Linear and Generalized Linear Models

Week 7, Lecture 2

Logistic regression, part 2

Saskia le Cessie

Leiden University Medical Centre

Yesterday:

- The basics of the logistic regression model

Today

- More on model building
- Why the logistic model is so often used
- Prosepctive and retrospective study designs

Problems with fitting logistic regression

- Sometimes estimates of regression coefficients are very big or small, or a warning like: `glm.fit: fitted probabilities numerically 0 or 1 occurred` is given

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -20.57    1023.66  -0.020    0.984
x              21.66    1023.66   0.021    0.983
```

- This indicates that the optimalisation algorithm failed to converge
- Reason: groups are linearly separable: in that case a perfect fit is possible (odds ratios are 0 or ∞).
- This does not imply that the model does not fit.
- On the contrary, the model predicts some observations perfectly.

$P(Y=1|X=0)=0.75$

$P(Y=1|X=1)=0$ X=1完美分割

可分的

$\exp(\beta)$ = odds ratio $= \frac{0.75/1-0.75}{0/1-0}$

$= +\infty$

$\beta = +\infty$

Confidence intervals for π_i

- Confidence interval for $\pi_i = \frac{\exp(\beta_0+\beta_1 x_{i1}+..+\beta_k x_{ik})}{1+\exp(\beta_0+\beta_1 x_{i1}+..+\beta_k x_{ik})} = \frac{\exp(x_i'\boldsymbol{\beta})}{1+\exp(x_i'\boldsymbol{\beta})}$
- First confidence interval for logit$_i$ = $x_i'\boldsymbol{\beta}$, then transform back to probability scale

- var($x_i'\widehat{\boldsymbol{\beta}}$) = x_i'var($\widehat{\boldsymbol{\beta}}$) x_i

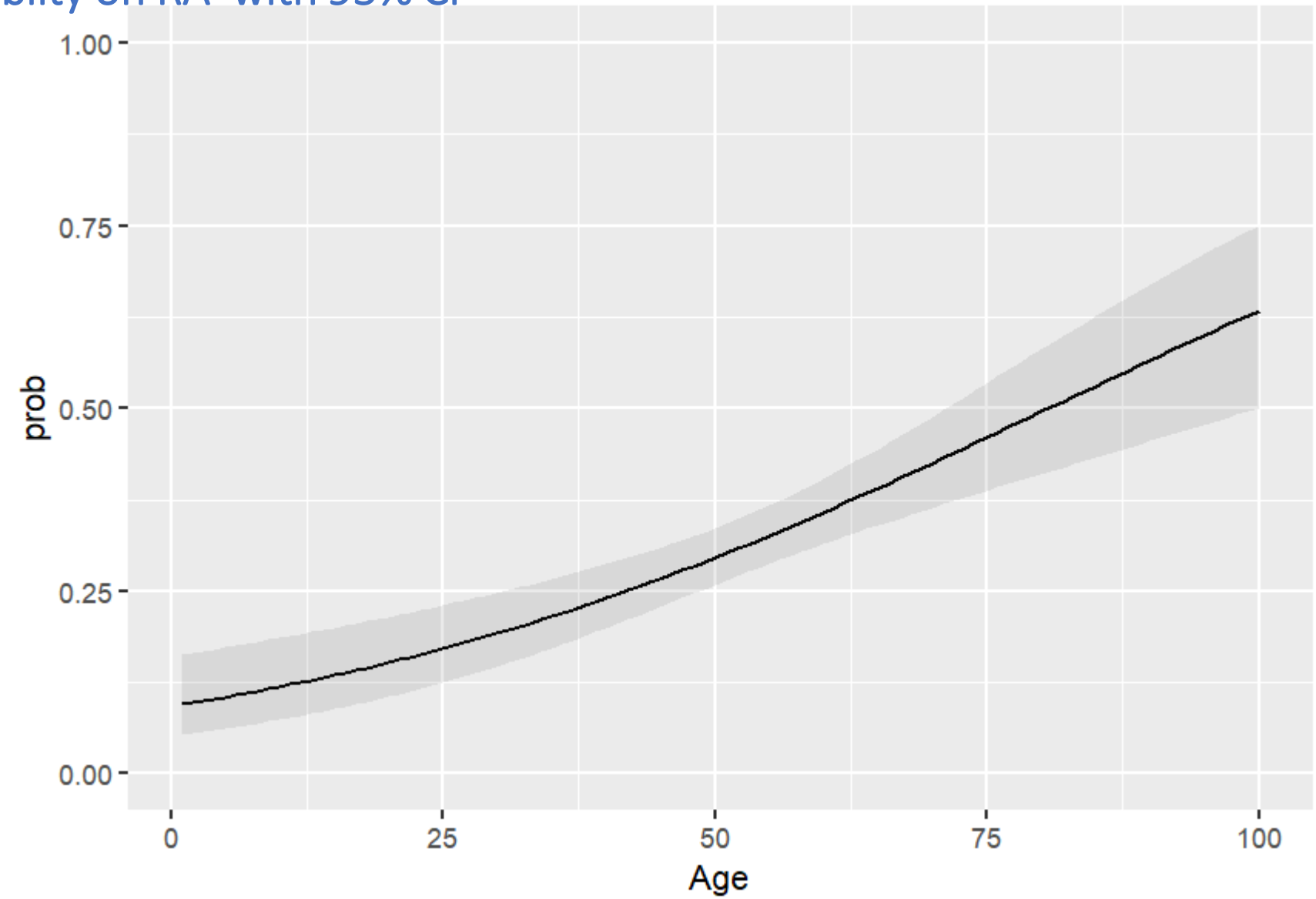
  $\approx x_i'(X'V\ X)^{-1}x_i$ → estimate of se ($x_i'\widehat{\boldsymbol{\beta}}$)

- CI for $x_i'\boldsymbol{\beta}$: ($x_i'\widehat{\boldsymbol{\beta}}$ - $z_{\alpha/2}$ se ($x_i'\widehat{\boldsymbol{\beta}}$) , $x_i'\widehat{\boldsymbol{\beta}}$ + $z_{\alpha/2}$ se ($x_i'\widehat{\boldsymbol{\beta}}$))

  (LWB, UPB)

- Use this to obtain CI for π_i: $(\frac{\exp(LWB)}{1+\exp(LWB)}, \frac{\exp(UPB)}{1+\exp(UPB))})$

Probabilty on RA with 95% CI



Model building issues

Model building in logistic regression

- Many topics discussed for linear regression carry over to logistic regression
  - Categorical variables with dummy variables
  - Interaction terms
  - Stepwise selection

Deviance

- Deviance D = -2 log likelihood fitted model - -2 log likelihood saturated model

饱和

- Saturated model is the perfect fitted model, a model with a parameter for every observation so that the data are fitted exactly

$$\sum y_i \log \pi_i + (1-y_i) \log(1-\pi_i)$$

- For logistic regression the saturated model has *n* parameters (unless all X-variables are categorical with few levels). In this case, the likelihood of the saturated model equals 1, the log likelihood = 0

- In that case D = -2 log likelihood of the fitted model

Comparing two nested models

- Compare a restricted model (RM) to full model (FM)
- Use likelihood ratio test; calculate difference in deviance.
- $D_{RM} - D_{FM} = -2\log\big(L(FM)\big) - -2\log\big(L(RM)\big) = 2\log\big(L(RM)\big) - 2\log\big(L(FM)\big)$
- Test statistic has approximately χ^2_{l-s} distribution with l number of fitted parameters Full Model and s number of parameters restricted model.
- Similar to considering differences of residual sums of squares in linear regression.

Example:

- RA example
- Model with age, sex smoking and rheumafactor: deviance= 599.85
- Model with only age and rheumafactor: deviance= 616.22

- Difference: 616.22-599.85 =16.36
- Difference in number of fitted parameters: 5-3 = 2

- Compare to χ_2^2 → p-value of 0.0003

Does the model fit?

- If all X variables are categorical with few levels, deviance can be used as goodness of fit test.
- Otherwise: assess the fit by comparing the model to a more complex model
- For example: compare model with age linear to a model with age and age*age (polynomials), or a model with age modeled with splines.
- The Hosmer –Lemeshow test is an overall goodness of fit test but it is not very powerful

Akaike Information Criterium

Measurement of fit of the model

Deviance = -2 log L

- AIC = p – 2 log L , with p number of parameters estimated (k+1)

  = Deviance + p

- Smaller values are better
- AIC Penalizes models with a large number of parameters

Residuals

- $e_i = y_i - \hat{\pi}_i$ (unstandardised)
- $r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$ (Pearson residuals)
- $dev_i = \text{sign}(e_i)\sqrt{-2(y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i))}$

dev_i is standard residual in R

Usually not very informative, because y can take only two values . You may make smoothed plots

Leverage and Influencial points

- In linear regression: the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$,
- In logistic regression: hat matrix $\mathbf{H} = \boldsymbol{V}^{\mathbf{1/2}}\mathbf{X}(\mathbf{X}'\mathbf{V}\,\mathbf{X})^{-1}\mathbf{X}'\,\boldsymbol{V}^{\mathbf{1/2}}$, with V the diagonal matrix with diagonal elements π_i(1- π_i)
- h_{ii} , the diagonal element of $\mathbf{H}$ indicates high 'leverage'. Interpretation like in linear regression (points with large h_{ii} are extreme in the covariate space), if estimated probabilities are not to close to 0 or 1 (between 0.1 and 0.9).
- As in linear regression, influential points can be determined by calculating $\Delta\hat{\beta}_i$(change in parameter estimates if we remove observation i), or Cook's distance

Link functions

- The logistic function is a function from (0,1) to $(-\infty, \infty)$.
- It links π_i to the linear predictor $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$
- Other examples of functions which link π_i to the linear predictor η_i.
  - Probit: $\eta_i = \Phi^{-1}(\pi_i)$, which is the inverse of normal cdf.
  - Other cumulative density functions
- Probit models are sometimes used an alternative to logistic models. Predicted probabilities are very similar. Coefficients are different but there exist conversion rules.

Why is the logistic model so popular?

- Nice mathematical properties
- It yields estimates of odds and odds ratios
- Odds and odds ratios approximate risks and relative risks if the outcome Y=1 is rare.
- It can be used to analyze data from studies with retrospective sampling (case-control data).

Prospective and retrospective sample designs

追踪研究（疾病出现前分组）

Cohort study (prospective sampling design)

群组研究

1. Start by collecting data on predictor variables $(x_1, \ldots, x_k)$ in a well defined population

2. Follow subjects over a certain period and record their outcome status (Y)

Cohort study

$(x_1, \ldots, x_k)$  Y

- The predictors are fixed, the outcome is random
- Sampling from Y | x

Types of cohort studies

- Concurrent/Prospective study: first sample data on X and subsequently follow individuals over time while recording outcome(s) of interest
  - These studies often take years to conduct
  - You determine what data to collect and when to collect data

- Non concurrent/ historical/retrospective cohort study. Population is assembled from available data records
  - Can be conducted in quite a short time
  - But not all information of interest may have been recorded

In cohort studies

It is possible to estimate risks: $\pi = P(Y = 1|x)$

It is possible to estimate :

- risk differences, risk ratio and odds ratio

Outcome based sampling (case-control studies, retrospective sampling)

Sampling is carried out separately for those with the outcome (cases) and without the outcome (controls)

1. Identify two subgroups based on presence or absence of Y
2. Take a random sample from separately for those with Y=1 and with Y=0.
3. Measure X-variables in both samples

Outcome based sampling/ case-control study

$(x_1, \ldots, x_k)$  Y

- The outcome is fixed
- The x-variables are random
- Sampling from X|y

Examples of outcome dependent sampling

- What are characteristics that have influenced people's decision to vote for the PVV party?
- Case-control design: select 1000 PVV voters and 1000 non PVV voters, and collect data on predictors (age, income, gender, ……)
- Is smoking affecting the risk on lung cancer?
- Case-control design: select 100 patients with lunc cancer and 100 patients without and collect data on smoking

Outcome based sampling

No longer possible to estimate risks: $\pi = P(Y = 1|x)$

No longer possible to estimate risk differences and risk ratio

But: still possible to estimate odds ratio's

	Y	
X	RA	No RA
Rheumafactor	84	56
No rheumafactor	93	336

Look at Y|X

P(Y=1|X=1) =

P(Y=1|X=0) =

- Odds ratio $\frac{P(Y=1|X=1)/(1-P(Y=1|X=1)}{P(Y=1|X=0)/(1-P(Y=1|X=0)}$

Look at X|Y

P(X=1|Y=1) =

P(X=1|Y=0) =

- Odds ratio $\frac{P(X=1|Y=1)/(1-P(X=1|Y=1)}{P(X=1|Y=0)/(1-P(X=1|Y=0)}$

Property of the odds ratio:

- Interchanging Y and X has no influence on the value of the odds ratio.
- The odds ratio is invariant under changes in study design.

Nice property of logistic regression

- Estimates odds ratios which are invariant under the sampling design
- Applying the logistic regression model (which models Y|X) to data obtained from outcome based sampling (with sampling from X|Y) still yields consistent estimates of odds ratios and the correct standard errors
- General proof is not trivial ??
- Note: other binary regression models (like the probit model) do not have this property

Back to the Rheuma example

The first steps of building a prediction model

1. Select possible predictors based on clinical knowledge
   - Not too many if you want to perform standard logistic regression
2. Find a way to deal with missing values
3. Build model
   1. Check the shape of the continuous covariates, (testing quadratic terms or making (smoothed) residual plots)
   2. Add some (sensible) interaction terms
   3. Outliers and overall goodness-of-fit
4. Variable selection ?

Variable selection performed in two steps

Step 1 Compare the candidate predictors one by one between patients with and without RA

- With unpaired t-test for numerical variables.
- With chi-square test for categorical variables

Variables which had some relation with the outcome ($p<0.10$) were selected for step 2

Selecting variables

- Performed different backward selections
- We compared the fit of different models with Akaike information criterion
- We looked if model was plausible (correct sign of coefficients)
- We categorized continuous variable is that was possible. Assumption: coefficients for categorical variables are easier to interpret

Table 2. Independent predictive variables for development of RA based on results of multivariate regression analysis*

Variable	B	OR	95% CI	*P*
Sex	0.8	2.1	1.3–3.6	0.003
Age	0.02	1.02	1.01–1.04	0.011
Localization in small joints hand/feet	0.6	1.8	1.1–3.1	0.024
Symmetric localization	0.5	1.6	1.0–2.8	0.075
Localization in upper extremities	0.8	2.1	1.1–4.4	0.04
Localization in both upper and lower extremities	1.3	3.5	1.7–7.5	0.001
Morning stiffness score on 100-mm VAS				
0–25	–	–	–	–
26–50	0.9	2.4	1.2–4.5	0.009
51–90	1.0	2.7	1.3–5.6	0.006
>90	2.2	9.3	3.0–28.7	<0.001
Number of tender joints				
0–3	–	–	–	–
4–10	0.6	1.8	0.9–3.3	0.082
>10	1.2	3.3	1.5–7.0	0.003
Number of swollen joints				
0–3	–	–	–	–
4–10	0.4	1.5	0.8–2.7	0.18
>10	1.0	2.8	1.1–7.6	0.038
CRP level, mg/liter				
0–4	–	–	–	–
5–50	0.6	1.6	0.9–3.0	0.13
>50	1.6	5.0	2.0–12.1	0.00
RF positivity	0.8	2.3	1.2–4.2	0.009
Anti-CCP positivity	2.1	8.1	4.2–15.8	<0.001

* B values are regression coefficients. RA = rheumatoid arthritis; OR = odds ratio; 95% CI = 95% confidence interval; VAS = CRP = C-reactive protein; RF = rheumatoid factor; anti-CCP = anti–cyclic citrullinated peptide.
† For the simplified prediction rule derived from the regression coefficient.

We simplified the score

- Regression coefficients were rounded to the nearest 0 or 0.5.

Variable	B	Points†
Sex	0.8	1
Age	0.02	0.02/year
Localization in small joints hand/feet	0.6	0.5
Symmetric localization	0.5	0.5
Localization in upper extremities	0.8	1
Localization in both upper and lower extremities	1.3	1.5
Morning stiffness score on 100-mm VAS		
0–25	–	–
26–50	0.9	1
51–90	1.0	1
>90	2.2	2
Number of tender joints		
0–3	–	–
4–10	0.6	0.5
>10	1.2	1
Number of swollen joints		
0–3	–	–
4–10	0.4	0.5
>10	1.0	1
CRP level, mg/liter		
0–4	–	–
5–50	0.6	0.5
>50	1.6	1.5
RF positivity	0.8	1
Anti-CCP positivity	2.1	2

year/0.02

1. What is the age in years? Multiply by 0.02.			_____
2. What is the sex?			
	In case female:	1 point	_____
3. What is the distribution of involved joints?			
	In case small joints hands/feet:	0.5 point	_____
	In case symmetric:	0.5 point	_____
	In case upper extremities:	1 point	_____
	In case upper and lower extremities:	1.5 points	_____
4. What is the score for morning stiffness on a 100-mm VAS?			
	In case 26-90 mm:	1 point	_____
	In case >90 mm:	2 points	_____
5. What is the number of tender joints?			
	In case 4-10:	0.5 point	_____
	In case 11 or higher:	1 point	_____
6. What is the number of swollen joints?			
	In case 4-10:	0.5 point	_____
	In case 11 or more:	1 point	_____
7. What is the C-reactive protein level?			
	In case 5-50 mg/liter:	0.5 point	_____
	In case 51 mg/liter or higher:	1.5 points	_____
8. Is the patient rheumatoid factor positive?			
	If yes:	1 point	_____
9. Are the anti-CCP antibodies positive?			
	If yes:	2 points	_____
		Total score	_____

How good predicts this model?

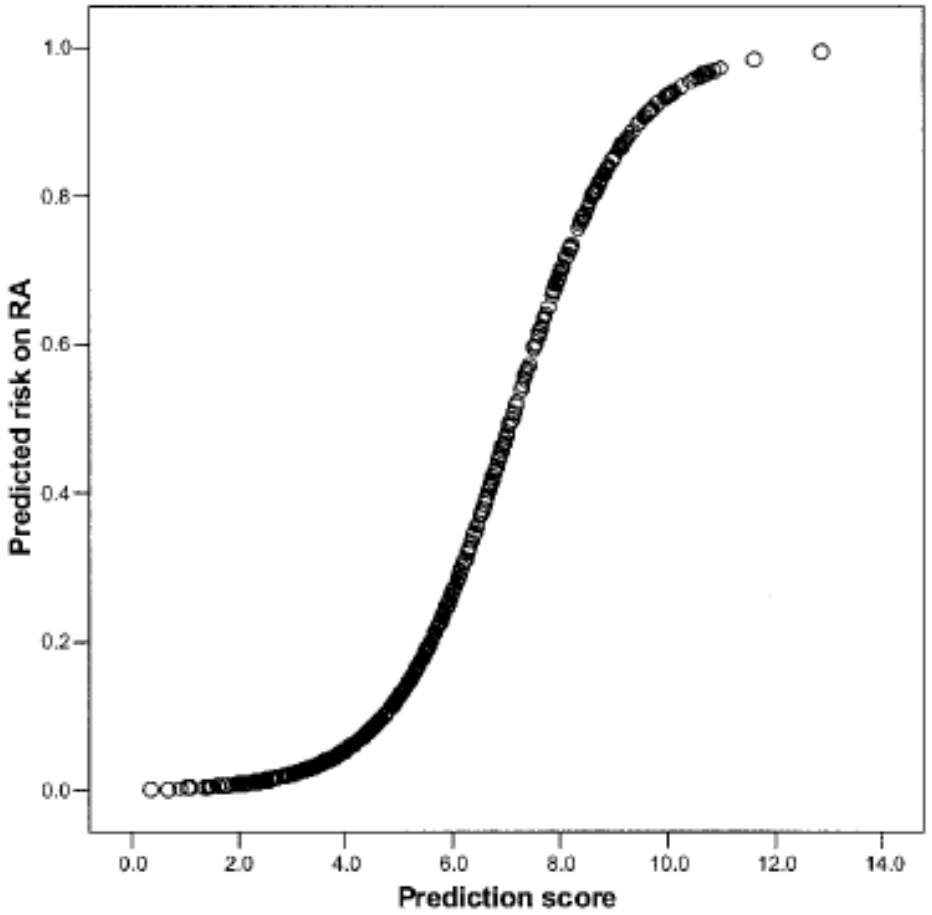


Figure 2. Predicted risk of rheumatoid arthritis (RA) as a function of the prediction score.

Table 3. Prediction scores and progression or nonprogression to RA*

Prediction score	No progression to RA (n = 387)	Progression to RA (n = 175)
0	1 (100)	0 (0)
1	8 (100)	0 (0)
2	42 (100)	0 (0)
3	58 (100)	0 (0)
4	78 (93)	6 (7)
5	73 (85)	13 (15)
6	63 (74)	22 (26)
7	37 (49)	38 (51)
8	16 (33)	33 (67)
9	6 (14)	36 (86)
10	5 (23)	17 (77)
11	0 (0)	8 (100)
12	0 (0)	1 (100)
13	0 (0)	1 (100)
14	0 (0)	0 (0)

* Values are the number (%) of patients with a given score. Scores were rounded to the nearest number ending in .5 or .0 (i.e., scores ≤0.5 are in the category 0, scores >0.5 and ≤1.5 are in the category 1, etc.). RA = rheumatoid arthritis.

Table 4. Cutoff values for prediction scores and risk of development of RA*

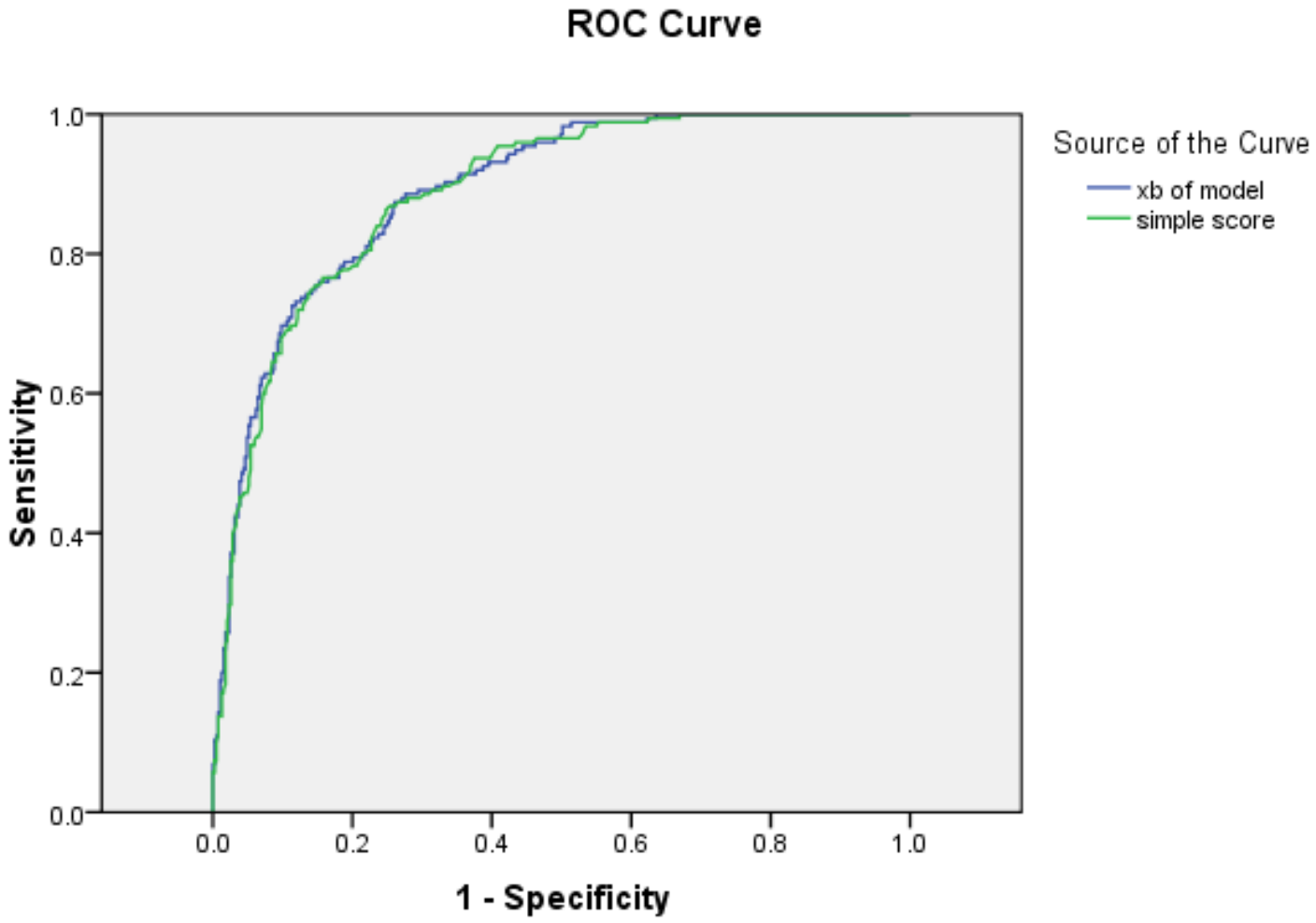
Cutoff values	No progression to RA	Progression to RA
Score ≤4.0	145 (99)	1 (1)
4.0–10.0	240 (60)	159 (40)
≥10.0	2 (12)	15 (88)
Score ≤5.0	223 (97)	8 (3)
5.0–9.0	157 (55)	131 (46)
≥9.0	7 (16)	36 (84)
Score ≤6.0	296 (91)	28 (9)
6.0–8.0	76 (52)	69 (48)
≥8.0	15 (16)	78 (84)

* Values are the number (%) of patients with a given score. Scores were rounded to the nearest number ending in .5 or .0. RA = rheumatoid arthritis.

How well does the model discriminate?

- Calculate for different cut-off values for prediction score sensitivity and specificity
- Make a ROC curve
- Calculate Area under the Curve, the so-called c- statistic.
- This c is usually between 0.5 (no discrimination) and 1 (perfect discrimination)
- c = proportion of case-non case pairs in which the case has indeed a larger prediction score

- AUC both models is 0.89



Validate on new data

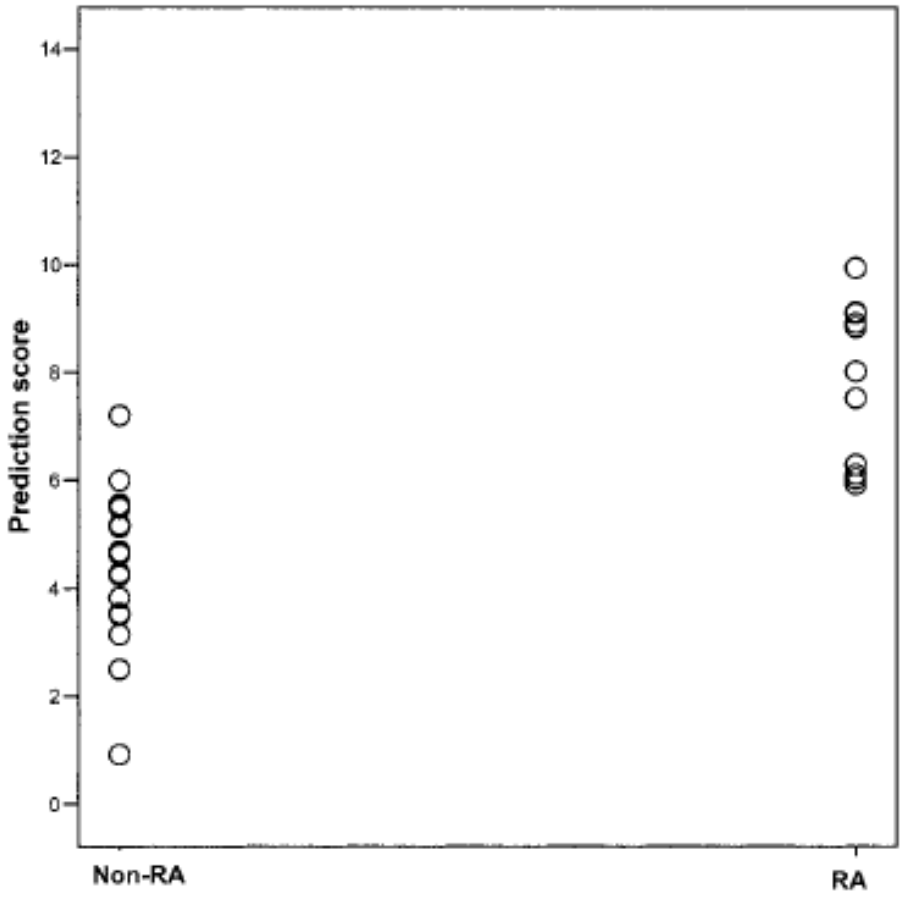


Figure 4. Prediction scores for patients with undifferentiated arthritis in whom rheumatoid arthritis (RA) did develop and those in whom RA did not develop.

Other logistic regression models

- For ordinal or nominal outcomes
  - ordinal or multinomial logistic regression

- For matched case-control data
  - Each case is matched to one or more controls

  - A matched design requires a matched analysis
    Ignoring the matching can yield biased estimates
  - Conditional logistic regression