table / confusion matrix** is useful as a first step to evaluate model performance
- **AUC** is a useful measure of the model **discrimination**
- Comparing observed and predicted risks is useful for model **calibration**
- Assessing **clinical usefulness** requires other approaches and often requires insights from ***beyond the data***

References and further reading

- Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014 Aug 1;35(29):1925–1931.
- Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015;162:W1–W73
- Comparison with machine learning:
  Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Stat Sci. Institute of Mathematical Statistics; 2001 Aug;16(3):199–231.
- ROC curve animation by Dariya Sydykova
- A Twitter thread on AUC by Cecile Janssens