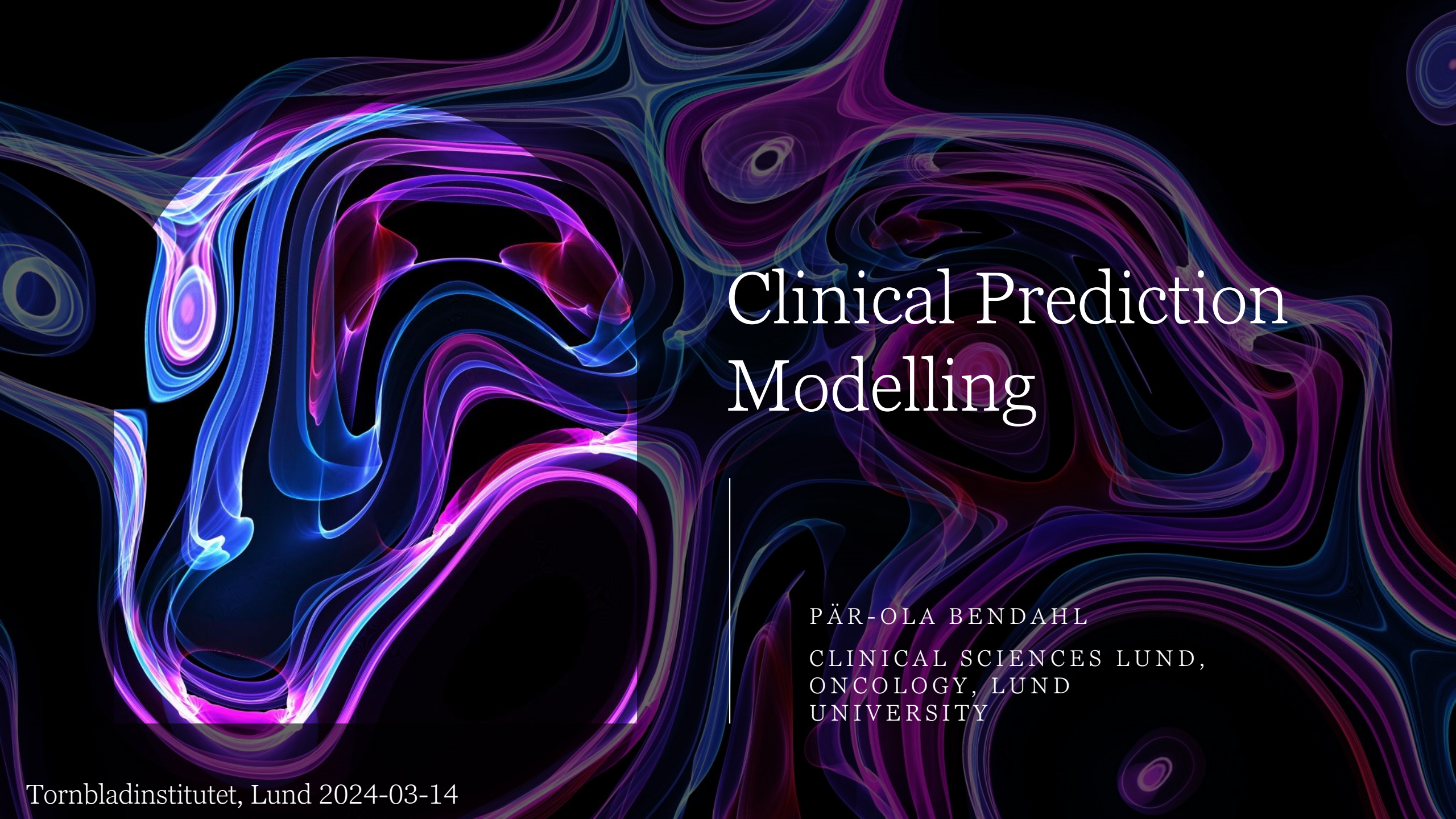


Clinical Prediction Modelling

PÄR-OLA BENDAHL

CLINICAL SCIENCES LUND, ONCOLOGY, LUND UNIVERSITY

Tornbladinstitutet, Lund 2024-03-14

Outline

- Examples
- The problem
- Methods
  - *Regression*
  - *Machine Learning*
- Validation
- (Clinical utility)

A multimarker diagnostic test

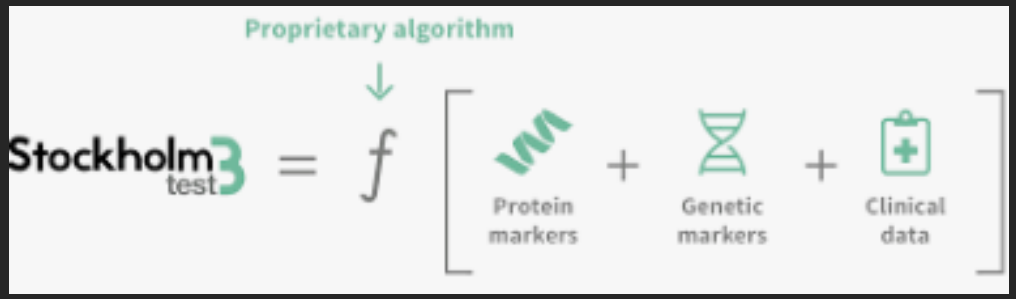
- Stockholm3

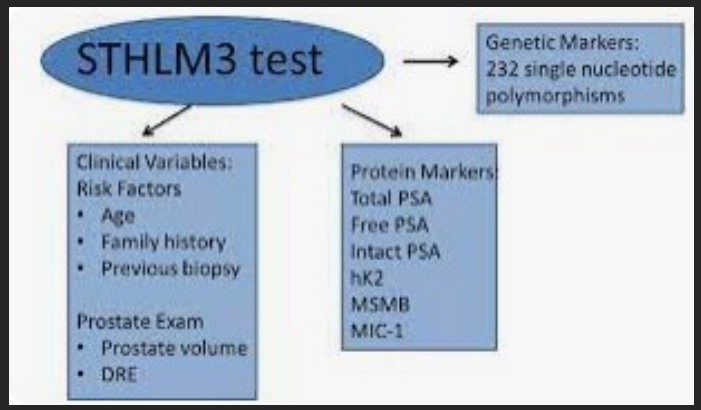
  Screening test for detection of advanced prostate cancer

  Proprietary algorithm

  More effective at predicting prostate cancer risk than PSA alone

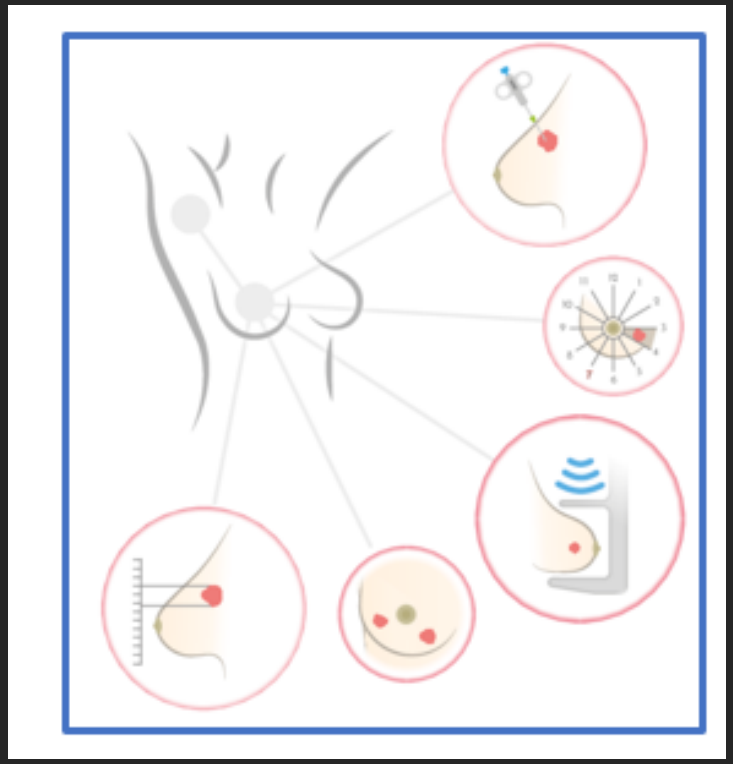
  Implemented in the Swedish health care system since 2016

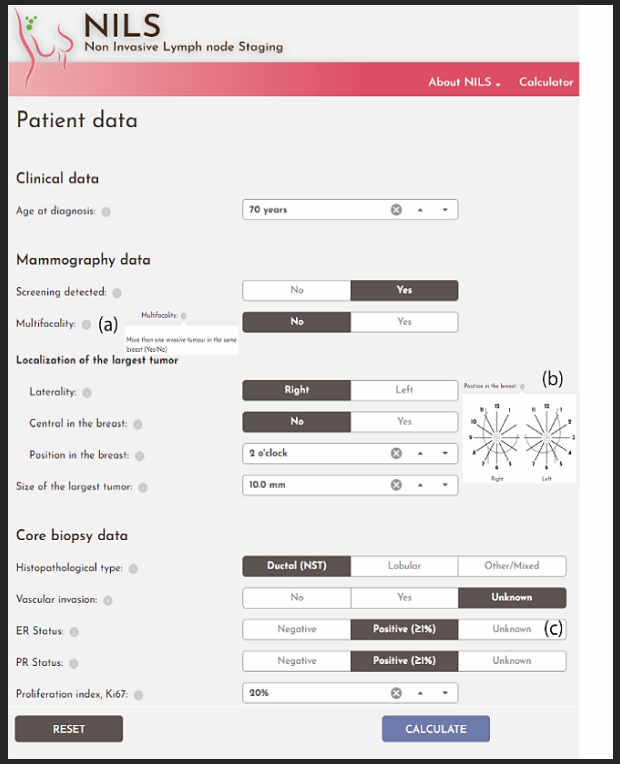


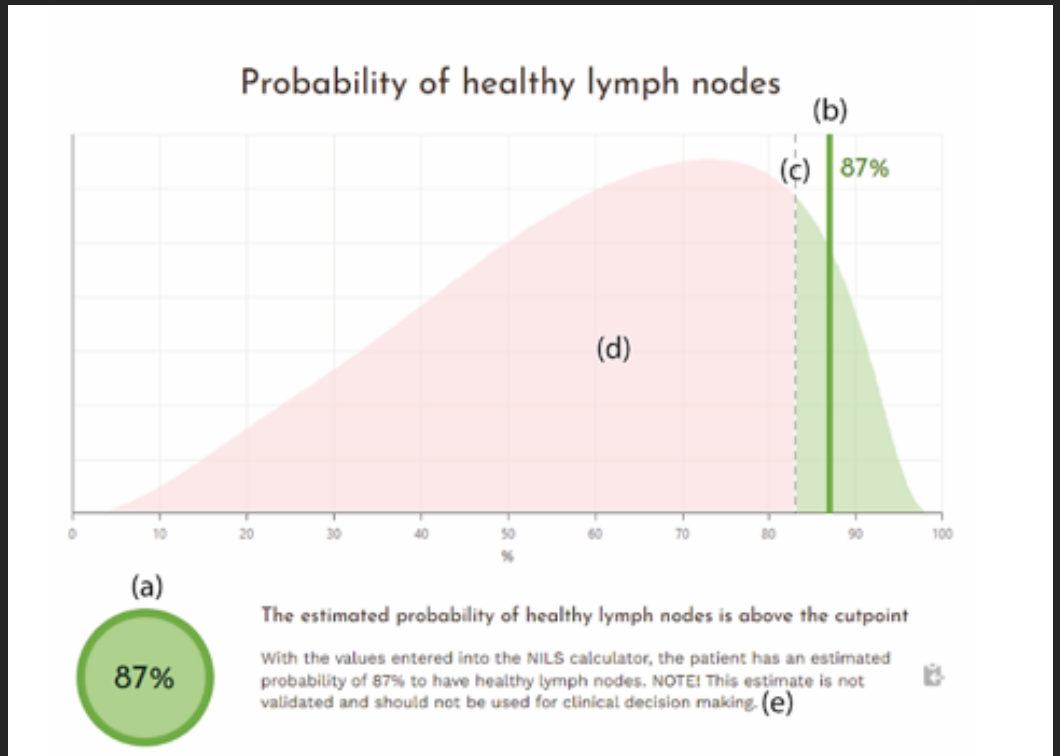


An example from our group (Rydén)

- NILS – Non-Invasive Lymph node Staging of the axilla (breast cancer)
- Knowledge gap: No existing prediction tools based on preoperatively available characteristics





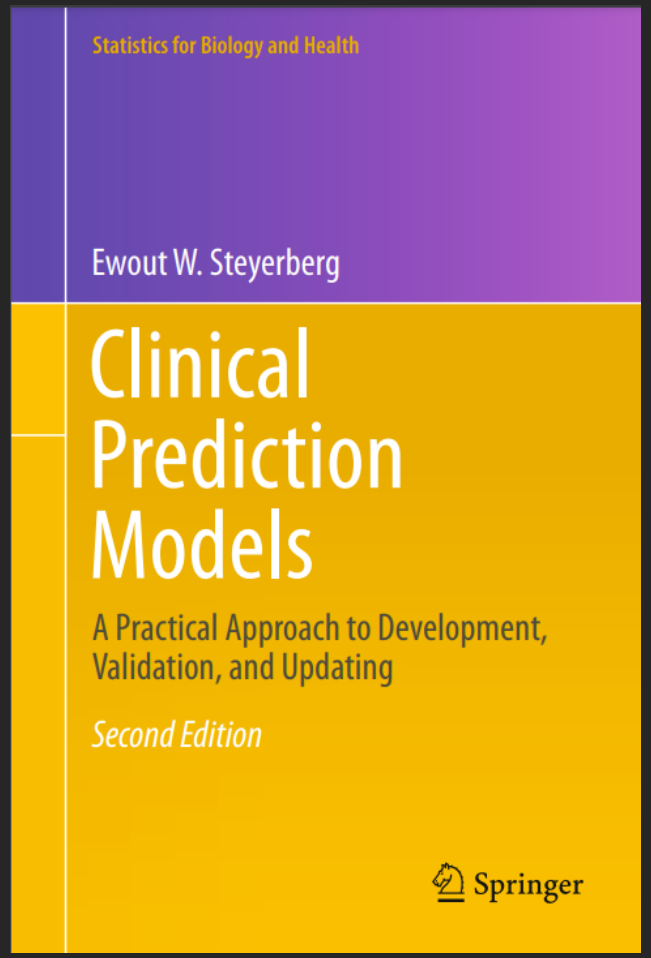


Figur: Looket Dihge

The steps

- Identify the need for a new decision support tool – extensive literature review
- Collect ”enough” relevant good quality data
- Develop and validate the model
- Assess its clinical value (prospective study)
- Implement the model

Recommended literature



The TRIPOD* guidelines

Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) : The TRIPOD statement

Reporting guideline provided for? (i.e. exactly what the authors state in the paper)	Reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. TRIPOD Checklist for Prediction Model Development: Word | PDF TRIPOD Checklist for Prediction Model Validation: Word | PDF TRIPOD Checklist for Prediction Model Development and Validation: Word | PDF
Full bibliographic reference	Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. This guideline was published simultaneously in 11 journals. You can read the guideline in any of these journals using the links below. Ann Intern Med. 2015;162(1):55-63. PMID: 25560714 Br J Cancer. 2015 Jan 6. PMID: 25562432 Circulation. 2015 Jan 13;131(2):211-9. PMID: 25561516 BMJ 2015; 350:g7594. PMID: 25569120 J Clin Epidemiol. 2015 Feb;68(2):134-43. PMID: 25579640 Eur Urol. 2014 Dec 9. PMID: 25572824 BMC Med. 2015 Jan 6;13(1):1. PMID: 25563062 Eur J Clin Invest. 2015;45(2):204-214. PMID: 25623047 Br J Surg. 2015;102(3):148-158. PMID: 25627261 BJOG. 2015;122(3):434-443. PMID: 25623578 Diabet Med. 2015;32(2):146-154. PMID: 25600898

* https://www.equator-network.org/reporting-guidelines/tripod-statement/

Well defined target population

- For which population is the prediction model intended?
- Stockholm 3: All men aged 45-74 with no previous prostate cancer diagnosis
- NILS: Female primary breast cancer patients

Well defined outcome

- Binary outcome in this lecture (most of the ideas applicable also to other types of outcomes)
- Advanced prostate cancer (yes/no)
- Breast cancer spread to the axillary lymph nodes (N+; yes/no)

    At least one macrometastasis (>2mm)

    At least one micrometastasis (>0.2mm)

    At least isolated tumor cells

The goal

To develop a model with good discrimination and calibration upon validation in independent datasets

To Capture the signal in the data used for model development, not the noise
A trade-off between bias and variance

All models are wrong but some are useful

George Box 1976

Since all models are wrong the scientist cannot obtain a ”correct” one by excessive elaboration

Just as the ability to devise simple but evocative models is a signature of the great scientist so overelaboration and overparametrization is often the mark of mediocrity

Problems to avoid

Overfitting (and underfitting)

- Overparametrization
- Overtraining
- Overoptimism

The principle of parsimony

—

Occham's razor

William Occham (1287-1347)

The problem-solving principle that recommends searching for explanations constructed with the smallest possible set of elements

Find a simple model that performs well upon external validation

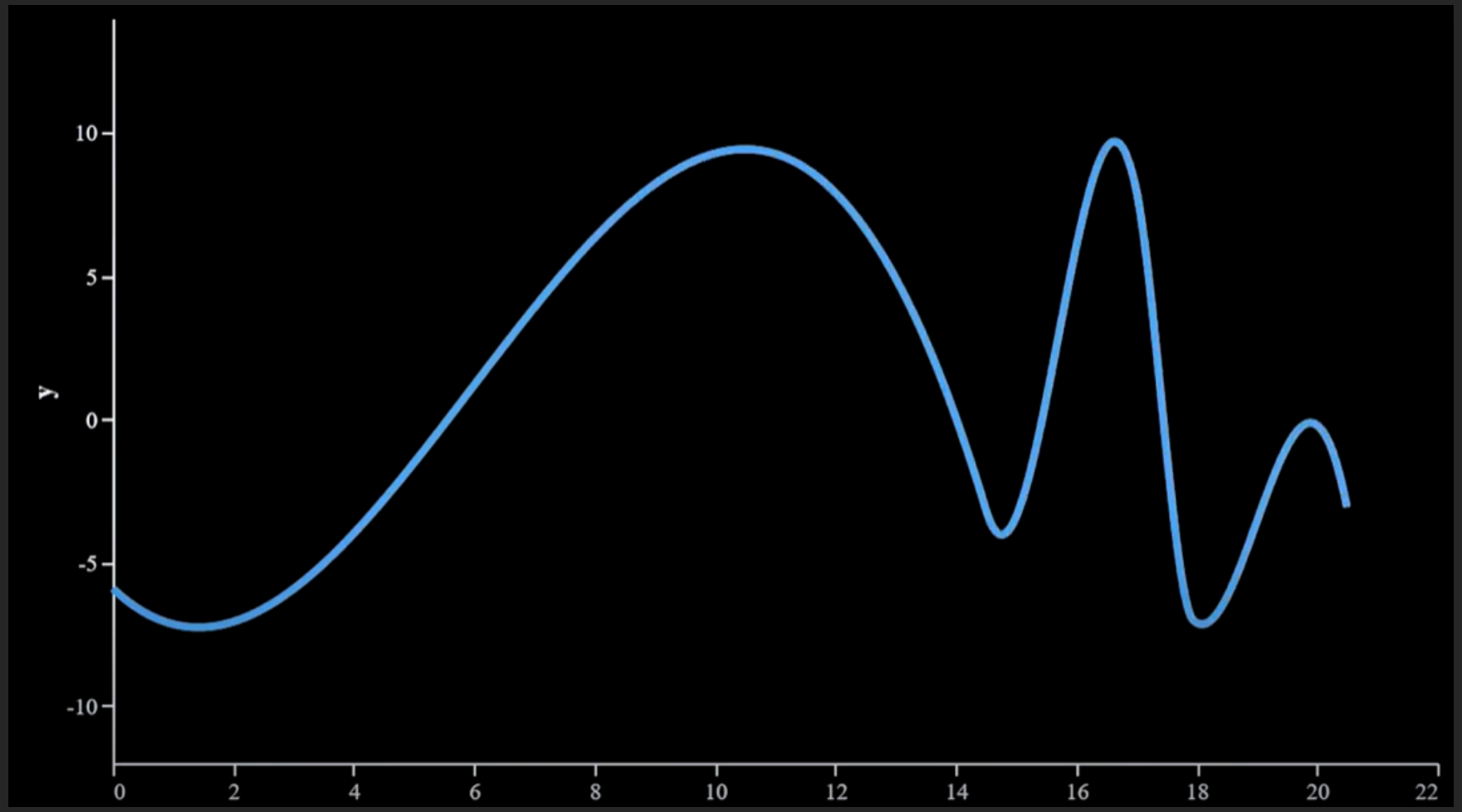
Break this rule if you have good reasons to do so

Ongoing project: Prediction of lymph node status using features from mammograpgy images

Convolutional neural networks with up to 2 million parameters

Hundreds of CPU hours to train a model

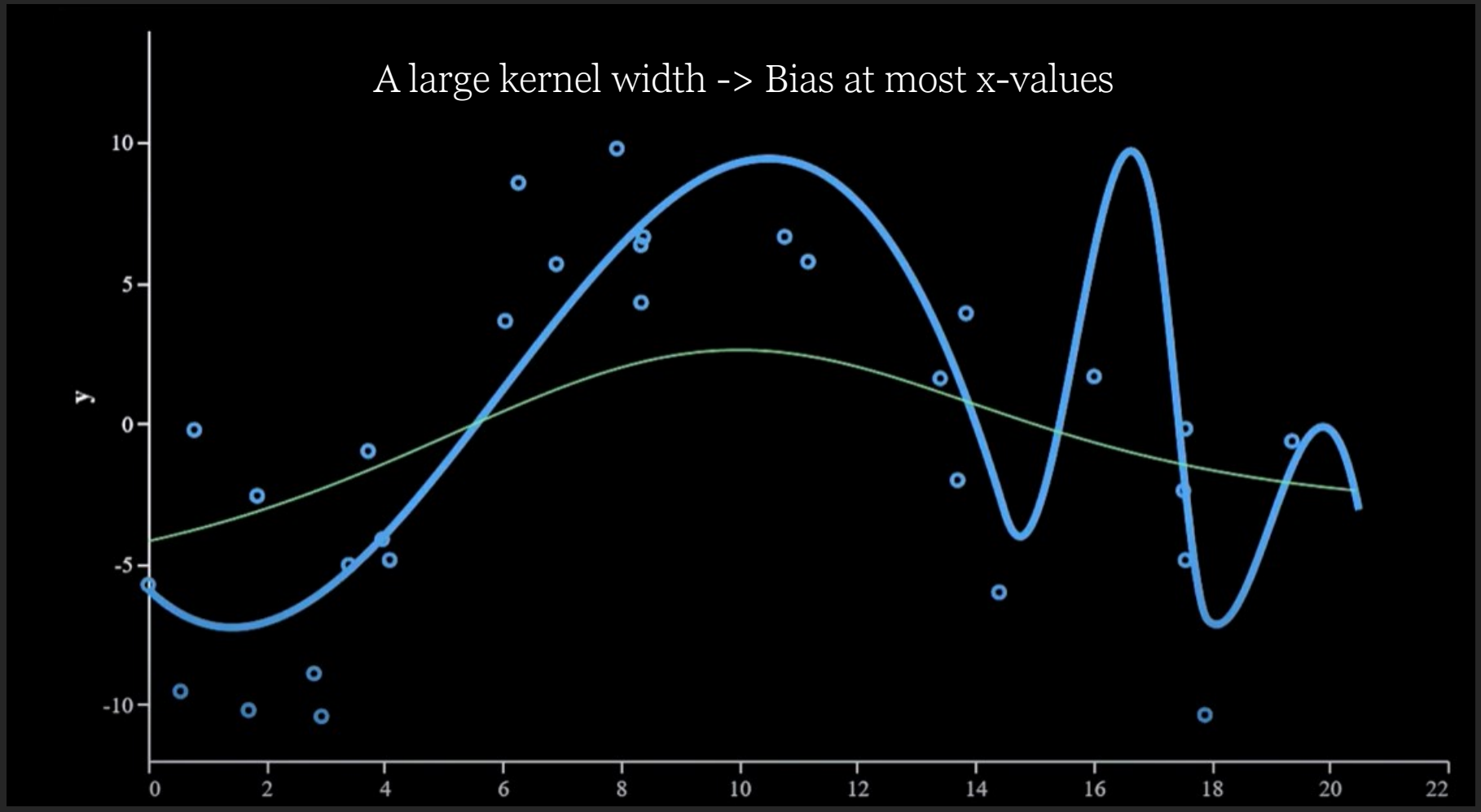
Example: Kernel estimate of a known function f



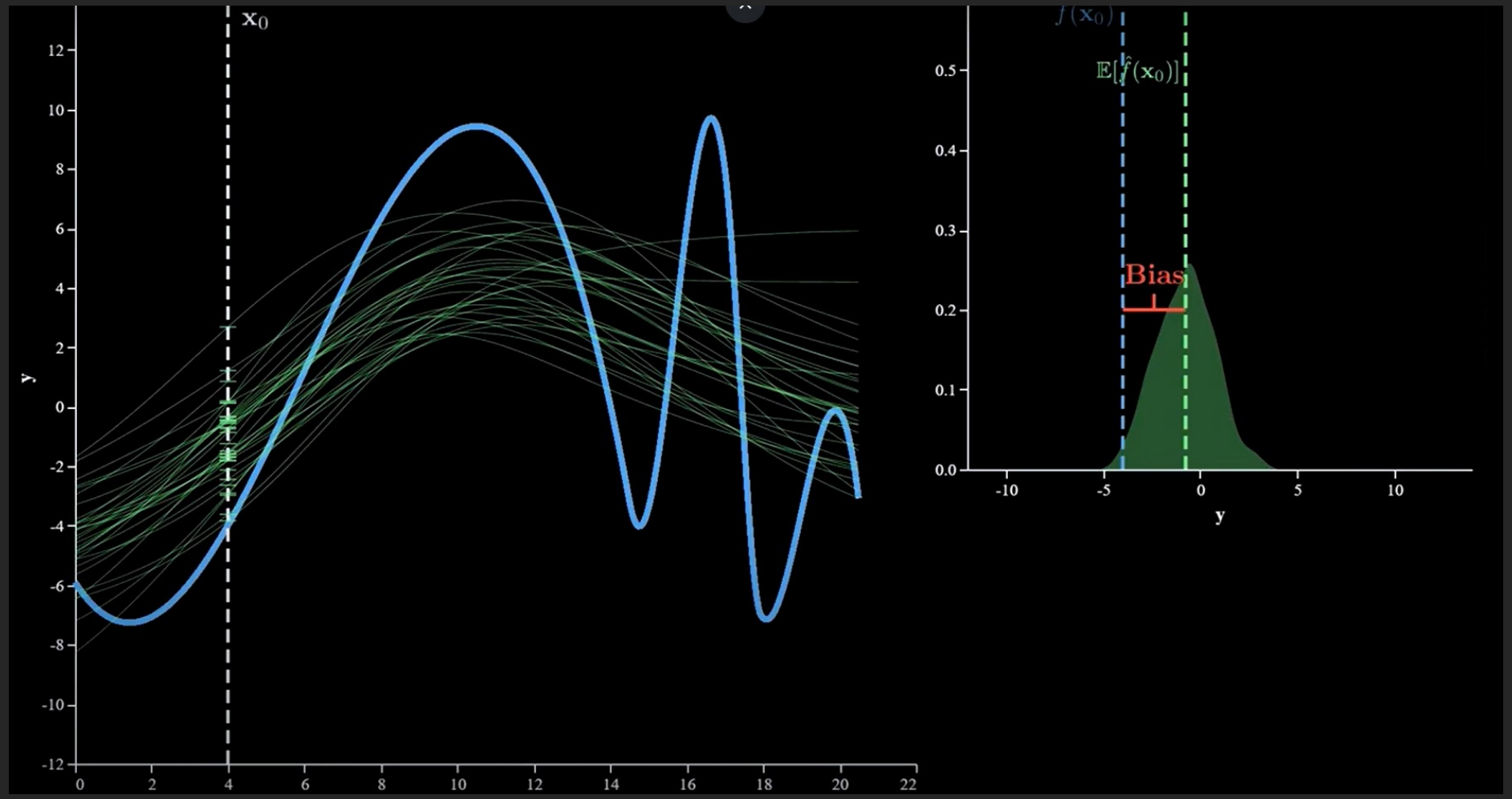
Source: The Bias Variance Trade-Off (youtube.com)

Simulate data from f and fit a Gaussian kernel regression model to the data

The complexity of the model is determined by a hyperparameter, the width of the regression kernel

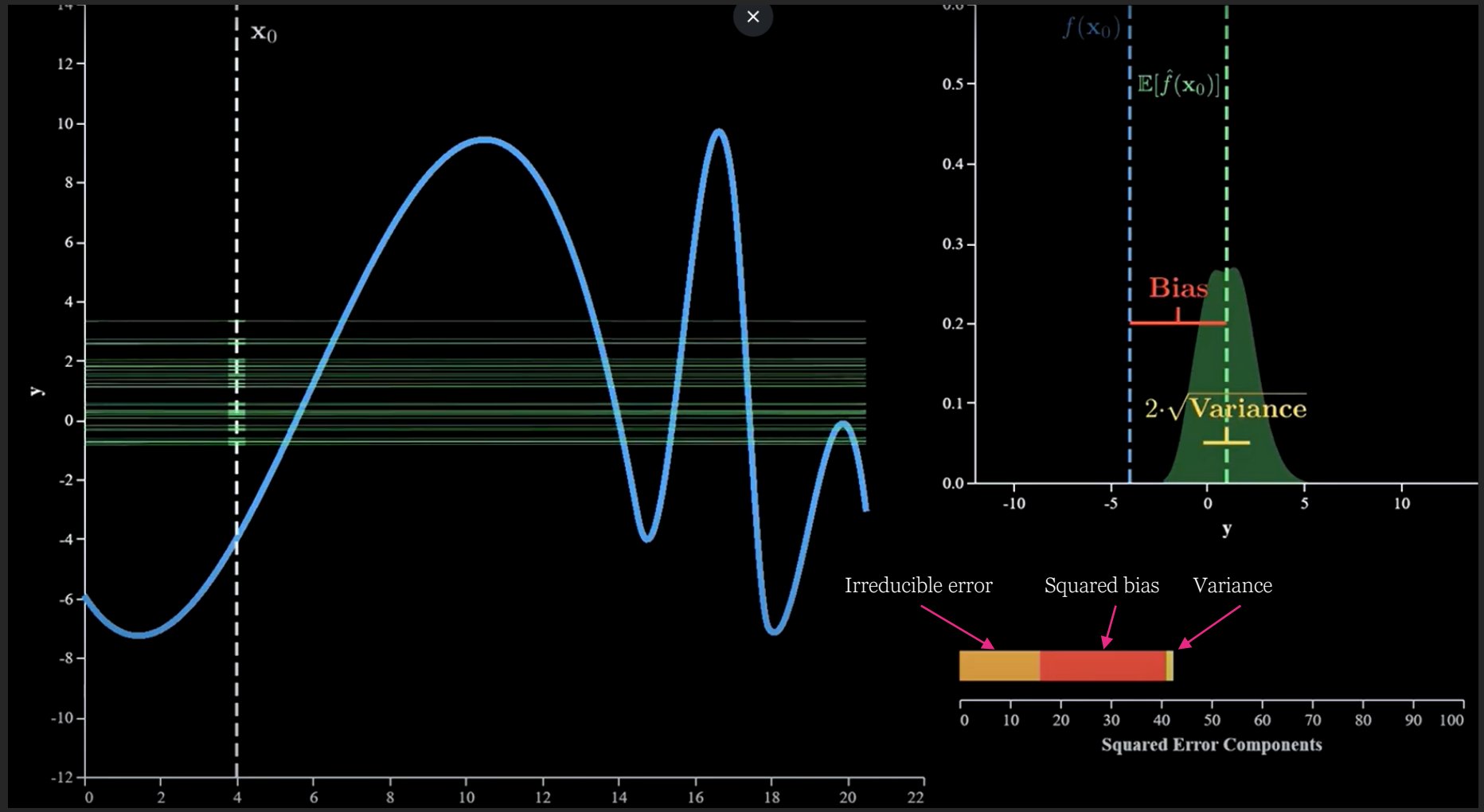


The average prediction at X_0 is biased for this large kernel width

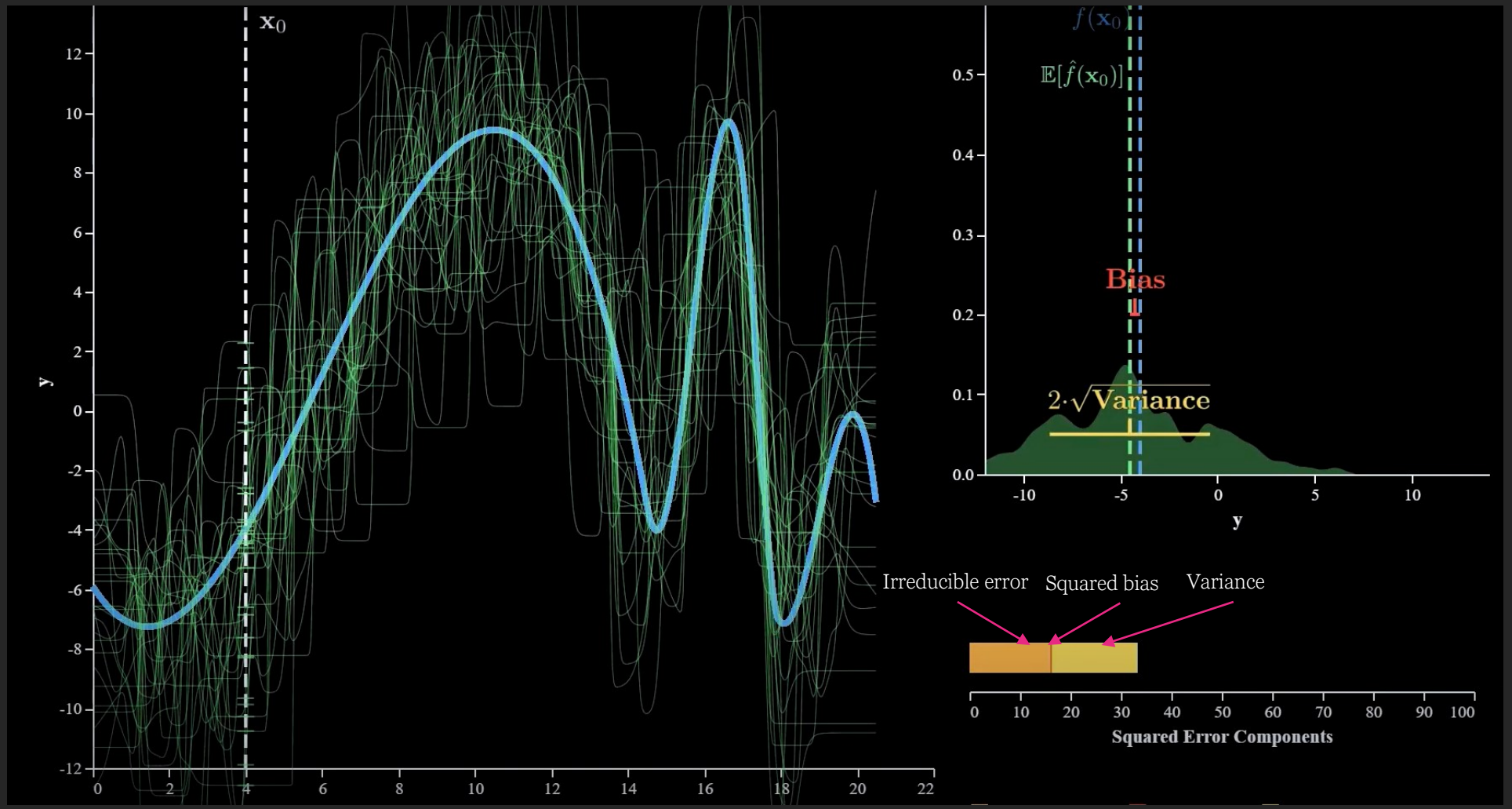


Even larger kernel width (underfit)

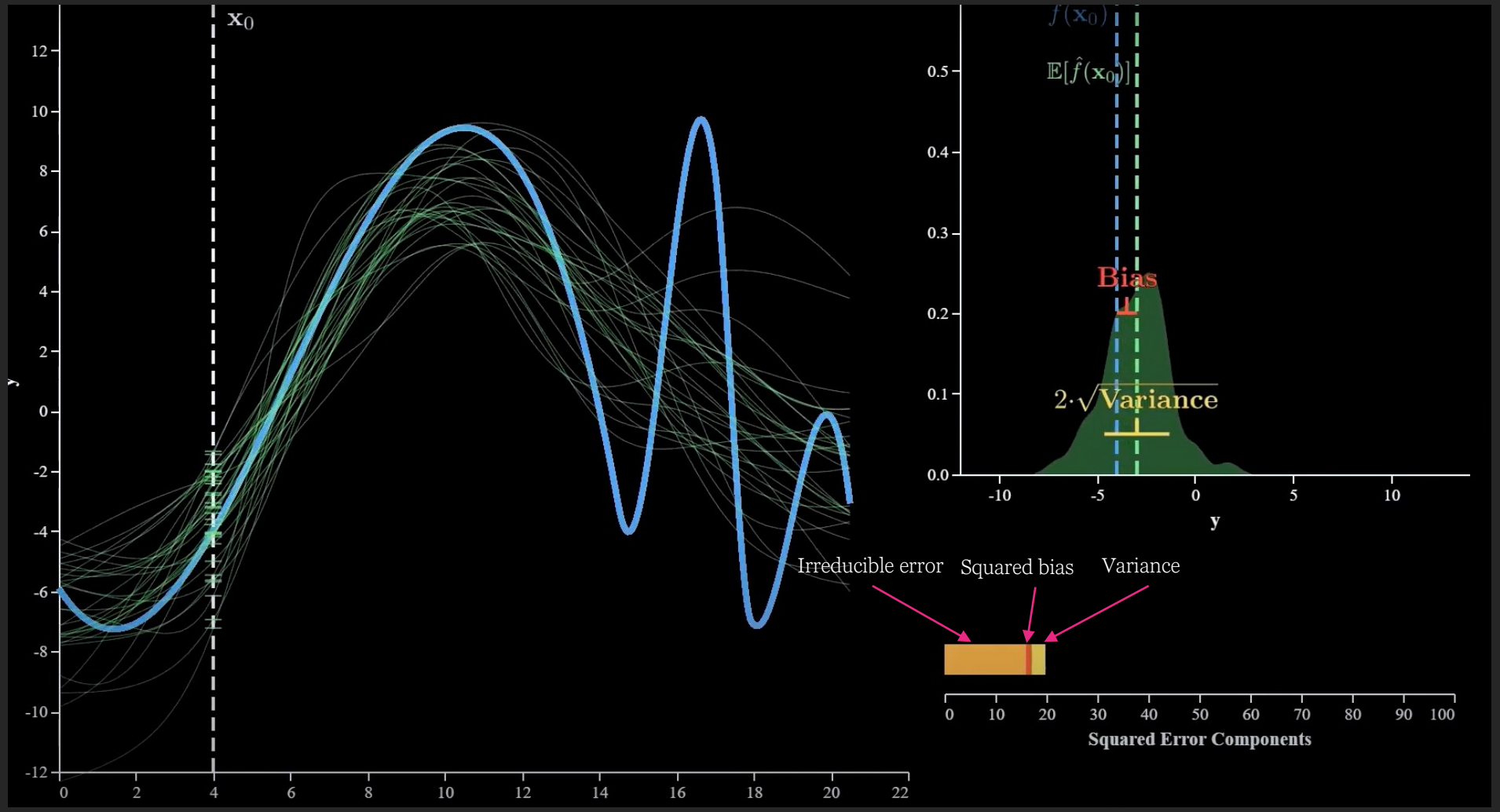
Total squared error = Irreducible error + Squared bias + Variance



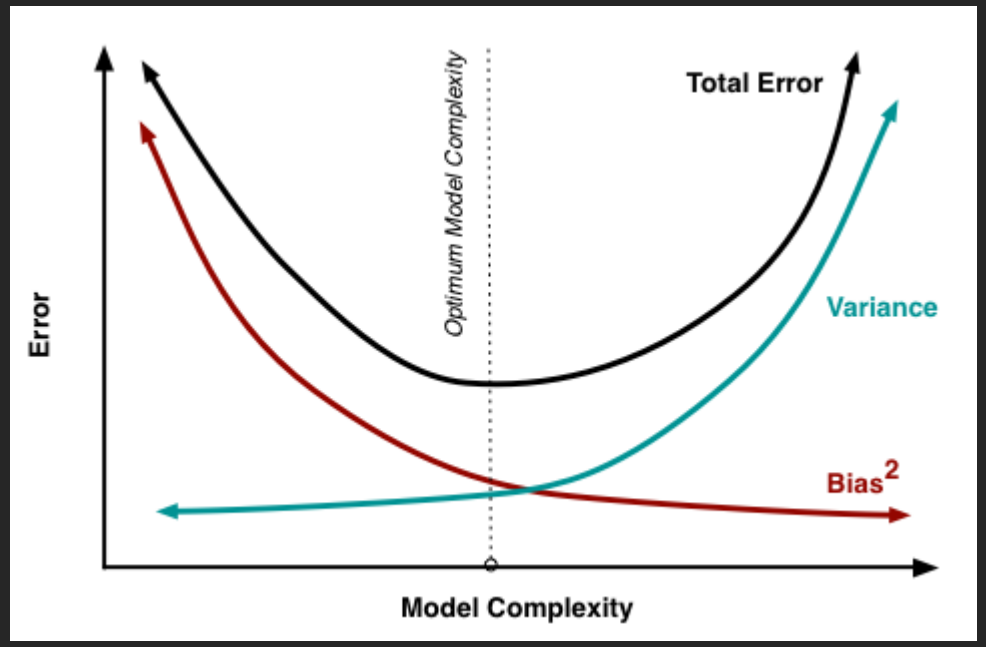
Small kernel width (overfit)



The perfect balance between bias and variance (at X_0)
Total squared error minimized

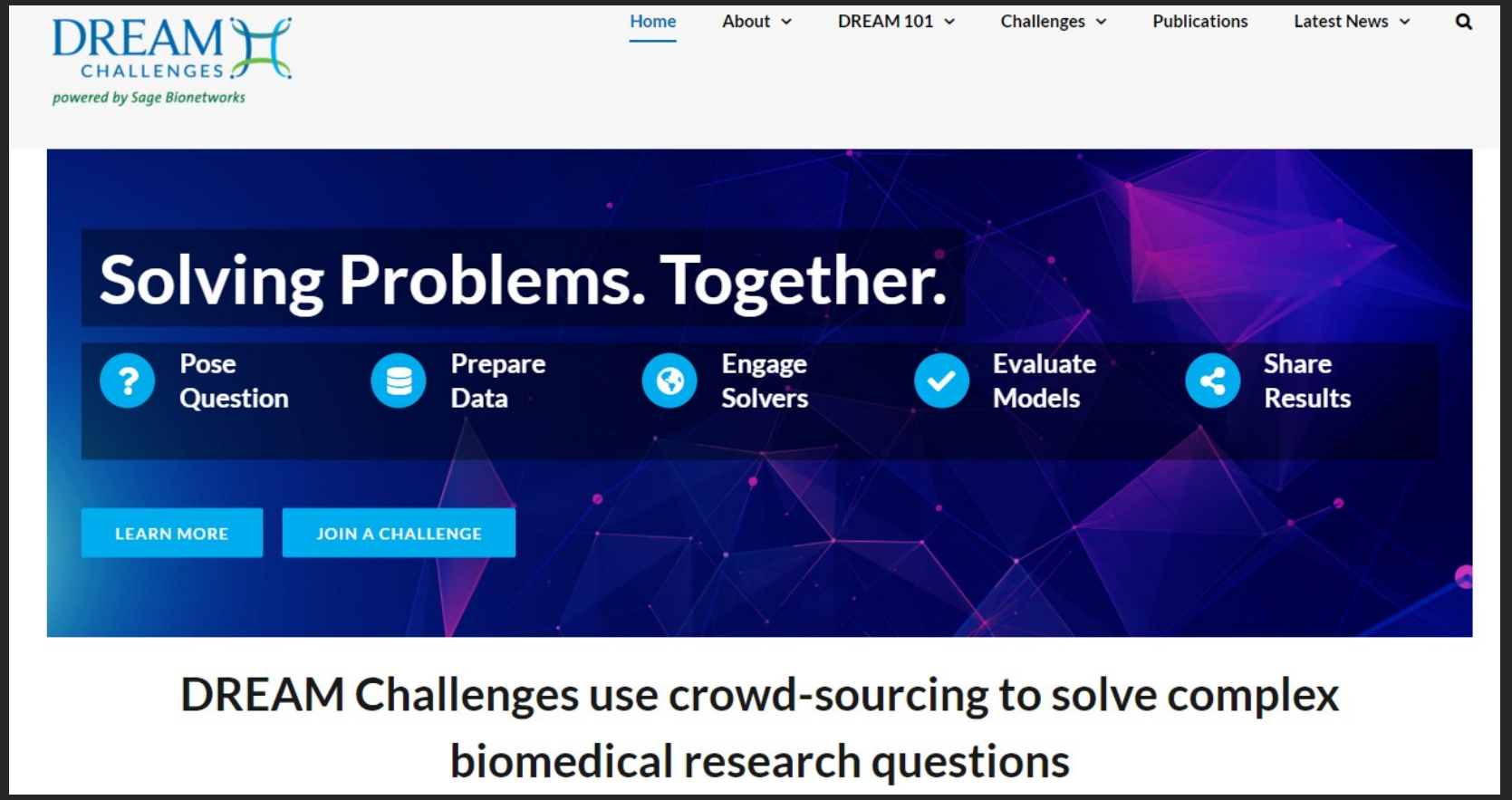


To sum up



Back to clinical prediction modelling

DREAM Challenges



Hard prediction problems

Large sample size

Many potential predictors

Often omics data

Cheap to participate

Top list of best fitting models

Always overfitted

Price money for best performing model on set-aside test data

Lessons from DREAM challanges

- Subject matter knowledge matters!
- 100% data driven models seldom winners

Regression Modelling or Machine Learning?

Dataset used to compare modelling strategies

The first version of the NILS model

Complete case analysis (n=588; 197 N+)

”The rule of 20” -> Up to 10 parameters in the model

Variable selection based mainly on subject matter knowledge (not data driven)

Nomograms for preoperative prediction of axillary nodal status in breast cancer

L. Dihge[1,3], P.-O. Bendahl[2] and L. Rydén[1,4]

Departments of [1]Surgery and [2]Oncology and Pathology, Clinical Sciences Lund, Lund University, Lund, and Departments of [3]Plastic and Reconstructive Surgery and [4]Surgery, Skåne University Hospital, Malmö, Sweden

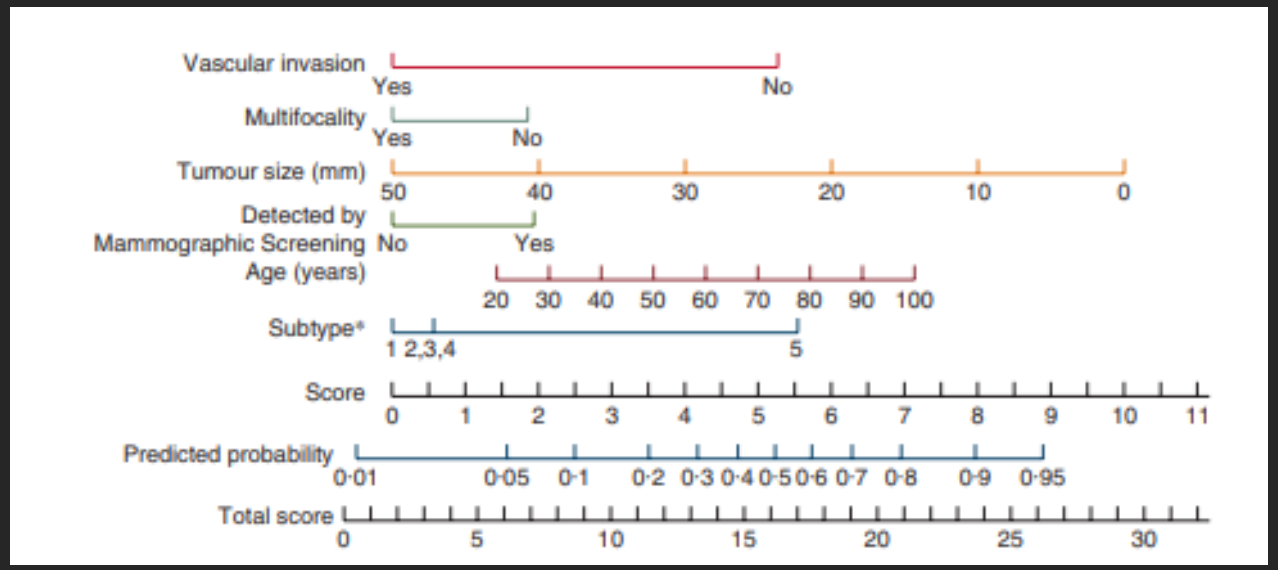
Correspondence to: Professor L. Rydén, Lund University, Faculty of Medicine, Department of Clinical Sciences Lund, Surgery, Medicon Village, SE-223 81, Lund, Sweden (e-mail: lisa.ryden@med.lu.se)

BJS 2017; 104: 1494-1505

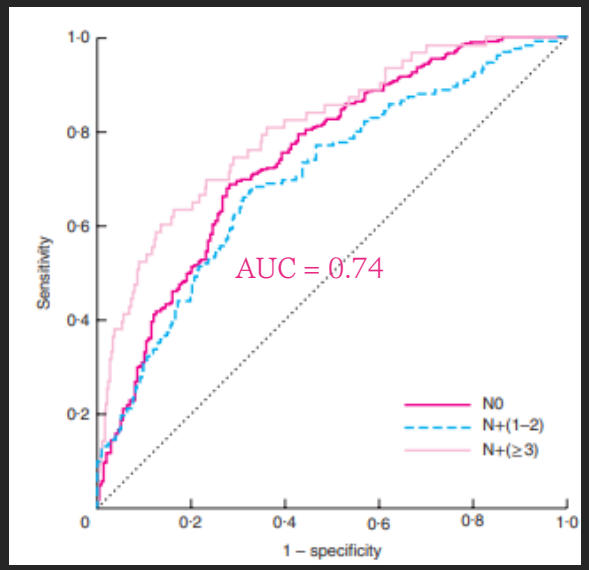
Logistic regression

	N0 *versus* N+ (*n* = 598)	
	Odds ratio	*P*
Subtype		0·031
LumA	1·00	
LumB/HER2−	1·18 (0·76, 1·84)	
LumB/HER2+	1·11 (0·56, 2·21)	
HER2+/non-luminal	1·48 (0·40, 5·44)	
Triple-negative	5·06 (1·89, 13·50)	
Age (per year)	1·02 (1·00, 1·04)	0·013
Mode of detection		0·006
Symptomatic	1·00	
Mammographic screening	1·75 (1·18, 2·61)	
Tumour size (per mm)	0·94 (0·92, 0·97)	< 0·001
Multifocality		0·015
Yes	1·00	
No	1·72 (1·11, 2·65)	
Vascular invasion		< 0·001
Yes	1·00	
No	4·67 (2·70, 8·09)	

Nomogram



ROC analysis



Model validated by Majid et al. BJS Open 2021: The same AUC = 0.74 and good calibration

Data split recommended for large dataset

1. Training set (for model development; 70%)
2. Validation set (first validation of performance; 15%; finetuning of hyperparameters)
3. Test set (final evaluation of performace; 15%)

Note 1: The N-status dataset is too small for splitting into three sets so I merge validation and test (30%)

Note 2: Split conditional on the outcome to guarantee the same outcome prevalence in the subsets

Model development in the training dataset

The Binary outcome: N_plus

```
. tabulate N_plus if set==1
```

N_plus	Freq.	Percent	Cum.
0	281	67.06	67.06
1	138	32.94	100.00
Total	419	100.00	

Evaluated predictors

• Patient age (years)	Continuous
• Mode of detectiong (screening/symtomatic)	Binary
• Tumor size (mm)	Continuous
• Multifocality (yes/no)	Binary
• Lymphovascular invasion (yes/no)	Binary
• Molecular subtype (factor, 5 levels)	4 dummy variables

Model 1: Logistic regression, the BJS model

```
. logistic N_plus *LumA age scr tum_size multifoc LVI

Logistic regression                             Number of obs =    419
                                                LR chi2(9)    =  89.37
                                                Prob > chi2   = 0.0000
Log likelihood = -220.84396                     Pseudo R2     = 0.1683
```

N_plus	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
LumB_HER2neg_vs_LumA	.7685602	.2104684	-0.96	0.336	.4493434	1.314551
LumB_HER2pos_vs_LumA	1.050976	.4896771	0.11	0.915	.4216927	2.619328
HER2pos_nonlum_vs_LumA	.5071292	.384929	-0.89	0.371	.1145583	2.244971
TN_vs_LumA	.1325429	.0879352	-3.05	0.002	.0361101	.4865012
age	.9750664	.0100907	-2.44	0.015	.9554883	.9950457
scr_det	.4459999	.1109787	-3.24	0.001	.2738604	.7263407
tum_size	1.048318	.0146677	3.37	0.001	1.01996	1.077464
multifoc	1.765435	.4636465	2.16	0.030	1.055129	2.953916
LVI	5.988399	2.132529	5.03	0.000	2.979815	12.03461
_cons	1.357134	.9775351	0.42	0.672	.3307558	5.568495

Model performance in the training set

```
. estat ic

Akaike's information criterion and Bayesian information criterion
```

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	419	-265.5292	-220.844	10	461.6879	502.0666

```
Note: BIC uses N = number of observations. See [R] IC note.

. predict phat
(option pr assumed; Pr(N_plus))

. roctab N_plus phat
```

Obs	ROC area	Std. err.	Asymptotic normal [95% conf.	interval]
419	0.7505	0.0259	0.69971	0.80129

```
. gen sq_error=(phat-N_plus)^2

. sum sq_error
```

Variable	Obs	Mean	Std. dev.	Min	Max
sq_error	419	.1756163	.2082388	.0003843	.8128378

Model 2: Logistic regression, the BJS model

Stepwise backward elimination, pr(0.157)

```
. stepwise, pr(.157): logistic N_plus *LumA age scr tum_size multifoc LVI

Wald test, begin with full model:
p = 0.9150 >= 0.1570, removing LumB_HER2pos_vs_LumA
p = 0.3633 >= 0.1570, removing HER2pos_nonlum_vs_LumA
p = 0.3630 >= 0.1570, removing LumB_HER2neg_vs_LumA

Logistic regression                               Number of obs =    419
                                                  LR chi2(6)    =  87.64
                                                  Prob > chi2   = 0.0000
Log likelihood = -221.70836                       Pseudo R2     = 0.1650
```

N_plus	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
tum_size	1.046248	.0143387	3.30	0.001	1.018519	1.074733
LVI	5.747705	2.006036	5.01	0.000	2.900095	11.39139
multifoc	1.767931	.4625329	2.18	0.029	1.058695	2.952297
TN_vs_LumA	.1517574	.0973275	-2.94	0.003	.0431761	.5334037
age	.9754421	.0099687	-2.43	0.015	.9560982	.9951773
scr_det	.4575946	.111754	-3.20	0.001	.2835307	.7385192
_cons	1.22745	.8600713	0.29	0.770	.310863	4.846617

Model performance in the training set

```
. estat ic

Akaike's information criterion and Bayesian information criterion
```

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	419	-265.5292	-221.7084	7	457.4167	485.6818

```
Note: BIC uses N = number of observations. See [R] IC note.

. predict phat
(option pr assumed; Pr(N_plus))

. roctab N_plus phat
```

Obs	ROC area	Std. err.	Asymptotic normal [95% conf.	interval]
419	0.7533	0.0255	0.70334	0.80334

```
. gen sq_error=(phat-N_plus)^2

. sum sq_error
```

Variable	Obs	Mean	Std. dev.	Min	Max
sq_error	419	.1768842	.2068361	.0004445	.7894332

Improved fit as measured by AIC and AUC, fewer parameters, more data driven, risk for overfit?

Model 3: Logistic regression, restricted cubic splines for tumor size Stepwise backward elimination, pr(0.157)

The hyperparameter number of knots varied for tumorsize

Three knots optimal as measured by AIC

Splines for age did not improve AIC

```
. makespline rcs tum_size, knots(3)

. stepwise, pr(.157): logistic N_plus *LumA age scr _r* multifoc LVI

Wald test, begin with full model:
p = 0.9219 >= 0.1570, removing LumB_HER2pos_vs_LumA
p = 0.4252 >= 0.1570, removing HER2pos_nonlum_vs_LumA
p = 0.3129 >= 0.1570, removing LumB_HER2neg_vs_LumA

Logistic regression                                   Number of obs =    419
                                                      LR chi2(7)    =  89.77
                                                      Prob > chi2   = 0.0000
Log likelihood = -220.64302                           Pseudo R2     = 0.1690
```

N_plus	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
_rcs_1_1	.0254117	.0649928	-1.44	0.151	.000169	3.819977
LVI	5.537328	1.941816	4.88	0.000	2.784857	11.01026
multifoc	1.741929	.4582316	2.11	0.035	1.040192	2.917074
TN_vs_LumA	.1644663	.1035914	-2.87	0.004	.0478559	.5652211
age	.9766961	.0100552	-2.29	0.022	.9571858	.996604
scr_det	.4693905	.114999	-3.09	0.002	.2903975	.7587096
_rs_rcs_1	177.9263	389.535	2.37	0.018	2.436125	12995.13
_cons	.6250788	.5389034	-0.55	0.586	.1153673	3.386779

Model performance in the training set

```
. estat ic

Akaike's information criterion and Bayesian information criterion
```

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	419	-265.5292	-220.643	8	457.286	489.589

```
Note: BIC uses N = number of observations. See [R] IC note.

. predict phat
(option pr assumed; Pr(N_plus))

. roctab N_plus phat
```

Obs	ROC area	Std. err.	Asymptotic normal [95% conf.	interval]
419	0.7574	0.0254	0.70754	0.80729

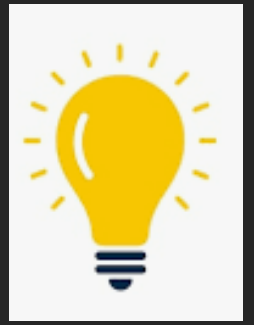
```
. gen sq_error=(phat-N_plus)^2

. sum sq_error
```

Variable	Obs	Mean	Std. dev.	Min	Max
sq_error	419	.1753244	.2056609	.0002379	.8230565

Best fit so far, more flexibility = more parameters, more data driven, higer risk for overfit (lower of underfit)

Idea – Punish complex models



AIC penalizes large model

Let k be the number of estimated parameters and L the estimated maximum likelihood for a model, then

$AIC = 2k - \ln(L)$

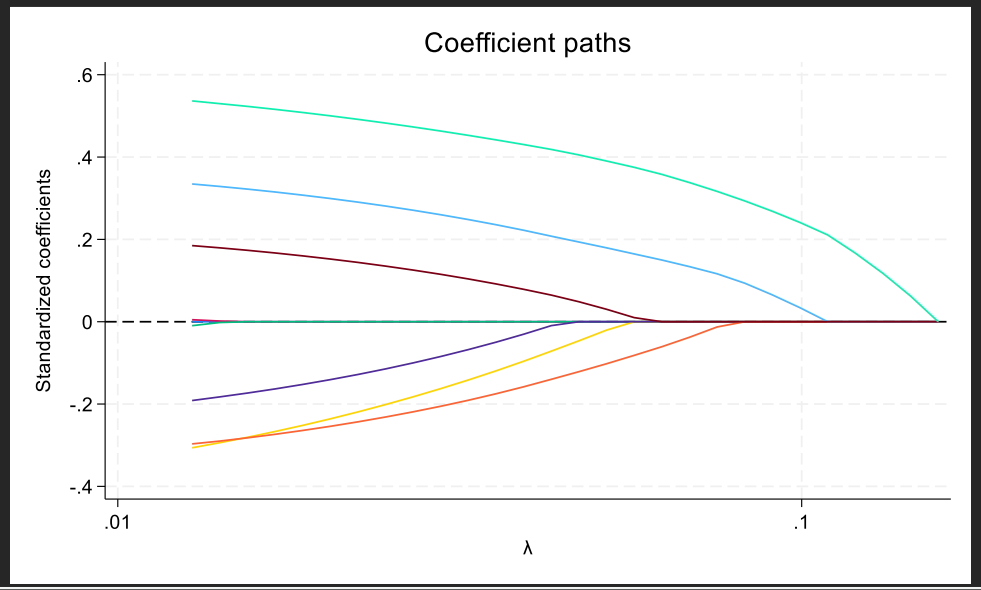
We want L to be as large as possible, hence $-\ln(L)$ to be as small as possible

Alternative – Penalized regression (shrinkage)

Maximize	$L - \lambda * F(\beta)$	
λ	hyperparemeter	
β	A vector of regression coefficients for the standardized predictors	
F	$\beta_1^2 + \ldots + \beta_k^2$	Ridge regression
F	$\lvert\beta_1\rvert + \ldots + \lvert\beta_k\rvert$	LASSO

Ridge *vs.* LASSO

- The Ridge penalty shrinks the parameters towards zero, but never all the way down to zero
- A large LASSO penalty leads to maximum on the boundary where some regression coefficients are = 0
- Hence LASSO can be used as a tool for variable selection



Ridge regression works best when most of the evaluated predictors are useful

LASSO regression works best when most of the evaluated preditors are useless

Lambda is often chosen by K-fold cross validation (default 10-fold in Stata)

Larger penalty

Model 4: Ridge regression, λ chosen by 10-fold cross validation

```
. elasticnet logit N_plus *LumA age scr tum_size multifoc LVI, rseed(1234) selection(cv) alpha(0)

Evaluating up to 100 lambdas in grid ...
```

```
Elastic net logit model                 No. of obs          =        419
                                        No. of covariates   =          9
Selection: Cross-validation             No. of CV folds     =         10
```

alpha	ID	Description	lambda	No. of nonzero coef.	Out-of-sample dev. ratio	CV mean deviance
0.000						
	1	first lambda	158.3391	9	-0.0052	1.273972
	86	lambda before	.0582432	9	0.1216	1.113365
	* 87	selected lambda	.053069	9	0.1216	1.113342
	88	lambda after	.0483545	9	0.1215	1.113445
	100	last lambda	.0158339	9	0.1162	1.120212

```
. lassocoef, display(coef)
```

	active
LumB_HER2neg_vs_LumA	-.039115
LumB_HER2pos_vs_LumA	.0543951
HER2pos_nonlum_vs_LumA	-.0578412
TN_vs_LumA	-.2835752
age	-.1968776
scr_det	-.2869108
tum_size	.3292159
multifoc	.1999622
LVI	.472077
_cons	-.7875241

Penalized standardized regression coefficients

Model performance in the training set

```
. predict phat
(options pr penalized assumed; Pr(N_plus) with penalized coefficients)

. roctab N_plus phat
```

Obs	ROC area	Std. err.	Asymptotic normal [95% conf.	interval]
419	0.7524	0.0257	0.70209	0.80263

```
. gen sq_error=(phat-N_plus)^2

. sum sq_error
```

Variable	Obs	Mean	Std. dev.	Min	Max
sq_error	419	.1770398	.1891919	.0030373	.7350298

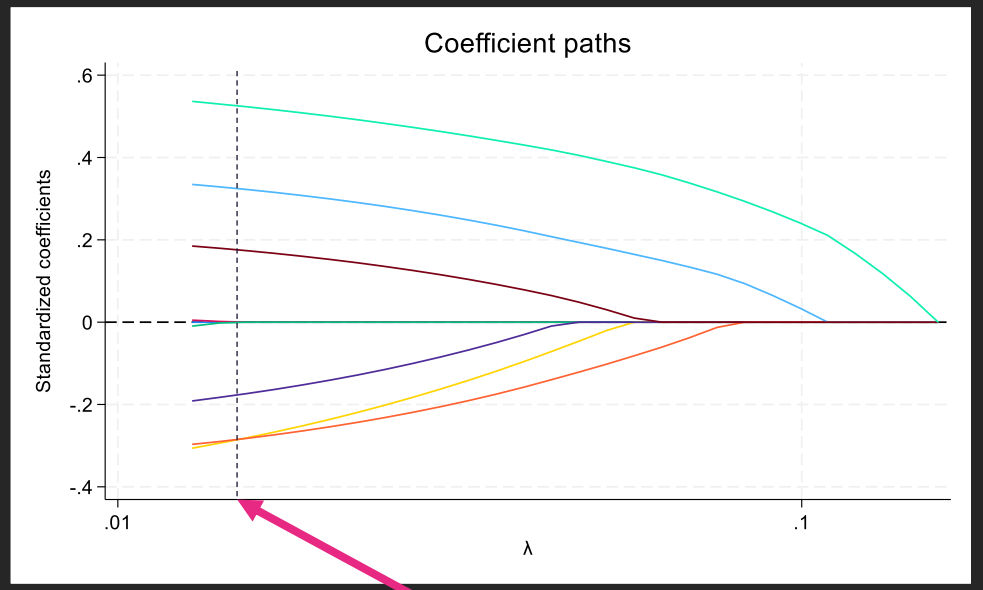
Ridge regression modelling with splines did not improve discrimination

Penalization should in theory safeguard against overfit

Model 5: LASSO regression, λ chosen by BIC criterion

```
. lasso logit N_plus *LumA age scr tum_size multifoc LVI, rseed(1234) selection(bic)

Evaluating up to 100 lambdas in grid ...
```



Selected λ = 0.1547; 6 nonzero coefficients

Model performance in the training set

```
. predict phat
(options pr penalized assumed; Pr(N_plus) with penalized coefficients)

. roctab N_plus phat
```

Obs	ROC area	Std. err.	Asymptotic normal [95% conf.	interval]
419	0.7571	0.0252	0.70768	0.80648

```
. gen sq_error=(phat-N_plus)^2

. sum sq_error
```

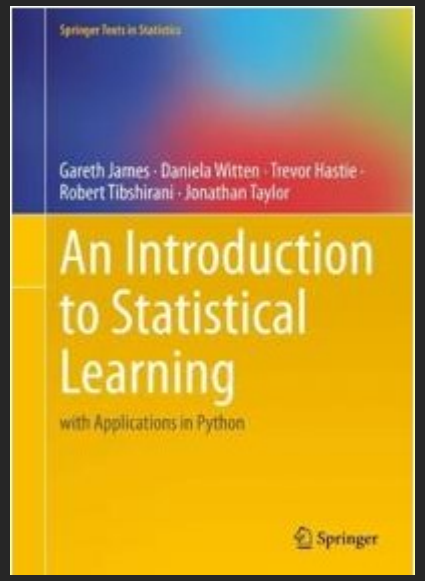
Variable	Obs	Mean	Std. dev.	Min	Max
sq_error	419	.1779631	.1913531	.0030763	.7229069

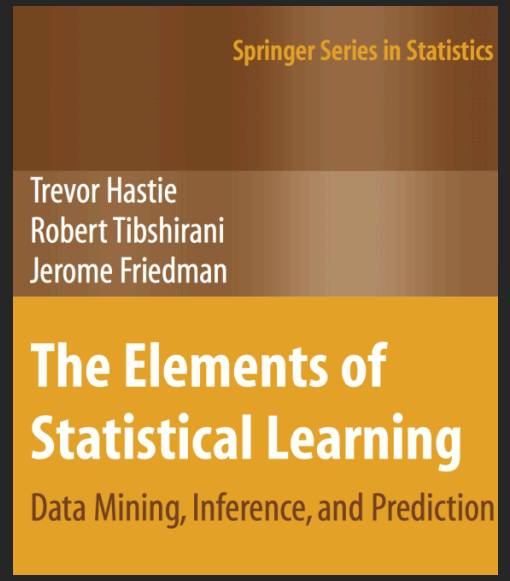
Other variants of penalized regression

- Other criteria for selection of λ
- Adaptive LASSO
- Backward selection LASSO
- Square root LASSO
- Relaxed LASSO
- Elastic net

Not evaluated in this presentation

Machine Learning





- Hyped
- Black box?
- Very flexible (too flexible?)
- Data hungry – Should not be applied to ”small datasets”
- Might be a good choice when both the number of patients and the number of potential predictors is very large

    Omics data

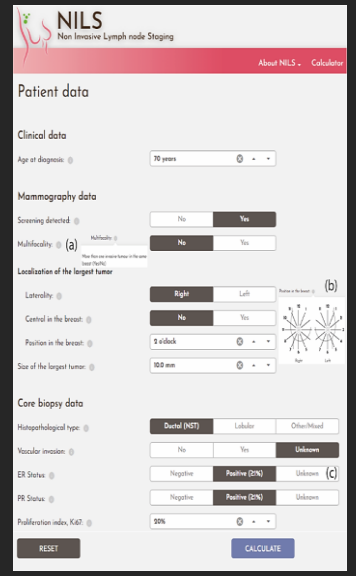
    Image data

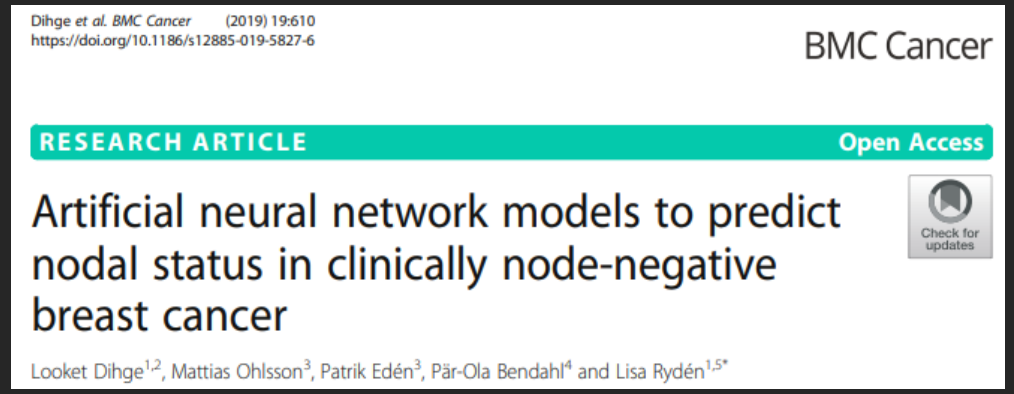
Examples of machine learning methods for prediction of a binary outcome

- Logistic regression? (NILS version 1)
- Classification/Decision trees

  Random Forest

  Boosting (Adaboost, Gradient boost, XGBoost)
- Support Vector Machines (SVM)
- K-Nearest Neighbours (KNN)
- Naive Bayes
- Artificial Neural Networks (ANN; NILS version 2)





Random Forest

1. Draw a bootstrap sample from the dataset

   On average 37% of the original samples will not be included in a bootstrap sample

   The clever idea is to use these samples to evaluate model performance (the Out-Of-Bag (OOB) error)

2. Build a decision tree for the bootstrap sample

   But evaluate only a random subset of the variables at each split

   Default fraction in R: Square root of the number of variables rounded downwards

3. Create a forest of n decision trees by repeating the steps 1-2 above

4. Classify all OOB samples for all trees, OOB error = fraction wrongly classified

Random forest modelling of Nodal status

Example code here: random_forest_demo/random_forest_demo.R at master · StatQuest/random_forest_demo · GitHub

```
library(randomForest)
```

Easy modelling once you have data in the right format

```
model <- randomForest(N_plus ~ ., data=data, proximity=TRUE)
```

Default settings often OK, but vary the hyperparameters

- number of trees
- number of variables evaluated at each split
- minimum number of samples per terminal node
- maximum number of terminal nodes

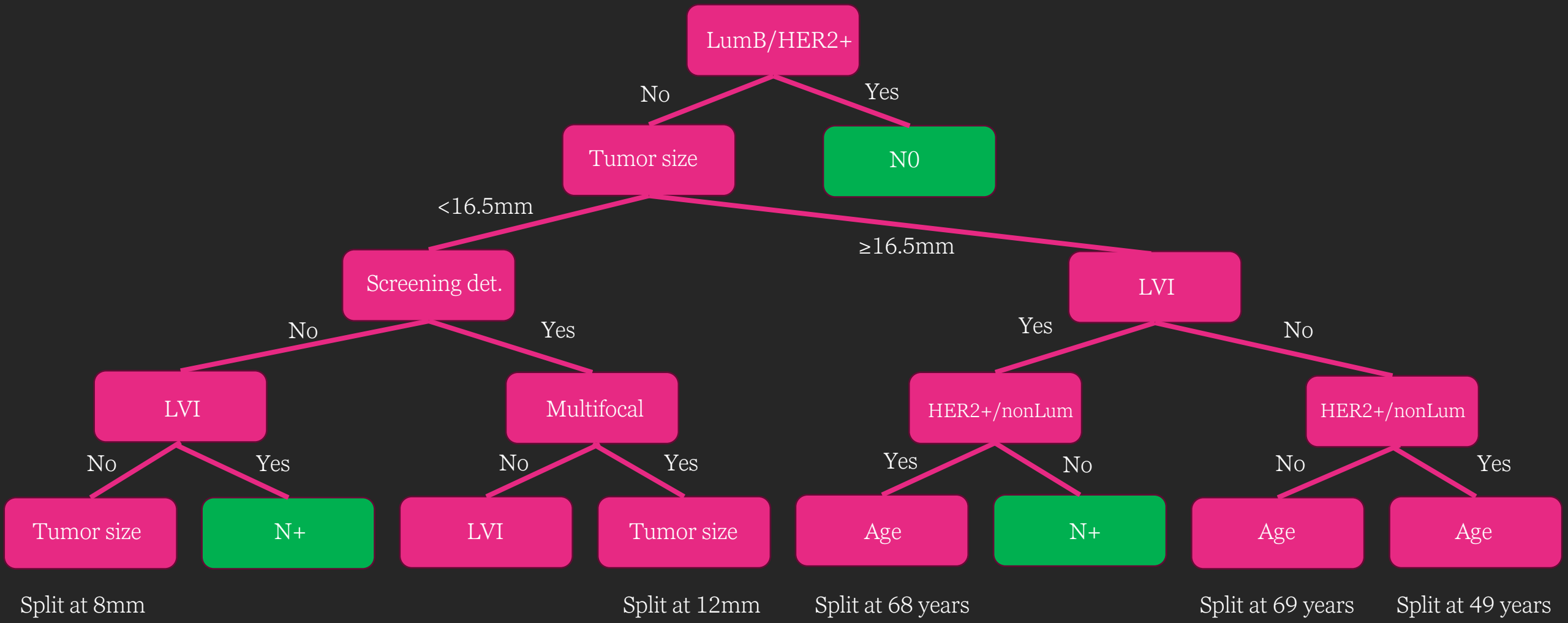
```
Call:
 randomForest(formula = N_plus ~ ., data = data, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 28.64%
Confusion matrix:
       N0 Nplus class.error
N0    252    29   0.1032028
Nplus  91    47   0.6594203
```

1-specificity

1-sensitivity

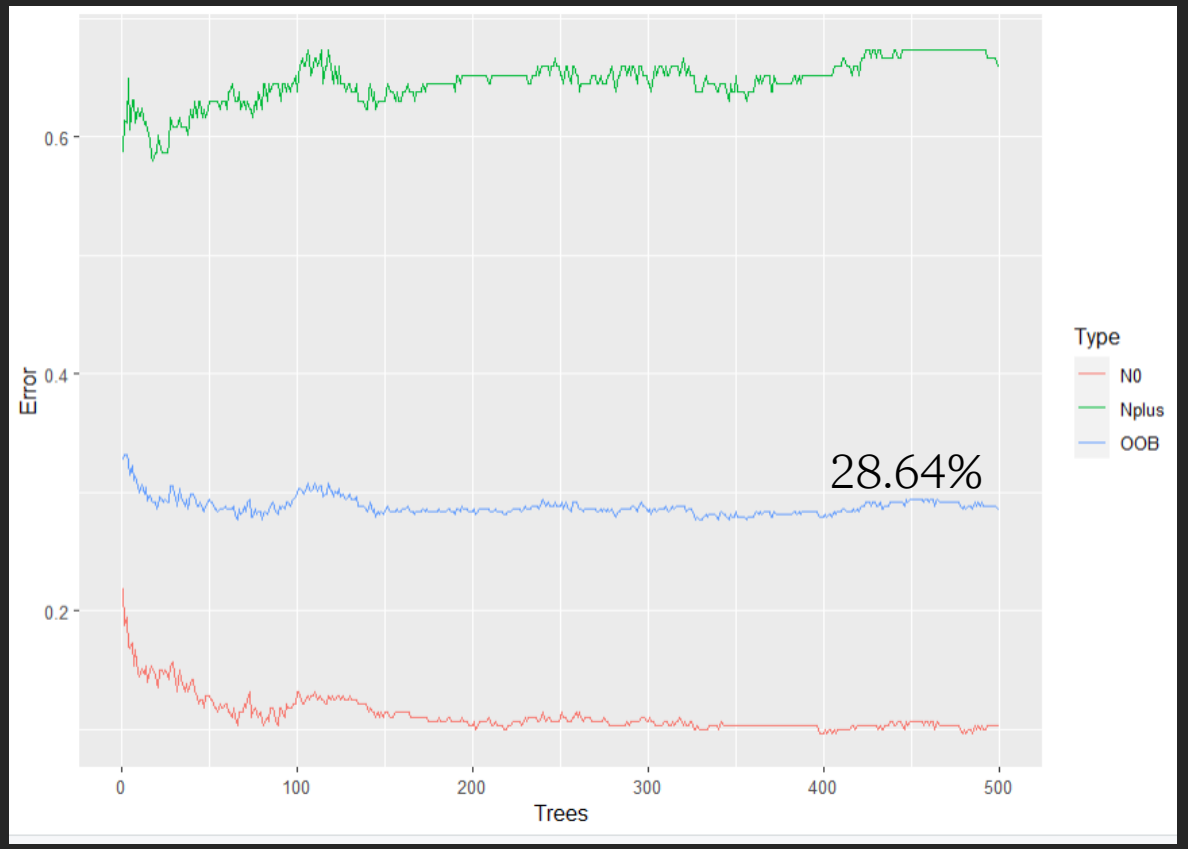
The upper part of the tree for the first bootstrap sample



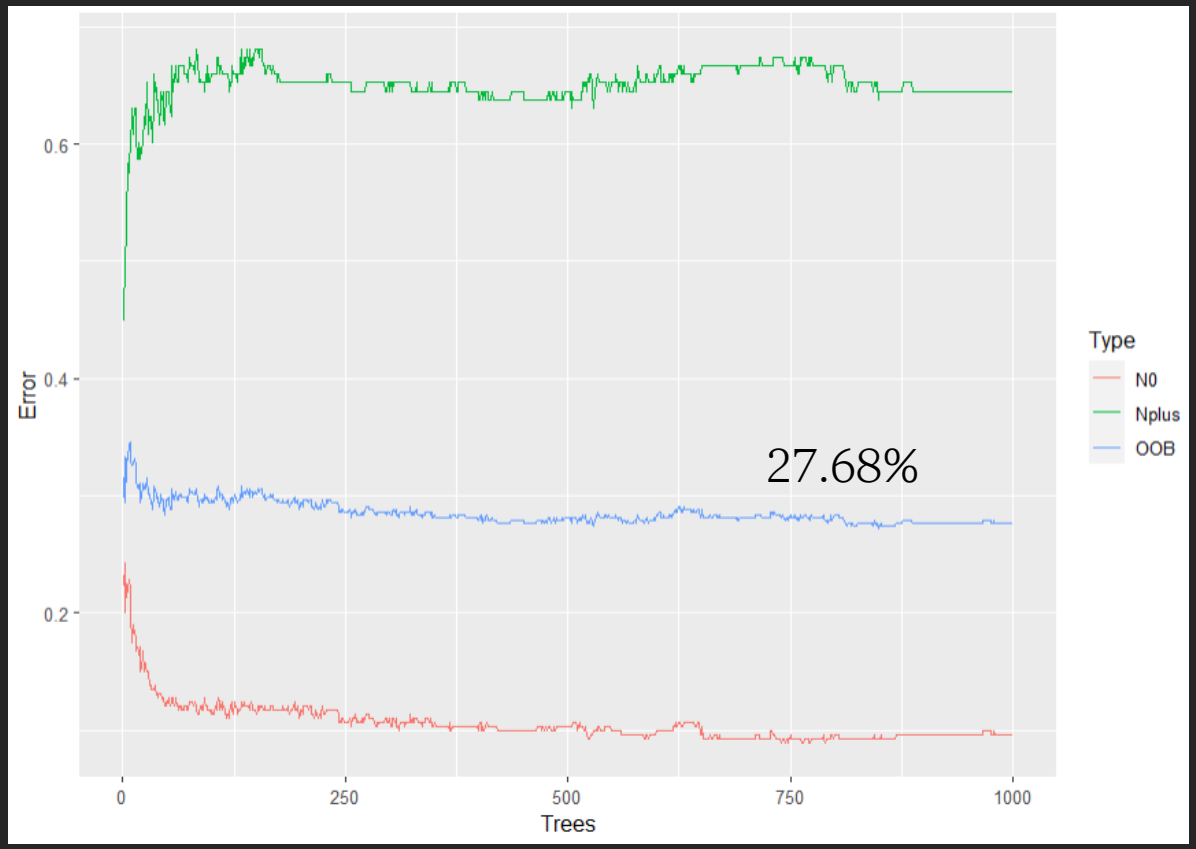
In total 111 nodes in the first tree of the forest

Out-of-bag Error

500 trees



1000 trees



Model 6: Random forest, default settings in R

- 500 trees
- 3 variables evaluated per split
- Min number of samples per terminal node = 1
- No limit on number of terminal nodes

```
Call:
 randomForest(formula = N_plus ~ ., data = data, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 28.64%
Confusion matrix:
       NO Nplus class.error
NO    252    29   0.1032028
Nplus  91    47   0.6594203
```

Model 7: Random forest, finetuned hyperparameters

- 1000 trees
- 2 variables evaluated per split
- Min number of samples per terminal node = 2
- Max 41 terminal nodes

```
Call:
 randomForest(formula = N_plus ~ ., data = data, ntree = 1000,
xnodes = 41)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 25.78%
Confusion matrix:
       NO Nplus class.error
NO    269    12  0.04270463
Nplus  96    42  0.69565217
```

Lower sensitivity but higher specificity

Validation of the random foresest models using data set aside (validation + test; 30%)

Model 6: Random forest, default settings in R

```
Call:
 randomForest(formula = N_plus ~ ., data = data, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 28.64%
Confusion matrix:
       NO Nplus class.error
NO    252    29   0.1032028
Nplus  91    47   0.6594203
```

Model 7: Random forest, finetuned hyperparameters

```
Call:
 randomForest(formula = N_plus ~ ., data = data, ntree = 1000,
xnodes = 41)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 25.78%
Confusion matrix:
       NO Nplus class.error
NO    269    12  0.04270463
Nplus  96    42  0.69565217
```

Validation+Test (n=179; 30%)

```
> set.seed(42)
> model <- readRDS(file="RF_default_dev")
> pred <- predict(model, data)
> table(data$N_plus,pred)
       pred
         NO Nplus
  NO    106    14
  Nplus  39    20
```

14+39=53 of 179 samples misclassified (29.61%)

```
> set.seed(42)
> model <- readRDS(file="RF_dev")
> pred <- predict(model, data)
> table(data$N_plus,pred)
       pred
         NO Nplus
  NO    112     8
  Nplus  41    18
```

8+41=49 of 179 samples misclassified (27.37%)

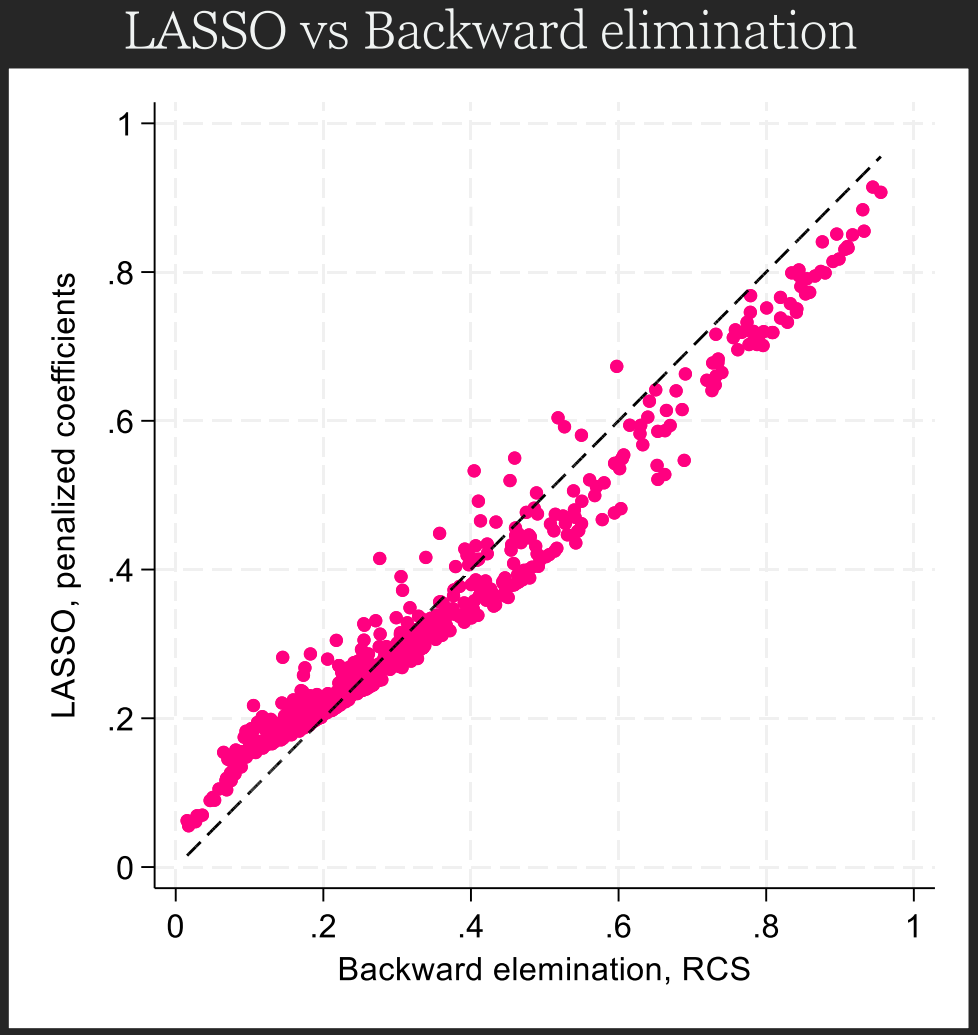
The five regression models

Model	AUC development	AUC validation+test	Drop
1. No selection	0.7505	0.7024	0.0481
2. Backward elimination	0.7533	0.7131	0.0402
3. Backward elimination + RCS	**0.7574**	**0.7158**	**0.0412**
4. Ridge, lambda selected by cv	0.7524	0.7062	0.0461
5. LASSO. Lambda selected by BIC	0.7571	0.7129	0.0442

General performance drop as measured bu AUC

Overfit?
Harder to predict N-status in validation+test?
Chance?

Predicted probabilies of N+ in validation+test



Tilted cloud

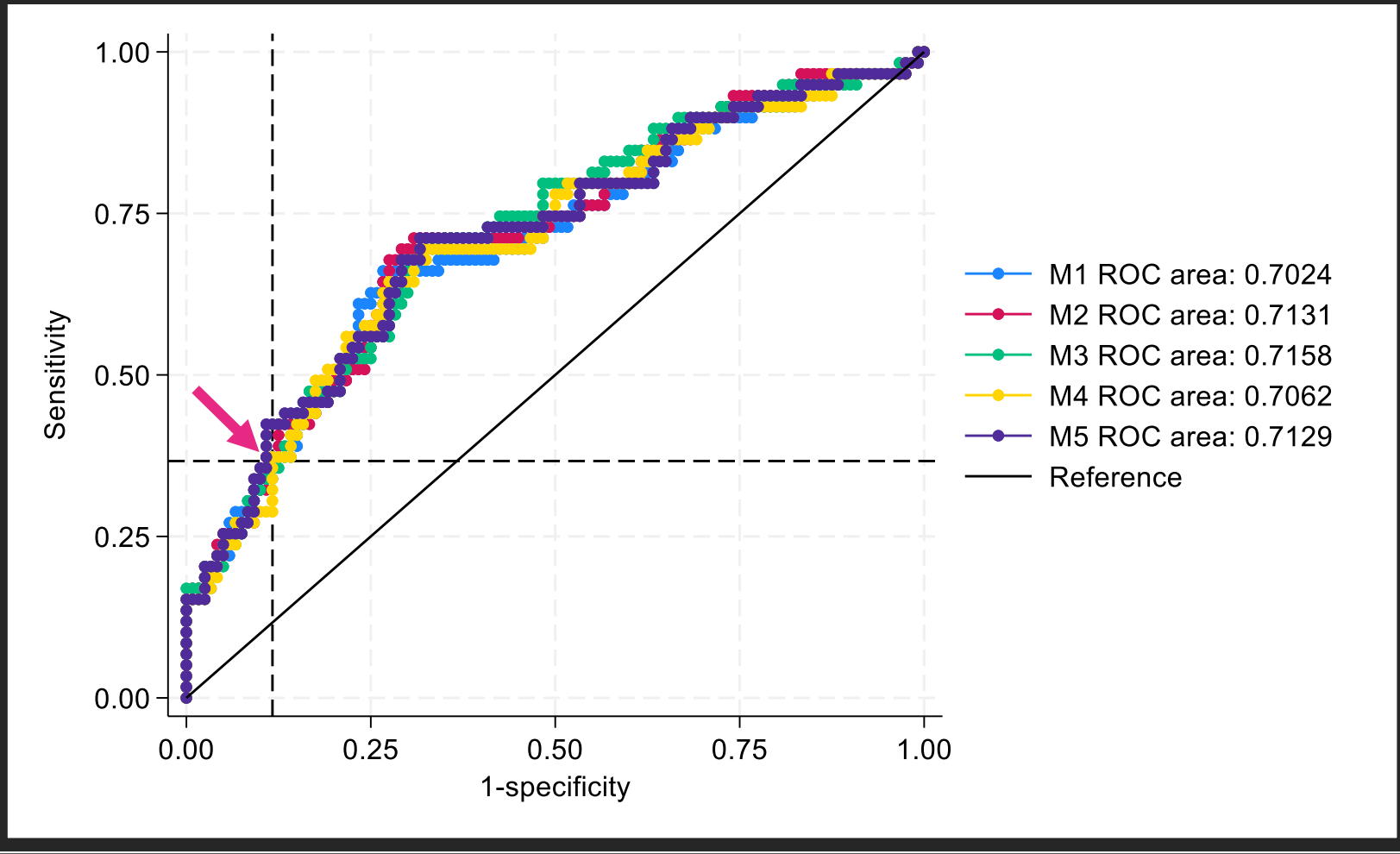
Effect of penalization (less extreme probabilities)

The five regression models after having added 50 random variables to the set of potential predictors

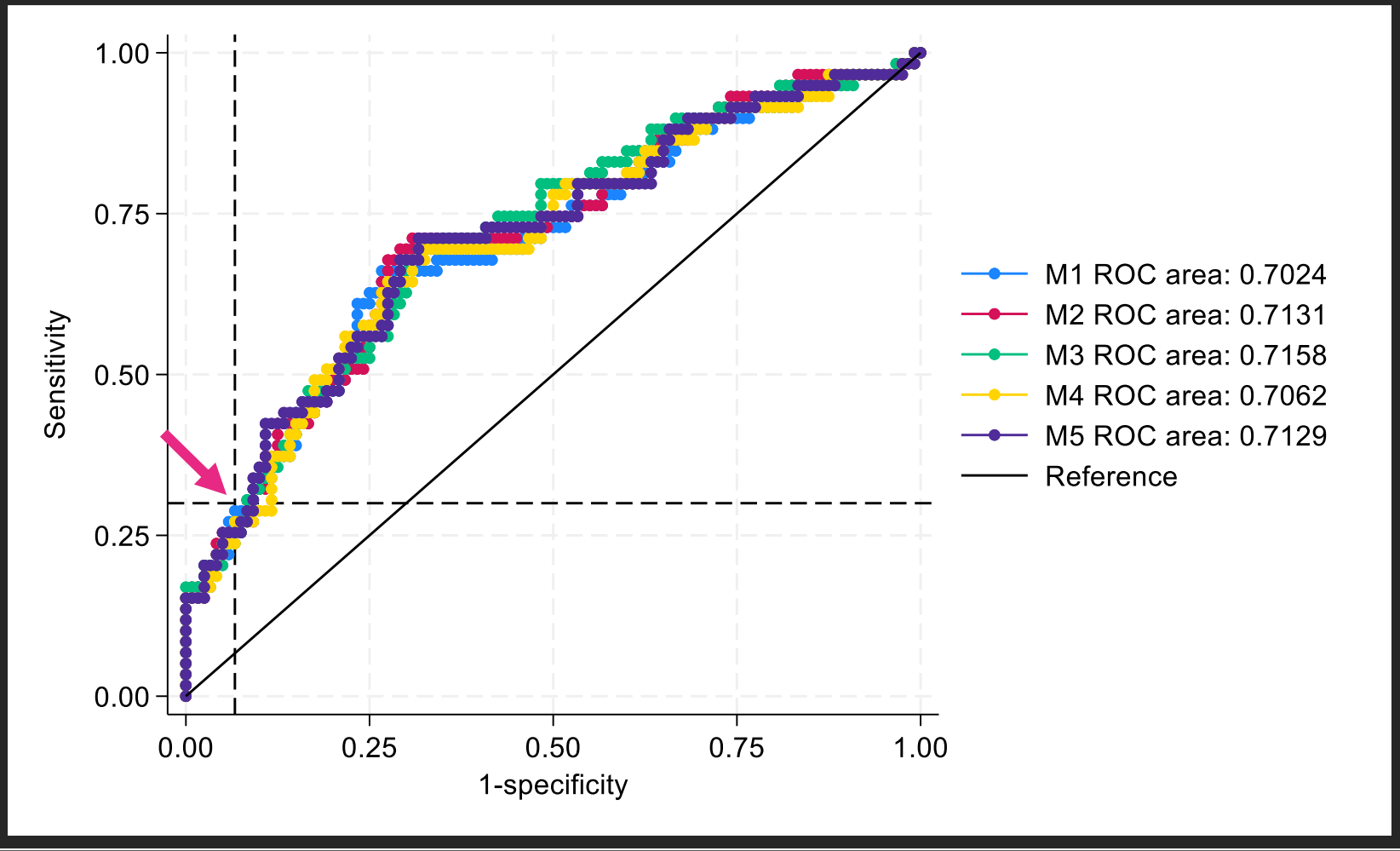
Model	AUC development	AUC validation+test	Drop	
1. No selection	0.8209	0.6592	0.1617	
2. Backward elimination	0.7959	0.6750	0.1209	8 random selected
3. Backward elimination + RCS	0.7998	0.6805	0.1192	
4. Ridge, lambda selected by cv	0.8128	0.6770	0.1408	
5. LASSO Lambda selected by BIC	**0.7352**	**0.7063**	**0.0289**	**Only 3 var selected**

Back to performance in validation+test without added random variables

Random Forest, default settings verus the five regression models – discrimination



Random Forest, finetuned hyperparameters verus the five regression models – discrimination



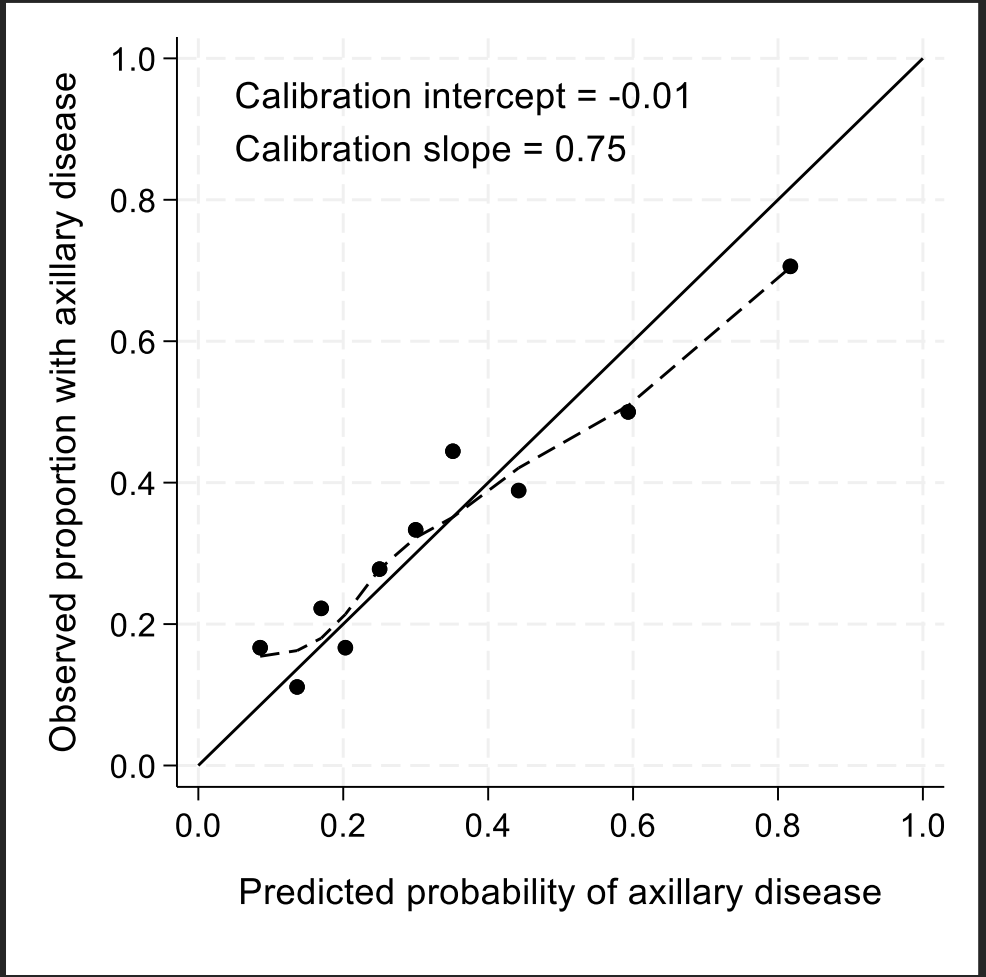
Can ROC-curves be drawn also for random forests?

Yes, based on the fraction of out-of-bag votes for N+ for each sample

Possible to draw in R?

Seems to be much easier i Python

Finally Calibration – Hosmer Lemeshow*



The backward elimination with splines for tumor size

Calibration in validation+test

* https://www.youtube.com/watch?v=KiON4m1JU14

Clinical utility

- Decison Curve Analysis (DCA)
- Health economy

Summary

Ways of minimizing the risk of overfit/overtraining

- Utilize expert knowledge to preselect relevant variables
- Adhere to the ”20 patients in the least common outcome class per parameter in the model” principle
- Collect more data (do not develop prediction models based on small datasets)
- Split the data if the sample size is large (Development, Validation, Test; e.g. 70/15/15)
- Use K-fold cross validation to finetune hyperparameters
- Bootstrap – use samples not selected to evaluate performance
- Penalized regression (Ridge regression, LASSO, Elastic Net)

Thanks!