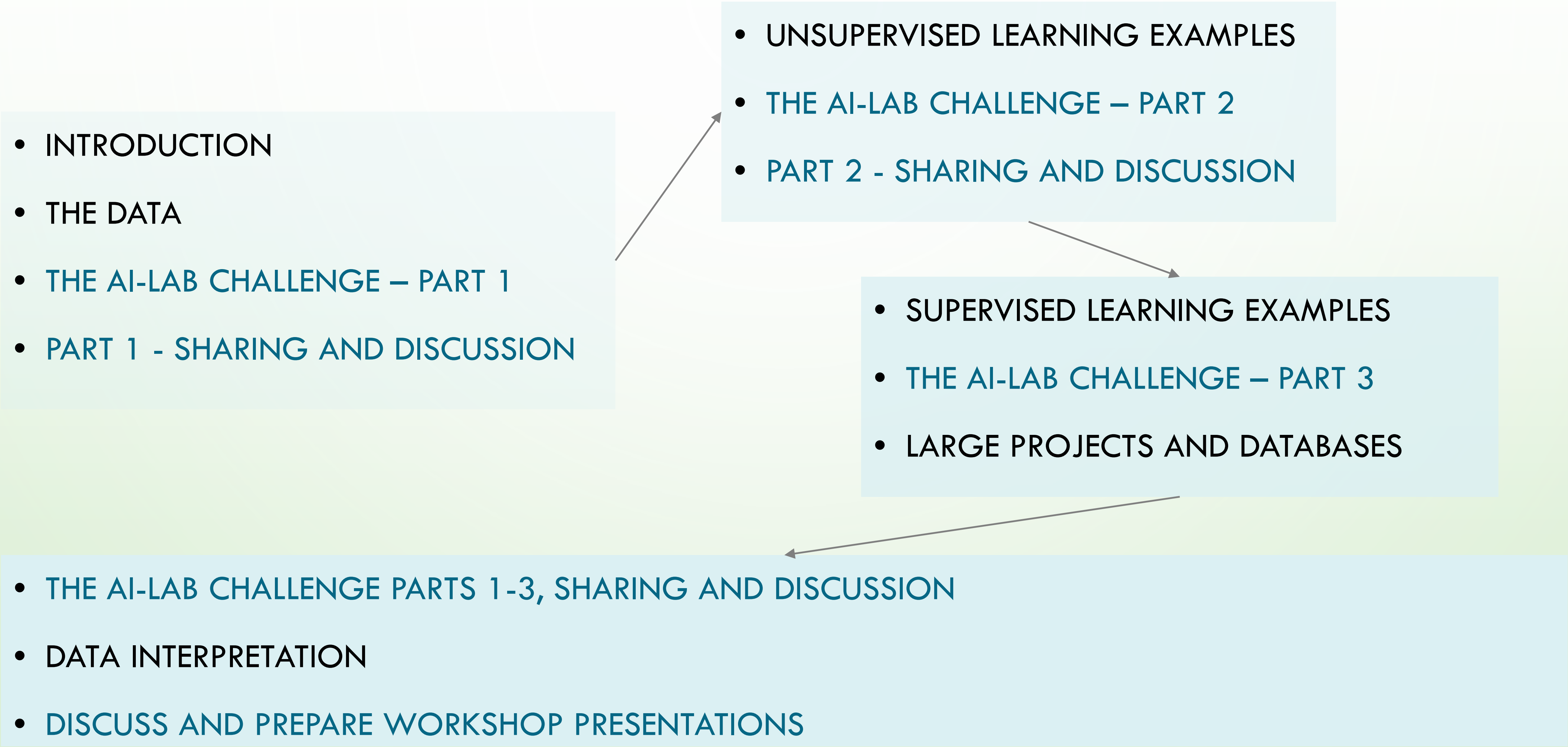
AI LAB

3.1 - SUPERVISED LEARNING

FRANCESCA M. BUFFA

LAB STRUCTURE



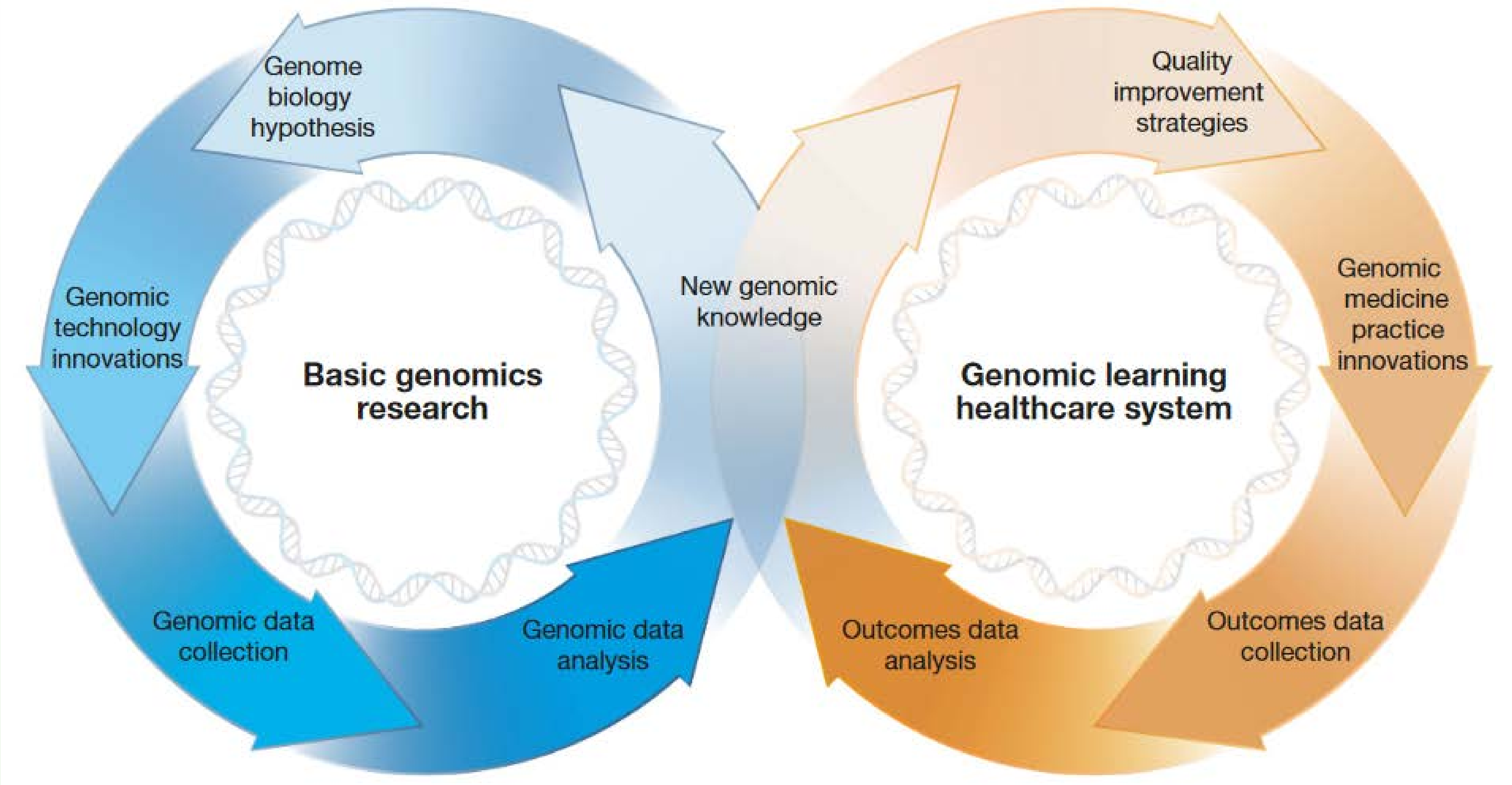
- INTRODUCTION
- THE DATA
- THE AI-LAB CHALLENGE – PART 1
- PART 1 - SHARING AND DISCUSSION

- UNSUPERVISED LEARNING EXAMPLES
- THE AI-LAB CHALLENGE – PART 2
- PART 2 - SHARING AND DISCUSSION

- SUPERVISED LEARNING EXAMPLES
- THE AI-LAB CHALLENGE – PART 3
- LARGE PROJECTS AND DATABASES

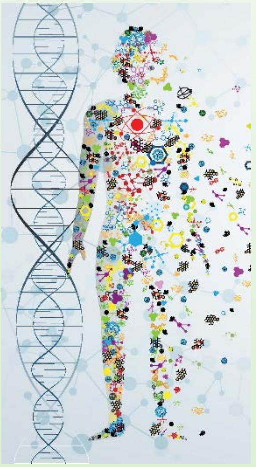
- THE AI-LAB CHALLENGE PARTS 1-3, SHARING AND DISCUSSION
- DATA INTERPRETATION
- DISCUSS AND PREPARE WORKSHOP PRESENTATIONS

VIRTUOUS CYCLES IN HUMAN GENOMICS RESEARCH AND CLINICAL CARE


Genome biology hypothesis
Genomic technology innovations
Basic genomics research
Genomic data collection
Genomic data analysis
New genomic knowledge
Quality improvement strategies
Genomic medicine practice innovations
Genomic learning healthcare system
Outcomes data collection
Outcomes data analysis

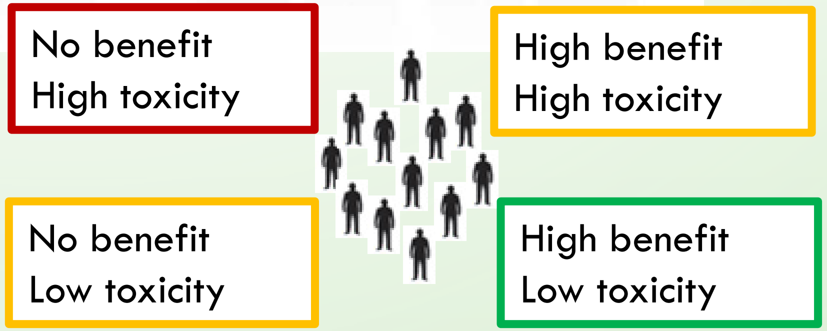
https://www.nature.com/articles/s41586-020-2817-4

Precision medicine paradigm



Biomarkers

Machine Learning



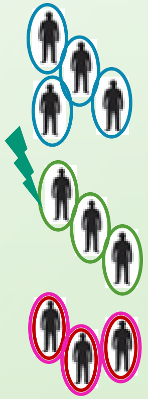
No benefit
High toxicity

High benefit
High toxicity

No benefit
Low toxicity

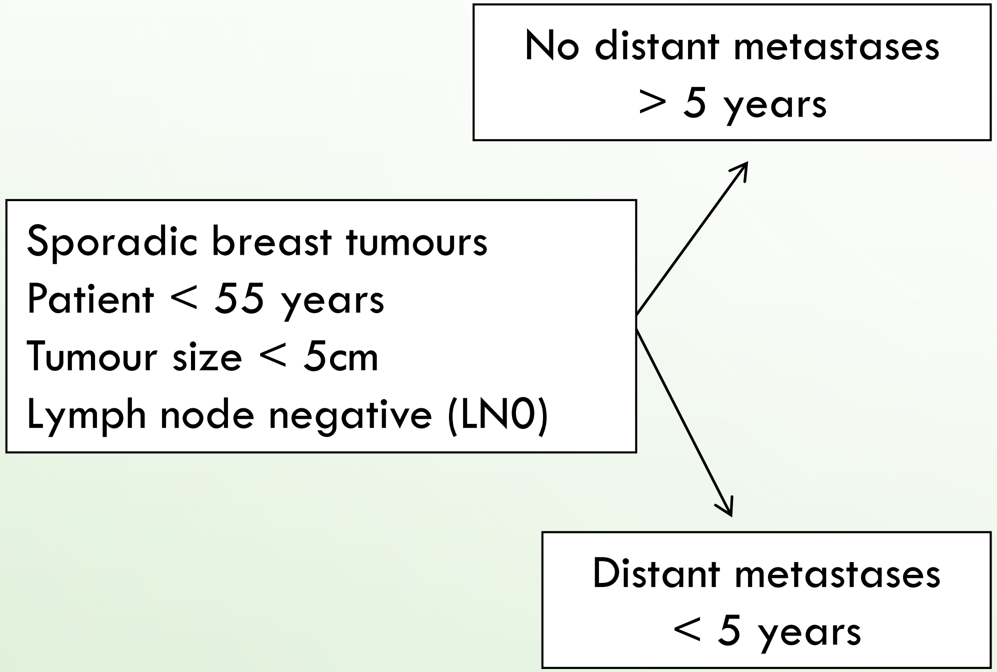
High benefit
Low toxicity

Prediction



Treatment

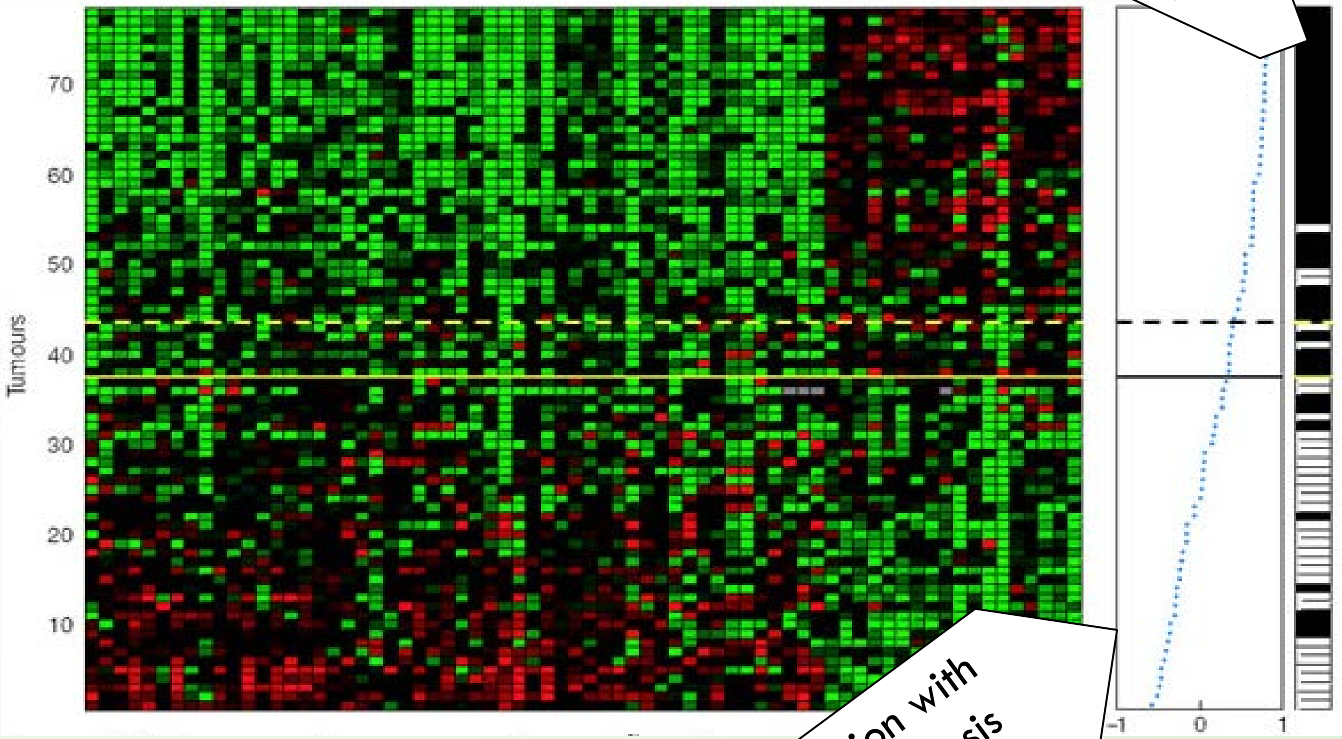
A 70-GENE SIGNATURE FOR RISK OF METASTASIS



Sporadic breast tumours
Patient < 55 years
Tumour size < 5cm
Lymph node negative (LN0)

No distant metastases
> 5 years

Distant metastases
< 5 years



Metastases

Correlation with good prognosis profile

Van 't Veer et al, Nature, 2002

A 70-GENE SIGNATURE FOR RISK OF METASTASIS

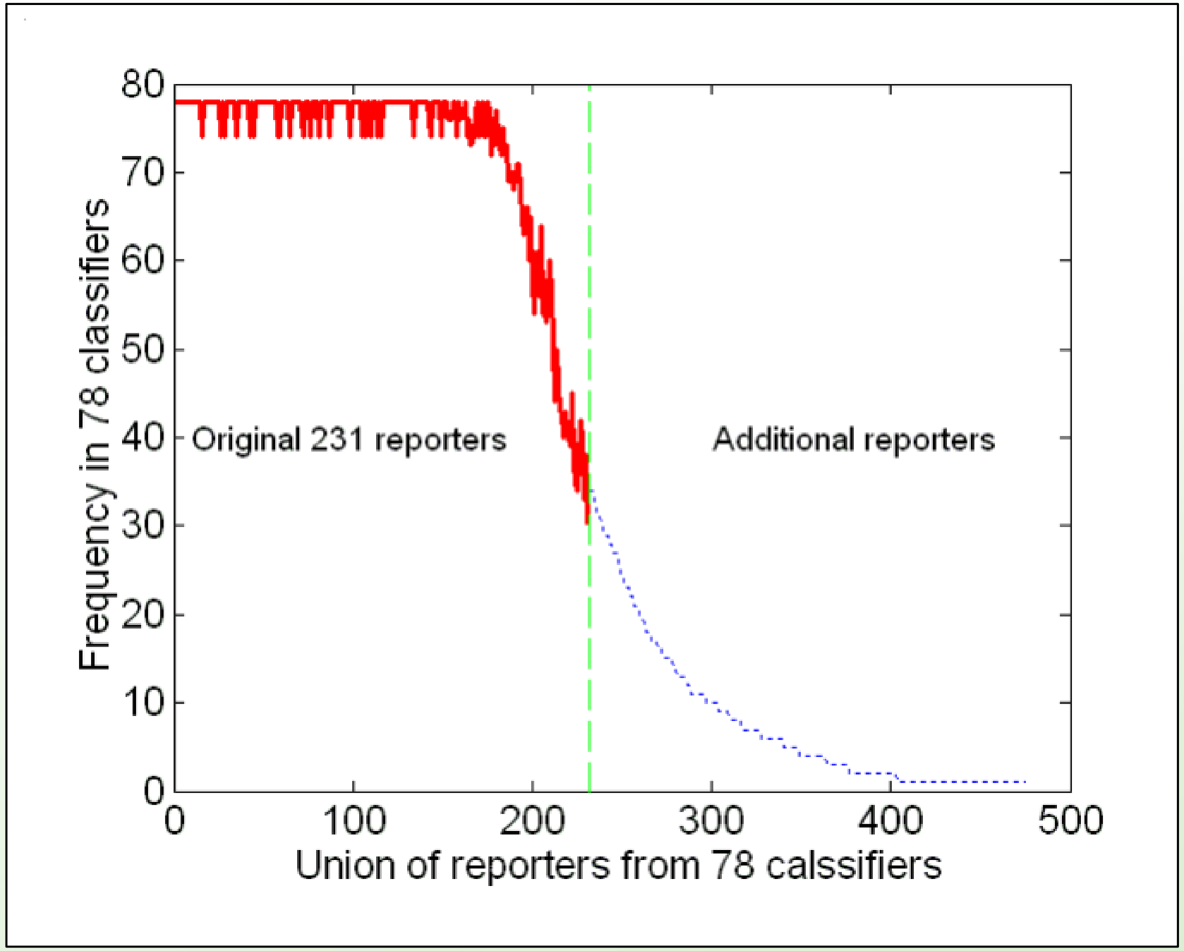
Correlation between the prognostic category (metastasis vs. no-metastasis) and the logarithmic expression ratio across all 78 samples for each individual gene in 5,000 significantly expressed genes.

Permutation to evaluate significance.

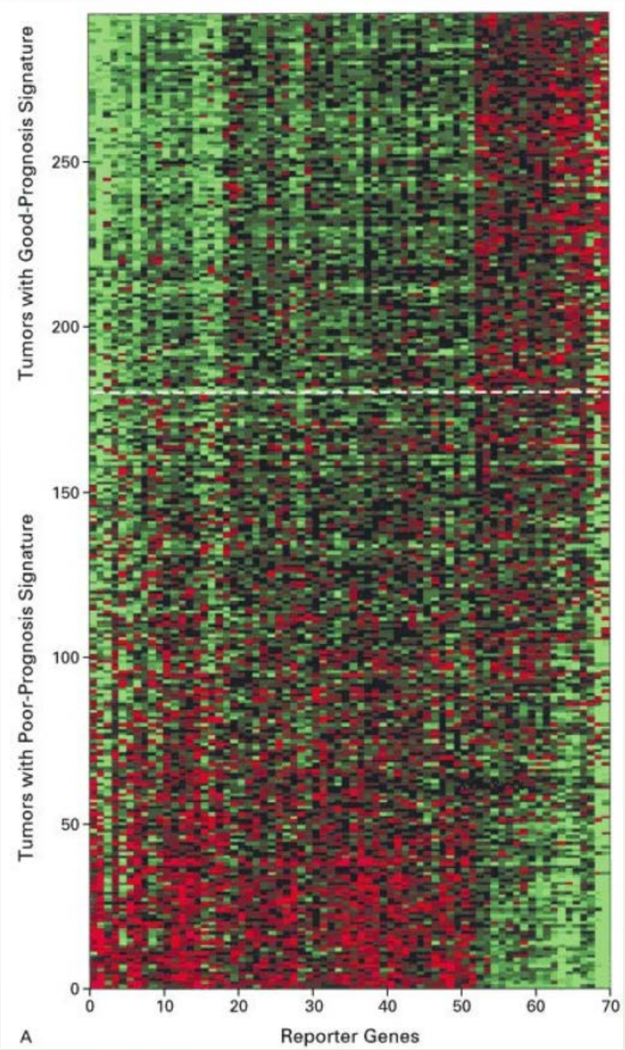
231 genes corr coeff > 0.3 or < – 0.3 selected (p= 0.3%)

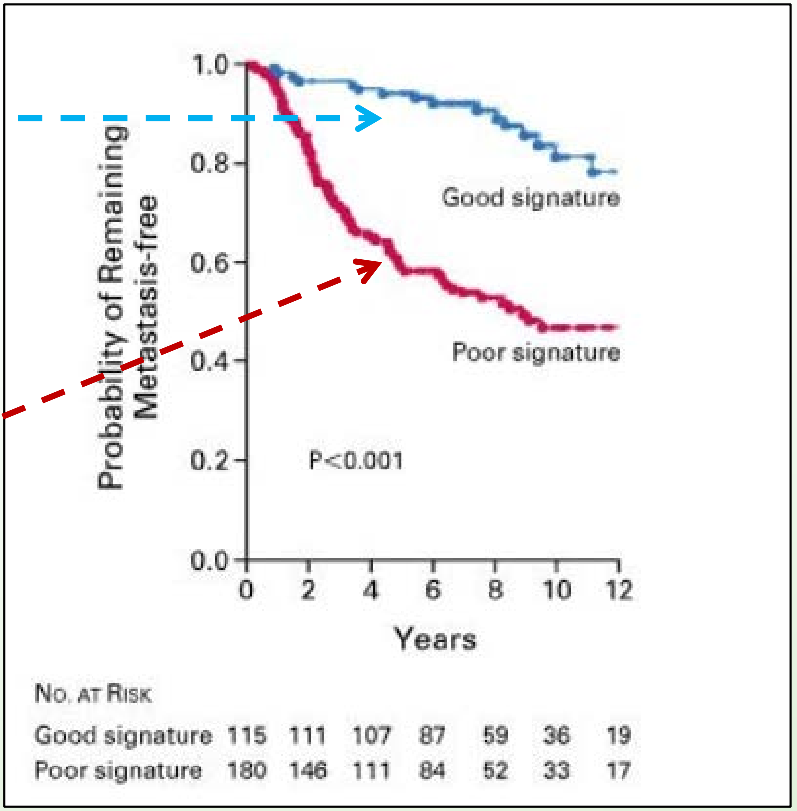
Leave-one-out: (1) leave one sample out, (2) define reporters based on the remaining 77 samples among the set of ~5000 significant genes, (3) use the reporters to predict the outcome of the one sample that was left out in step (1), (4) repeat steps (1)-(3) exhaustively for all 78 samples.

Select top 70



70-GENE SIGNATURE IS PREDICTIVE OF SURVIVAL AND RISK OF METASTASES [VAN DE VIJVER ET AL, N ENGL J MED, 2002]

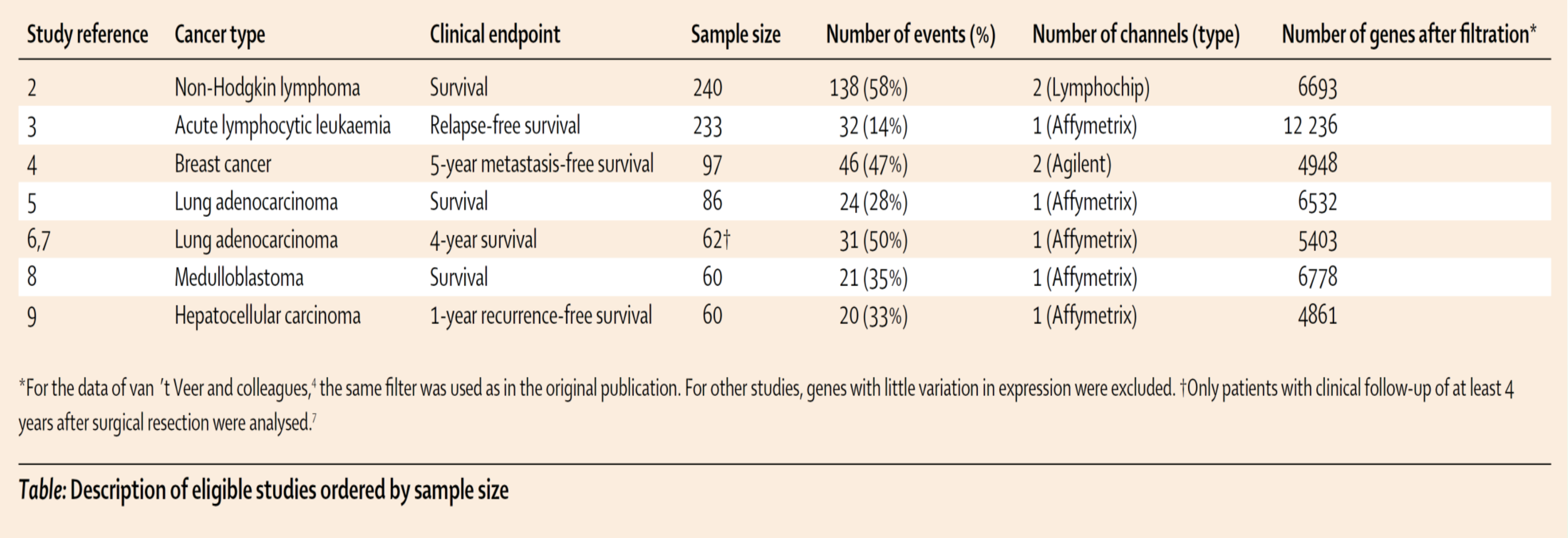




No. at Risk

Good signature	115	111	107	87	59	36	19
Poor signature	180	146	111	84	52	33	17

Performance of gene expression signatures in published studies



Study reference	Cancer type	Clinical endpoint	Sample size	Number of events (%)	Number of channels (type)	Number of genes after filtration*
2	Non-Hodgkin lymphoma	Survival	240	138 (58%)	2 (Lymphochip)	6693
3	Acute lymphocytic leukaemia	Relapse-free survival	233	32 (14%)	1 (Affymetrix)	12 236
4	Breast cancer	5-year metastasis-free survival	97	46 (47%)	2 (Agilent)	4948
5	Lung adenocarcinoma	Survival	86	24 (28%)	1 (Affymetrix)	6532
6,7	Lung adenocarcinoma	4-year survival	62†	31 (50%)	1 (Affymetrix)	5403
8	Medulloblastoma	Survival	60	21 (35%)	1 (Affymetrix)	6778
9	Hepatocellular carcinoma	1-year recurrence-free survival	60	20 (33%)	1 (Affymetrix)	4861

*For the data of van 't Veer and colleagues,[4] the same filter was used as in the original publication. For other studies, genes with little variation in expression were excluded. †Only patients with clinical follow-up of at least 4 years after surgical resection were analysed.[7]

Table: **Description of eligible studies ordered by sample size**

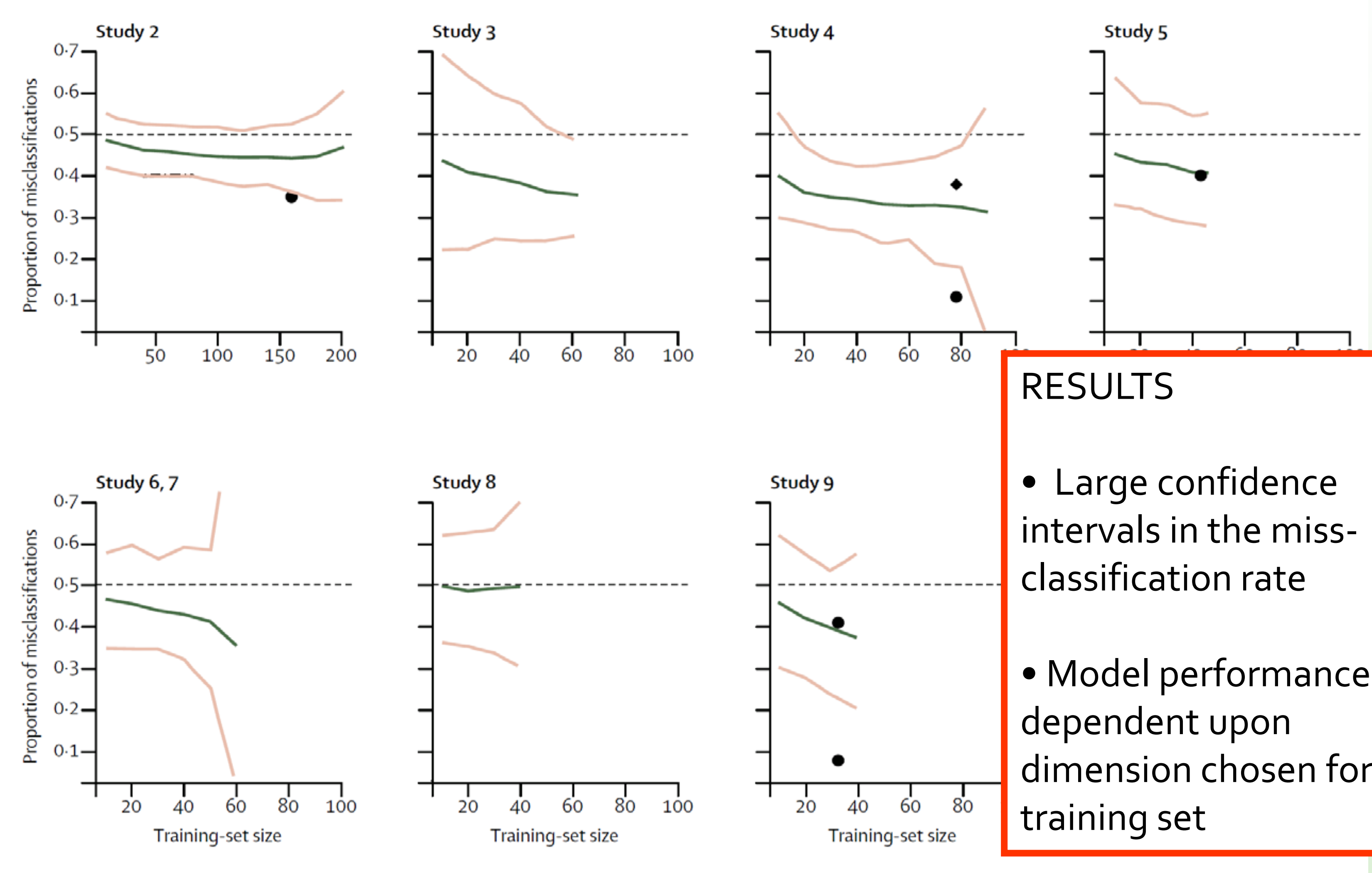
Michiels et al, Lancet 2005

Resampling strategy:

1) Sample a set of patients for training at random, and leave the rest for validation
2) Identify a gene expression signature (GES) for the clinical endpoint in the training set
3) Predict for the patients in validation set and estimate the proportion of misclassifications
4) Iterate on multiple random sets to study stability and performance of the GES

Performance of gene expression signatures in published studies

Misclassification rate from 500 random training-validation sets vs. training-set size (mean and 95% CIs)



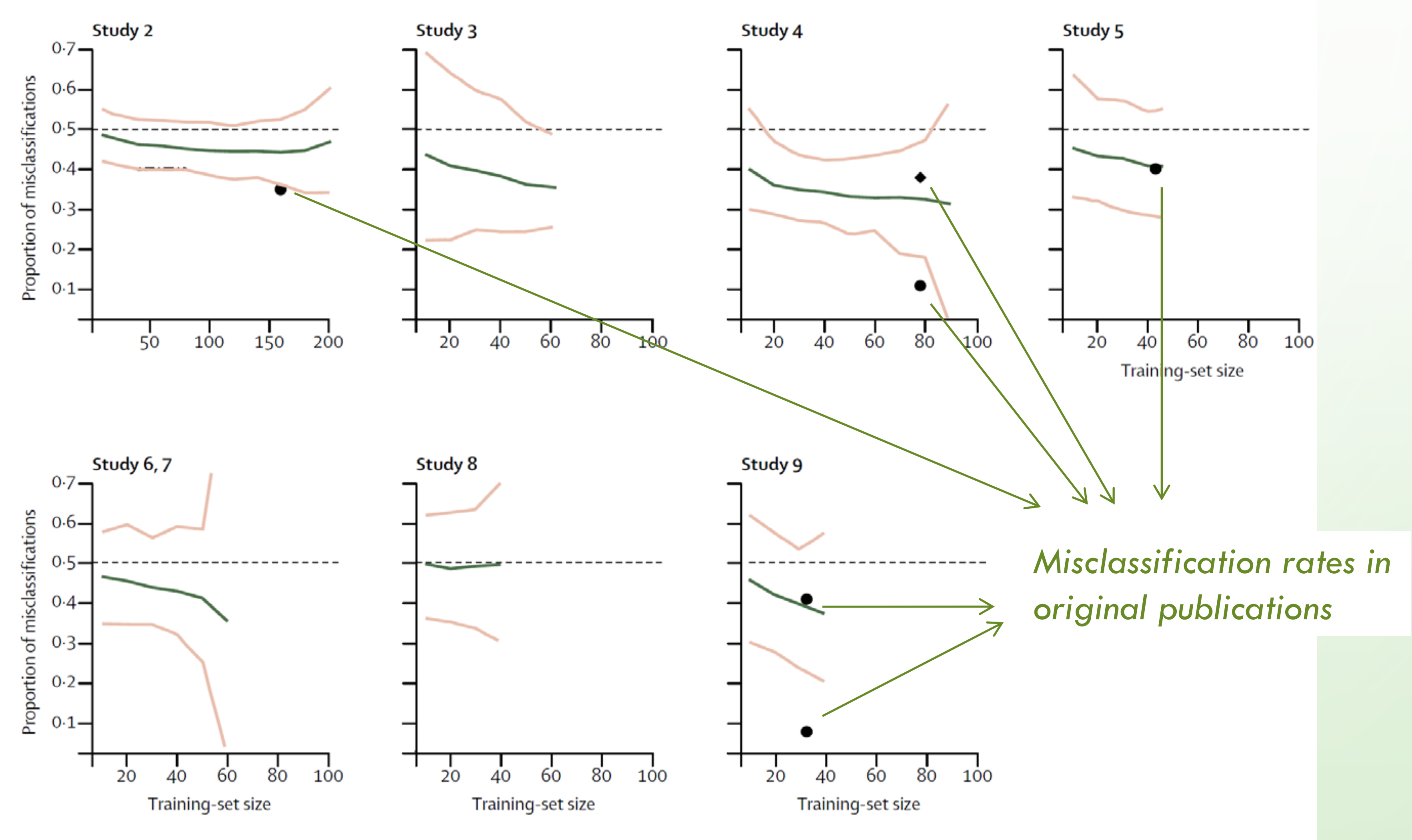
RESULTS

- Large confidence intervals in the miss-classification rate

- Model performance dependent upon dimension chosen for the training set

Performance of gene expression signatures in published studies

Misclassification rate from 500 random training-validation sets vs. training-set size (mean and 95% CIs)



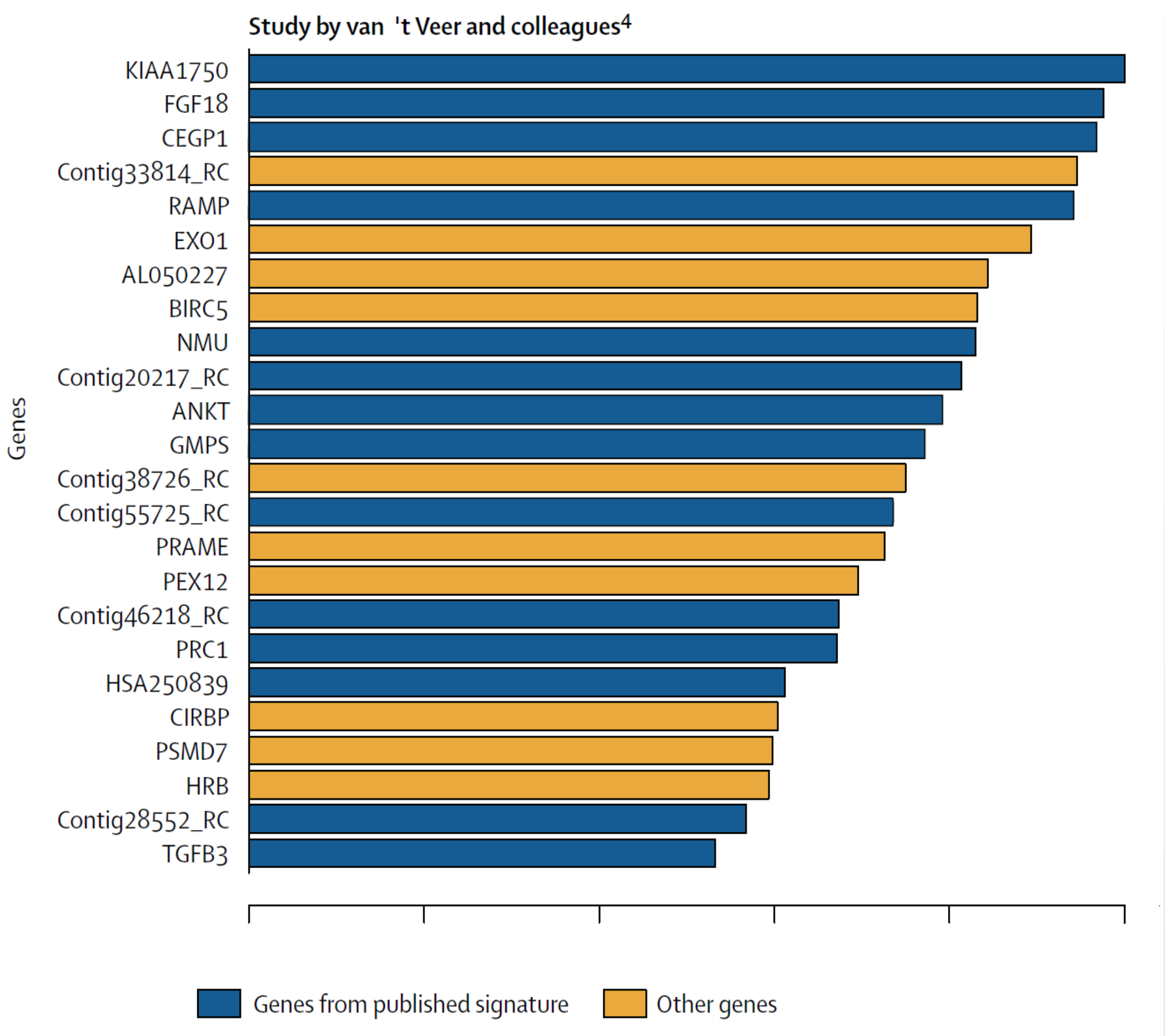
Misclassification rates in original publications

LARGE VARIABILITY IN SELECTED GENES

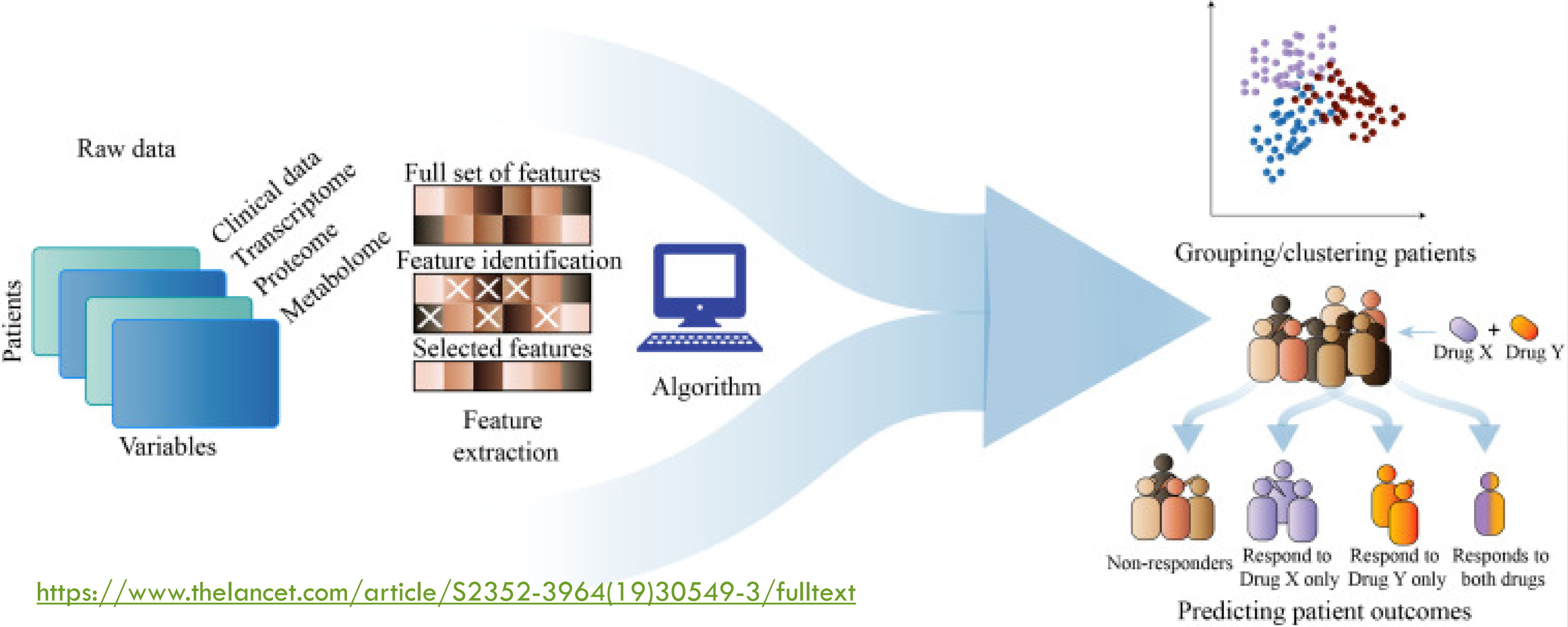
- *Several possible models with similar correlation with outcome*

- *Very little overlap in gene content between models*

24 genes within the common set were selected in at least 50% of the random simulations

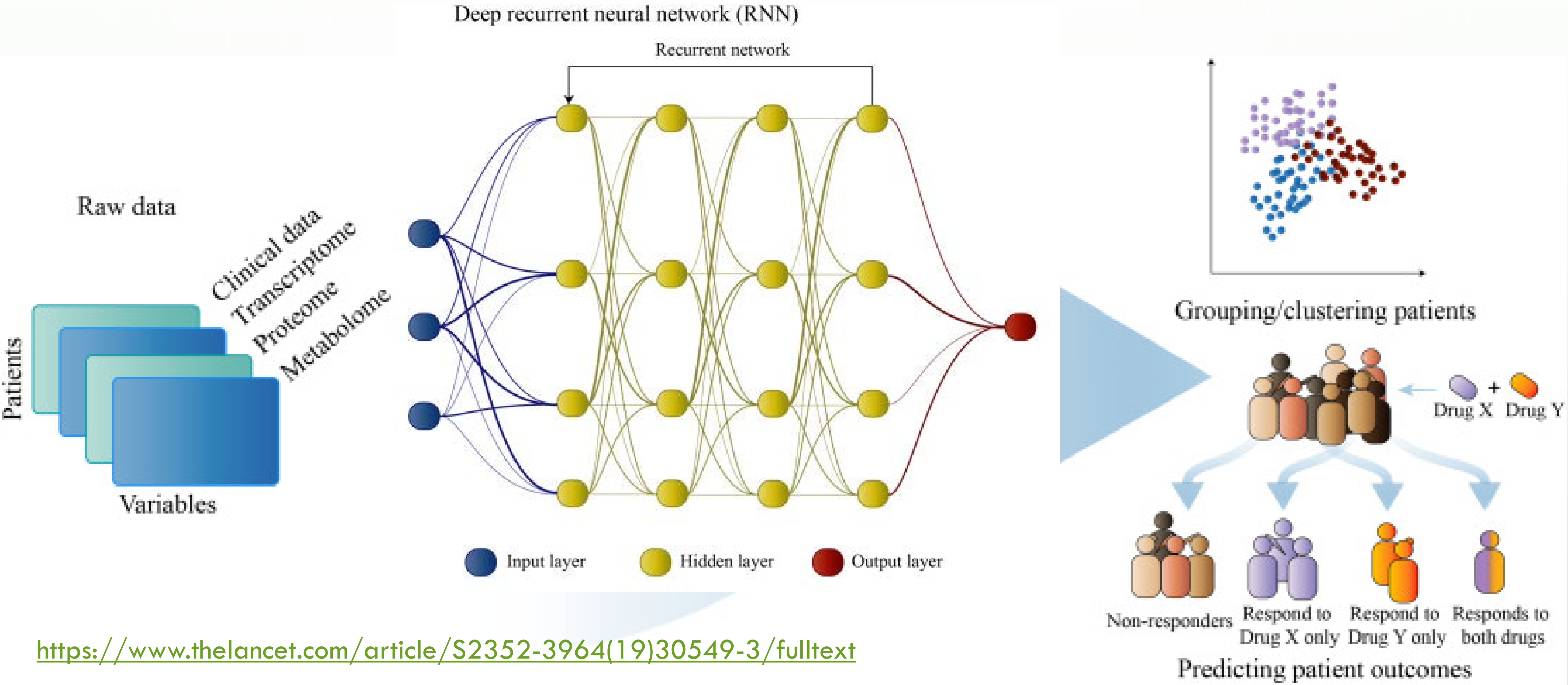


(DEEP) LEARNING TO IDENTIFY PREDICTIVE BIOMARKERS



https://www.thelancet.com/article/S2352-3964(19)30549-3/fulltext

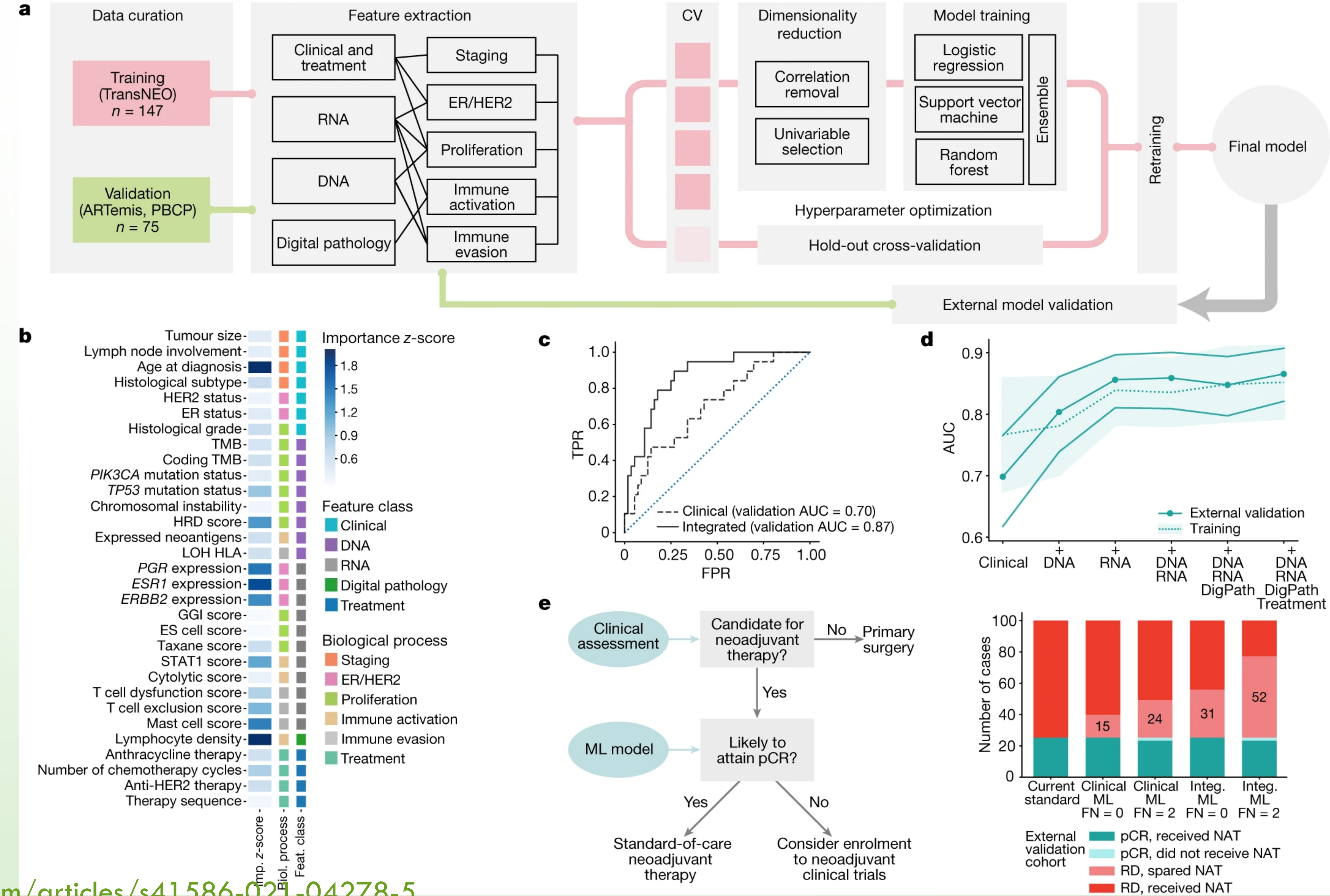
(DEEP) LEARNING TO IDENTIFY PREDICTIVE BIOMARKERS



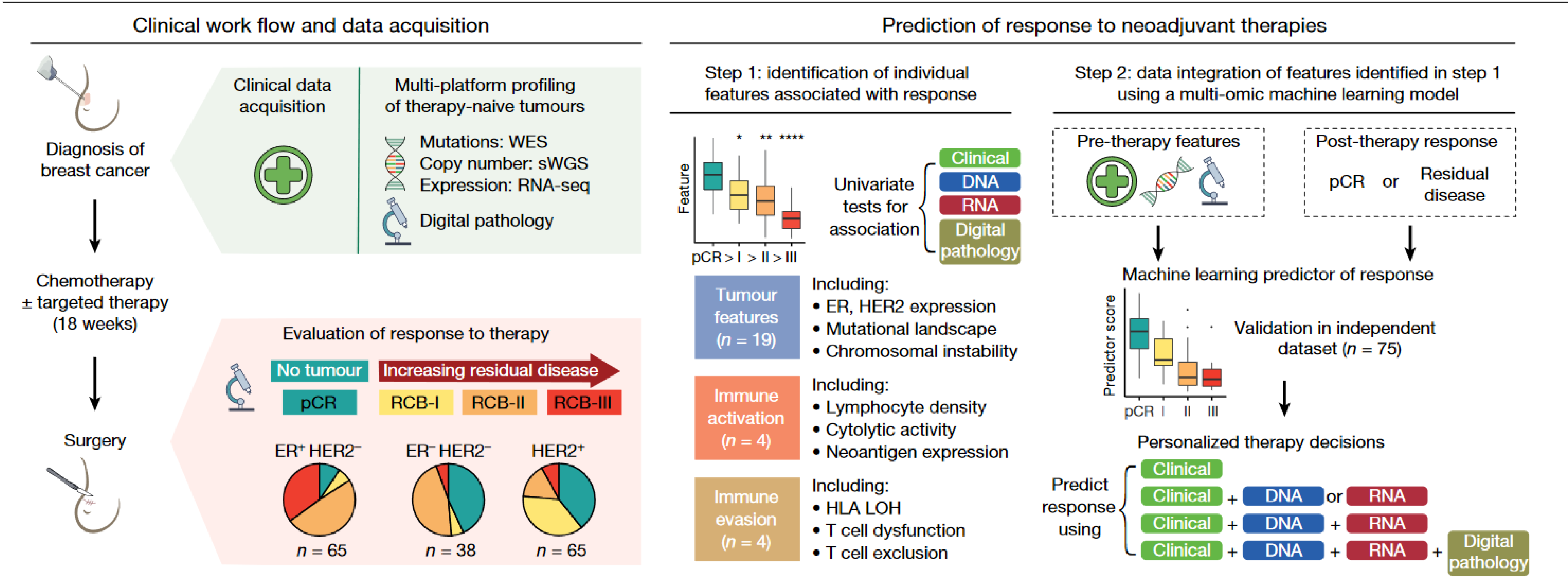
https://www.thelancet.com/article/S2352-3964(19)30549-3/fulltext

Multi-omic machine learning predictor of breast cancer therapy response

Nature, 623–629 (2022)



https://www.nature.com/articles/s41586-021-04278-5

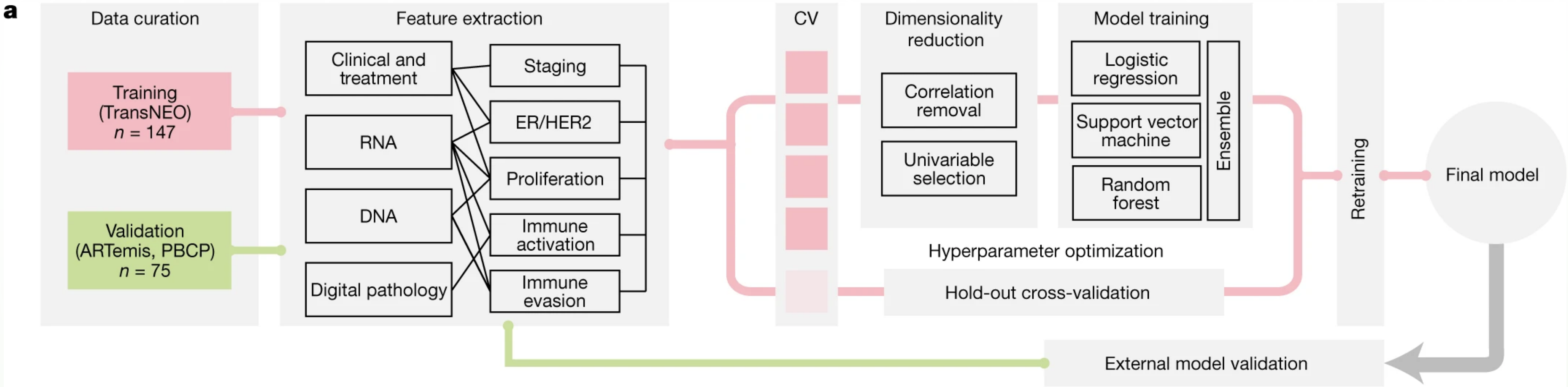

Clinical work flow and data acquisition
Diagnosis of breast cancer
Chemotherapy ± targeted therapy (18 weeks)
Surgery
Clinical data acquisition
Multi-platform profiling of therapy-naive tumours
Mutations: WES
Copy number: sWGS
Expression: RNA-seq
Digital pathology
Evaluation of response to therapy
No tumour
Increasing residual disease
pCR
RCB-I
RCB-II
RCB-III
ER+ HER2−
ER− HER2−
HER2+
n = 65
n = 38
n = 65
Prediction of response to neoadjuvant therapies
Step 1: identification of individual features associated with response
Feature
* ** ****
pCR > I > II > III
Univariate tests for association
Clinical
DNA
RNA
Digital pathology
Tumour features (n = 19)
Including:
• ER, HER2 expression
• Mutational landscape
• Chromosomal instability
Immune activation (n = 4)
Including:
• Lymphocyte density
• Cytolytic activity
• Neoantigen expression
Immune evasion (n = 4)
Including:
• HLA LOH
• T cell dysfunction
• T cell exclusion
Step 2: data integration of features identified in step 1 using a multi-omic machine learning model
Pre-therapy features
Post-therapy response
pCR or Residual disease
Machine learning predictor of response
Predictor score
pCR I II III
Validation in independent dataset (n = 75)
Personalized therapy decisions
Predict response using
Clinical
Clinical + DNA or RNA
Clinical + DNA + RNA
Clinical + DNA + RNA + Digital pathology

Methods

- clinical, digital pathology, genomic and transcriptomic profiles of pre-treatment biopsies of breast tumours from 168 patients treated with chemotherapy with or without HER2 (encoded by *ERBB2*)-targeted therapy before surgery.
- Pathology end points: complete response or residual disease
- Multi-omic features in these diagnostic biopsies

Multi-omic machine learning predictor of breast cancer therapy response

Nature, 623–629 (2022)



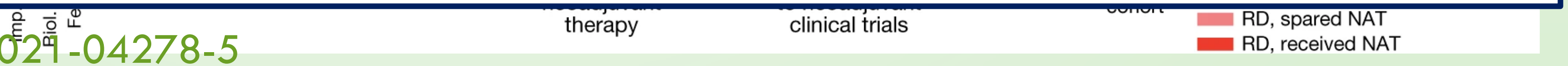
Methods

- Six pCR prediction models including different feature combinations were derived using:
  (1) clinical features only, and adding (2) DNA, (3) RNA, (4) DNA and RNA, (5) DNA, RNA and digital pathology, and (6) DNA, RNA, digital pathology and treatment.

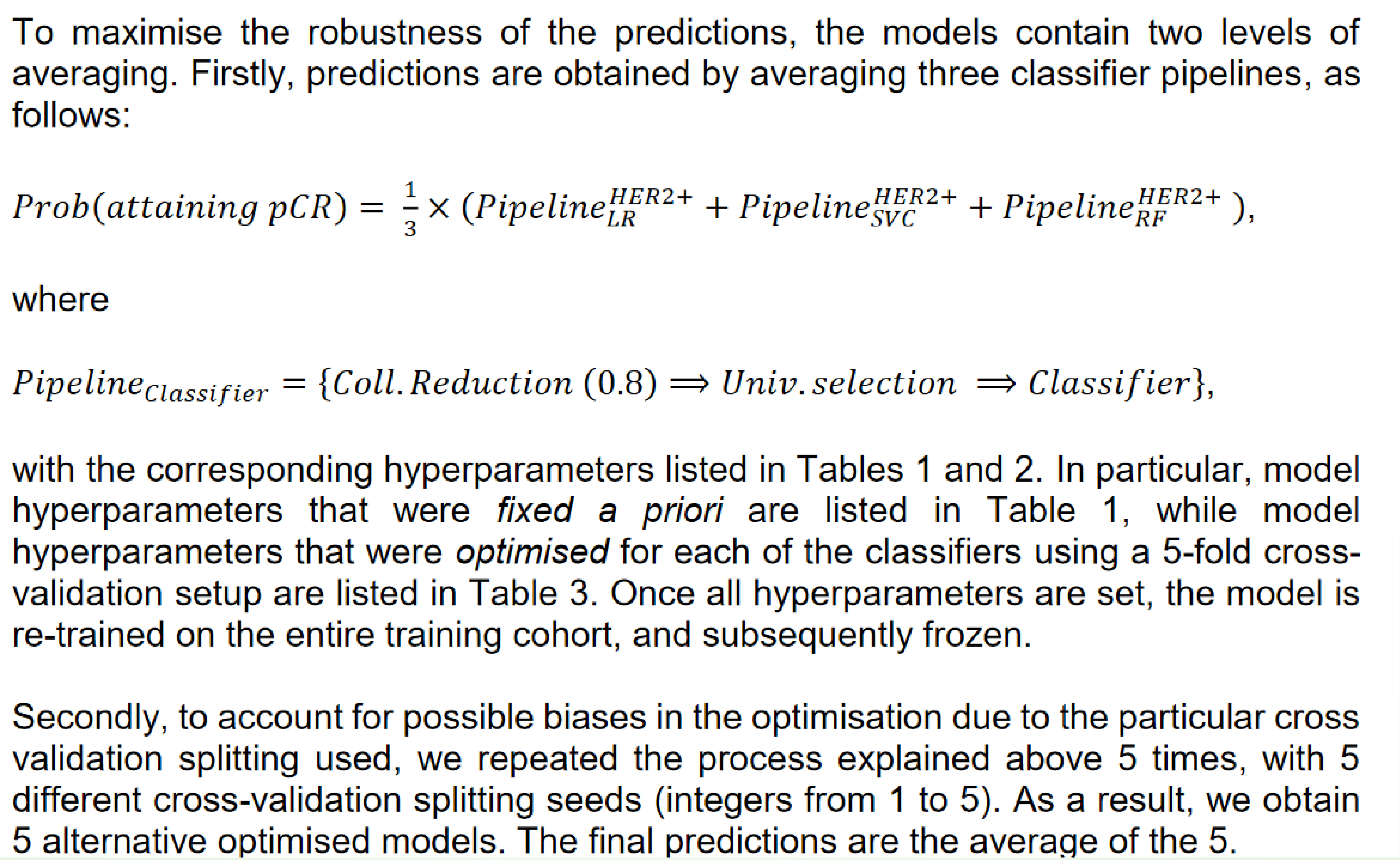
- The models were based on a multi-step predictor pipeline. Features were first filtered by univariable selection and collinearity reduction, and then fed into an unweighted ensemble classifier.

- Each ensemble consisted of three algorithms acting in parallel: logistic regression with elastic net regularization, a support vector machine and a random forest. The three algorithm scores were then averaged to form the predictor.

- A fivefold cross-validation scheme was used to optimize model hyperparameters



https://www.nature.com/articles/s41586-021-04278-5



To maximise the robustness of the predictions, the models contain two levels of averaging. Firstly, predictions are obtained by averaging three classifier pipelines, as follows:

$$Prob(attaining\ pCR) = \frac{1}{3} \times (Pipeline_{LR}^{HER2+} + Pipeline_{SVC}^{HER2+} + Pipeline_{RF}^{HER2+}),$$

where

$$Pipeline_{Classifier} = \{Coll.Reduction\ (0.8) \Longrightarrow Univ.selection\ \Longrightarrow Classifier\},$$

with the corresponding hyperparameters listed in Tables 1 and 2. In particular, model hyperparameters that were *fixed a priori* are listed in Table 1, while model hyperparameters that were *optimised* for each of the classifiers using a 5-fold cross-validation setup are listed in Table 3. Once all hyperparameters are set, the model is re-trained on the entire training cohort, and subsequently frozen.

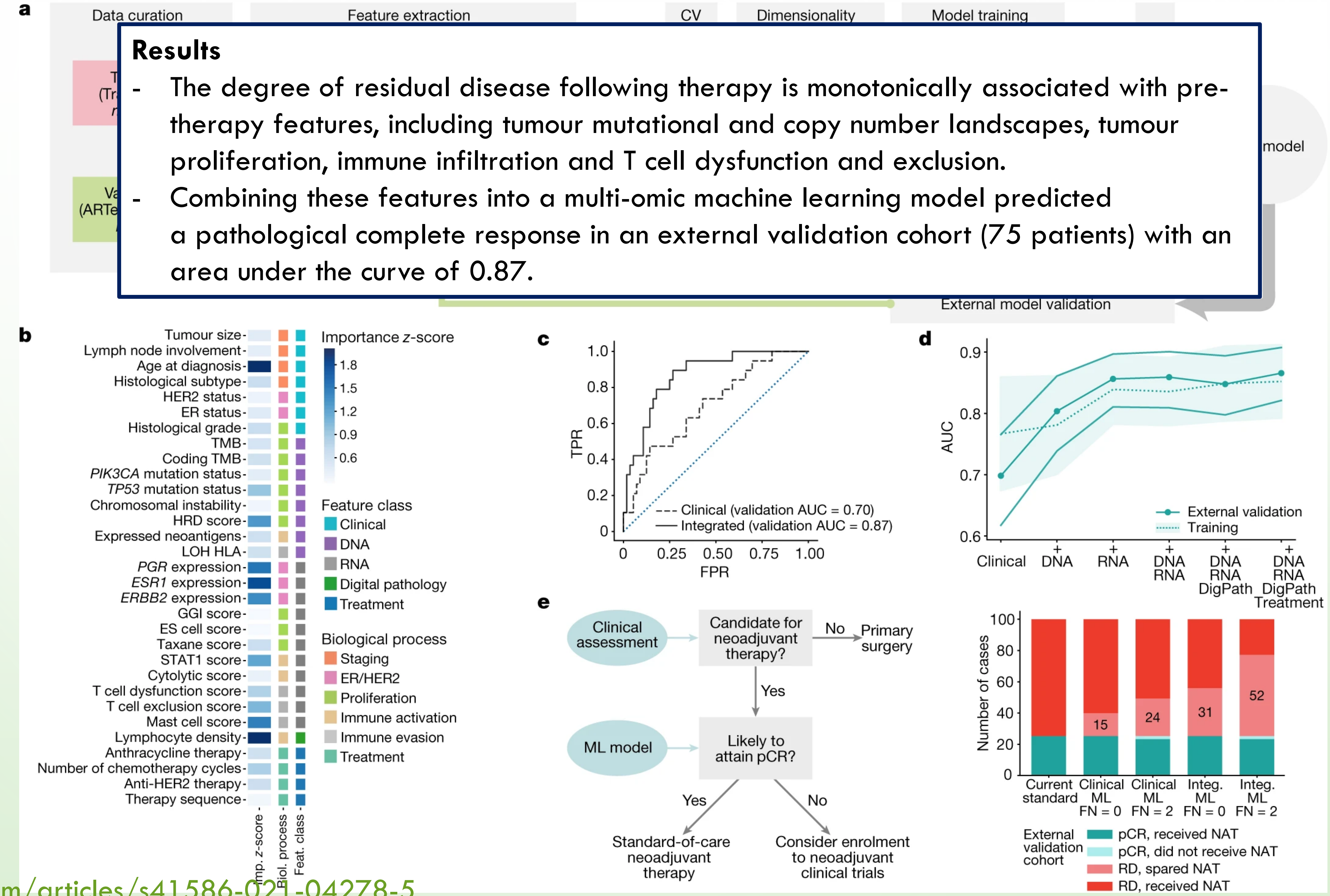
Secondly, to account for possible biases in the optimisation due to the particular cross validation splitting used, we repeated the process explained above 5 times, with 5 different cross-validation splitting seeds (integers from 1 to 5). As a result, we obtain 5 alternative optimised models. The final predictions are the average of the 5.

From Suppl Methods: https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-021-04278-5/MediaObjects/41586_2021_4278_MOESM1_ESM.pdf

Code used: https://github.com/cclab-brca/neoadjuvant-therapy-response-predictor/blob/master/R/06%20-%20ML%20predictor.R

Multi-omic machine learning predictor of breast cancer therapy response

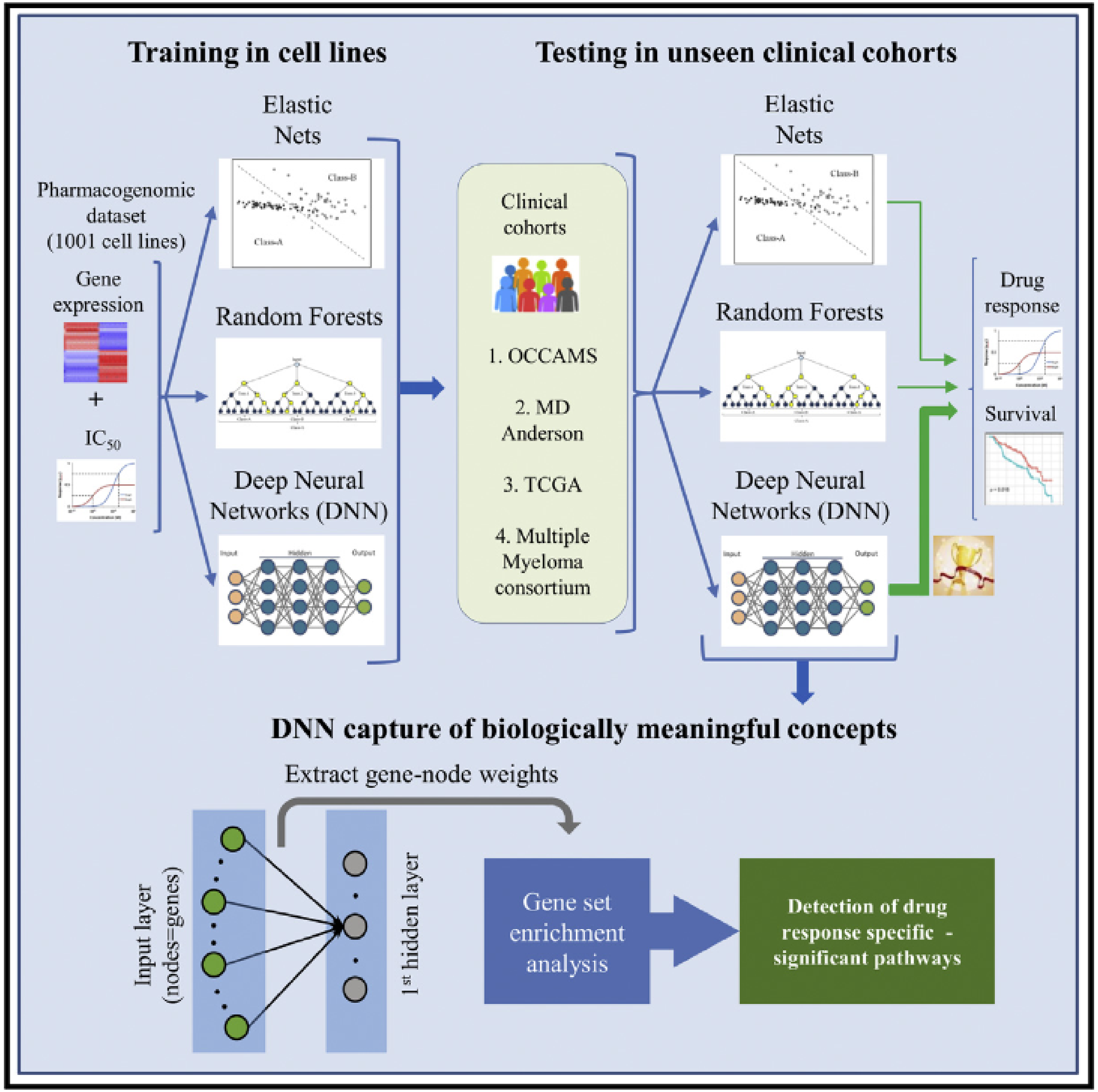
Nature, 623–629 (2022)



Results

- The degree of residual disease following therapy is monotonically associated with pre-therapy features, including tumour mutational and copy number landscapes, tumour proliferation, immune infiltration and T cell dysfunction and exclusion.
- Combining these features into a multi-omic machine learning model predicted a pathological complete response in an external validation cohort (75 patients) with an area under the curve of 0.87.

https://www.nature.com/articles/s41586-021-04278-5


Training in cell lines
Testing in unseen clinical cohorts
Pharmacogenomic dataset (1001 cell lines)
Gene expression
+
IC_{50}
Elastic Nets
Class-B
Class-A
Random Forests
Deep Neural Networks (DNN)
Input
Hidden
Output
Clinical cohorts
1. OCCAMS
2. MD Anderson
3. TCGA
4. Multiple Myeloma consortium
Elastic Nets
Random Forests
Deep Neural Networks (DNN)
Drug response
Survival
DNN capture of biologically meaningful concepts
Extract gene-node weights
Input layer (nodes=genes)
1st hidden layer
Gene set enrichment analysis
Detection of drug response specific - significant pathways

A Deep Learning Framework for Predicting Response to Therapy in Cancer

https://doi.org/10.1016/j.celrep.2019.11.017

Highlights

- A machine learning (ML) workflow is designed to predict drug response in cancer patients
- Deep neural networks (DNNs) surpass current ML algorithms in drug response prediction
- DNNs predict drug response and survival in various large clinical cohorts
- DNNs capture intricate biological interactions linked to specific drug response pathways