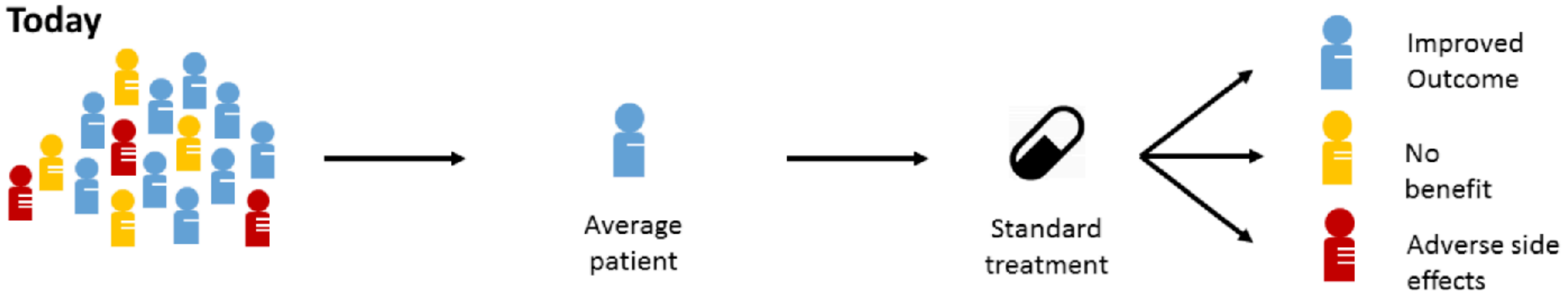
Kevin Wang

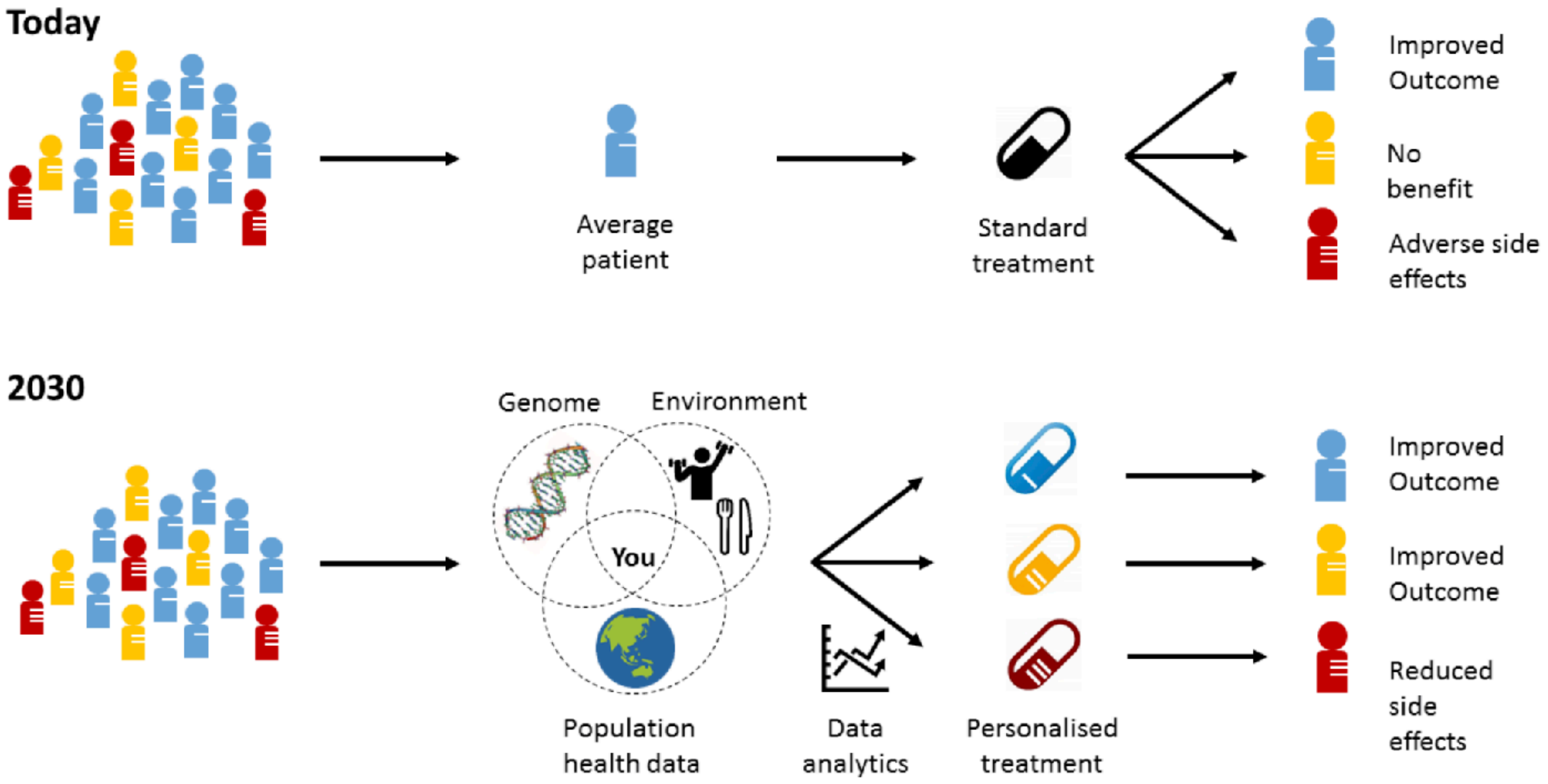
From linear regression to precision medicine

Precision medicine: predicting best cause of action using omics data



https://www.chiefscientist.gov.au/wp-content/uploads/Precision-medicine-final.pdf

Precision medicine: predicting best cause of action using omics data



https://www.chiefscientist.gov.au/wp-content/uploads/Precision-medicine-final.pdf

HOW DID WE GET HERE?

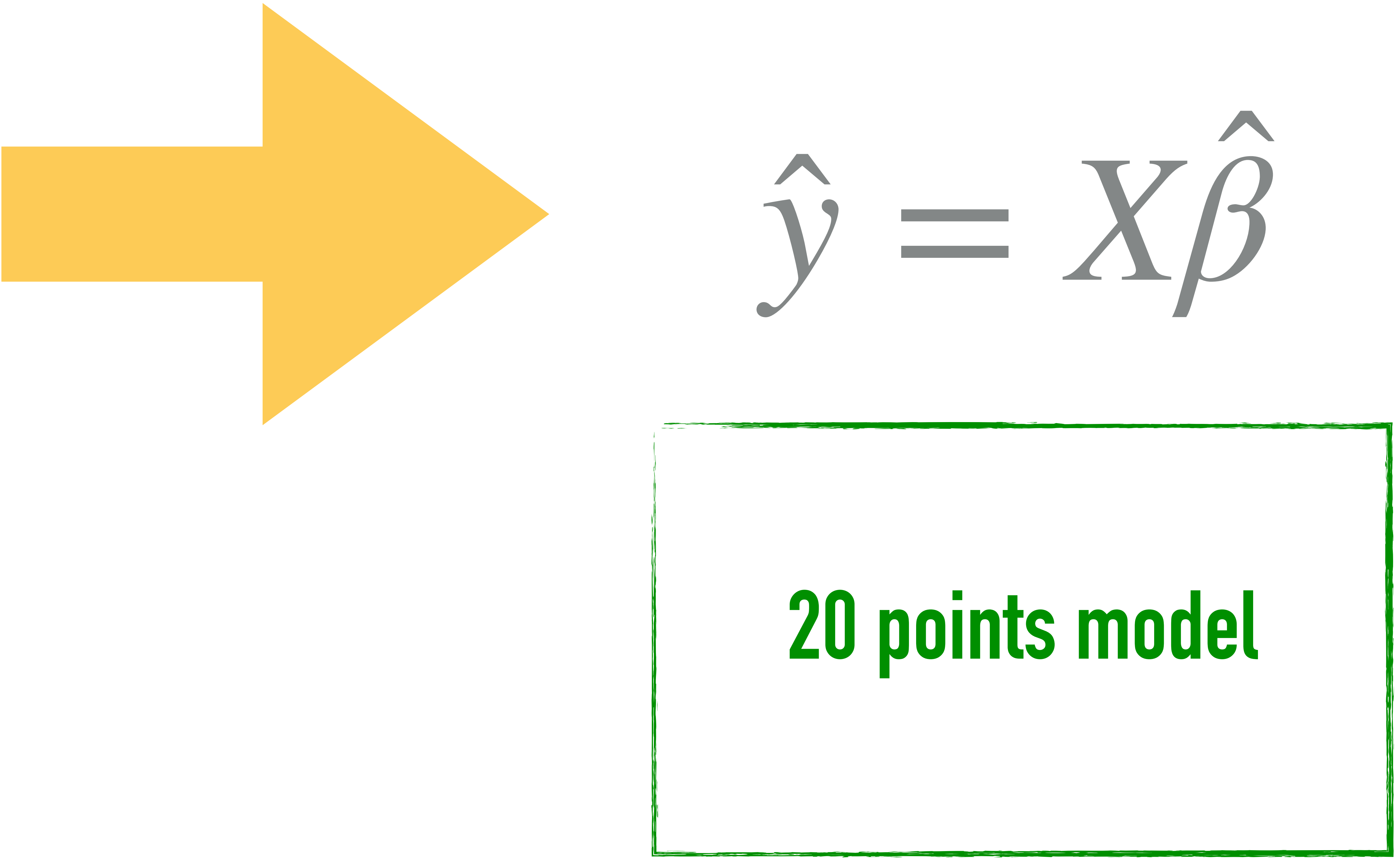
REGRESSION ANALYSIS BEFORE THE 1950’S

Linear regression

- A response variable often captures some complicated physical mechanism
- Predictor variables are usually quantities that are more manageable

Cox Proportional Hazard model

- Framingham heart disease risk score:
  - Age (Years)
  - Cholesterol (mg/dL)
  - If smoker (Yes/No)
  - HDL cholesterol (mg/dL)
  - Systolic blood pressure (mm Hg)



$$\hat{y} = X\hat{\beta}$$

20 points model

Least square regression is a projection

- A linear regression aims to explain as much complications in the response variable using the predictors as possible.
- L2 projection of y upon a subspace spanned by the column vectors of the predictor matrix.

Linear regression as minimisation problem

- A well studied problem

When the least squares solution fails

- When $n < p$, we have more parameters to be estimated (p) than observations we have collected (n). An overdetermined linear system.
- A simulation shows that the beta vector blows up!

CONSTRAINED REGRESSION ANALYSIS

When the beta blows up, we put a lid on it

- Ridge regression came around1945.

The same simulation, but with Ridge

REGRESSION OF THE 21ST CENTURY

Least Absolute Shrinkage and Selection Operator

- There are “only” three norms: L1, L2 and L-inf
- Tibshirani 1996 replaced L2 penalty with the L1.
- The original paper is now cited about 30,000 times.

Why variable selection

- Every variable you put into your model introduces variations into your estimation and prediction.
- For the informative variables, this is fine! In fact, you should be happy because it tells you a broad range of what to expect.
- But for non-informative variables, this variation is a nuisance, a noise that undermines your model.
- Variable selection aims to keep the informative ones and kill the latter.

Various wonderful properties of Lasso

Visualisation of the Lasso and the simulation

Robert Tibshirani

A GALLERY DEDICATED TO THE LASSO

Whenever Lasso fails, a small modification of the optimisation equation fixes it

- Highly correlated features
- Stability in selection
- Group structures
- Fused Lasso
-

CPOP: CROSS PLATFORM OMICS PREDICTION

Statistics is not invincible

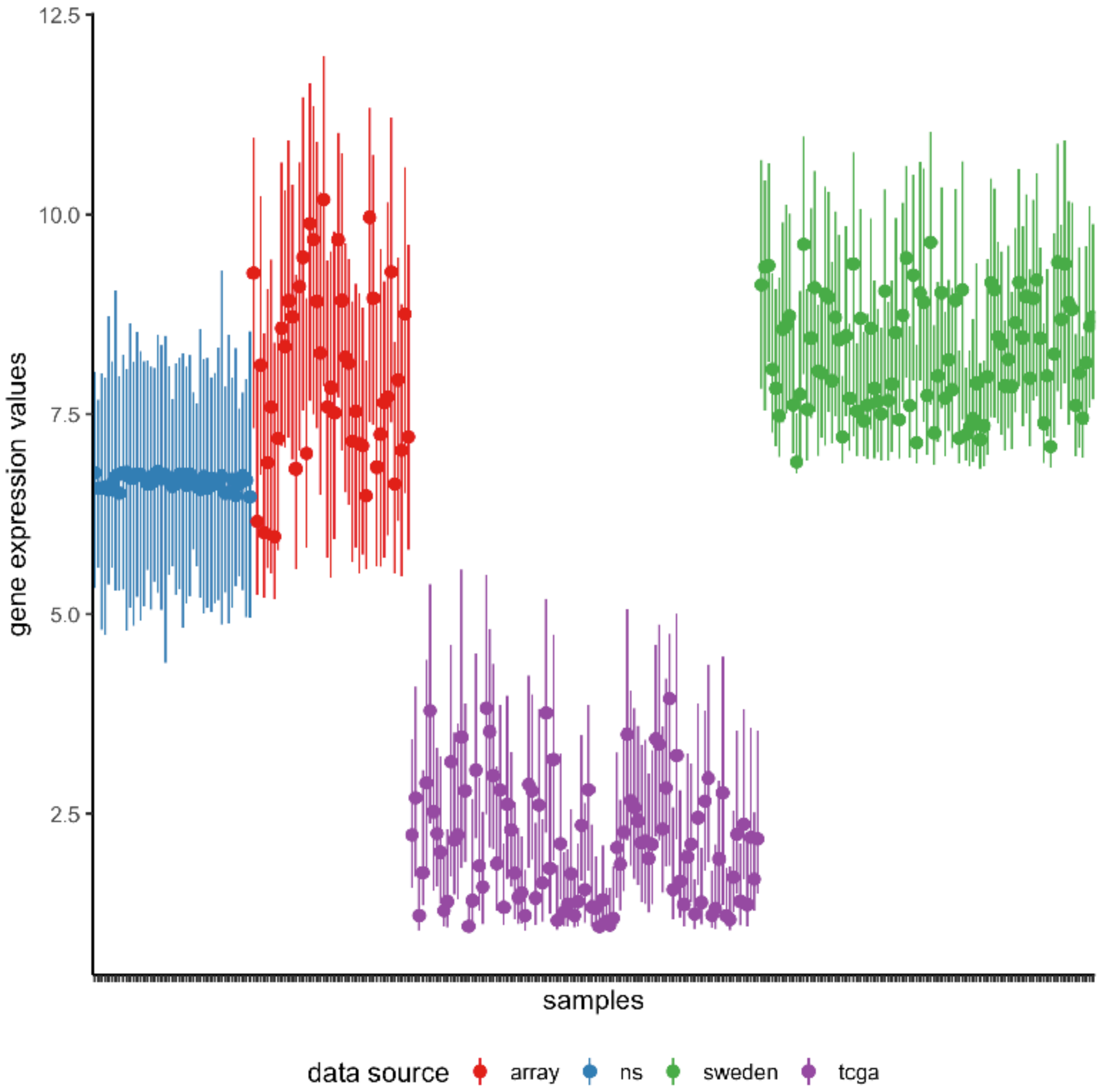
- When your training data and validation data are not of the same statistical properties, any model would do miserably.

Omics-based clinical risk score: what is so difficult?

Omics features are typically on a relative scale and unitless

$$\hat{y}_1 = X_1\hat{\beta}_1$$

$$\hat{y}_2 = X_2\hat{\beta}_1 = (X_1 + \mathbf{1})\hat{\beta}_1$$

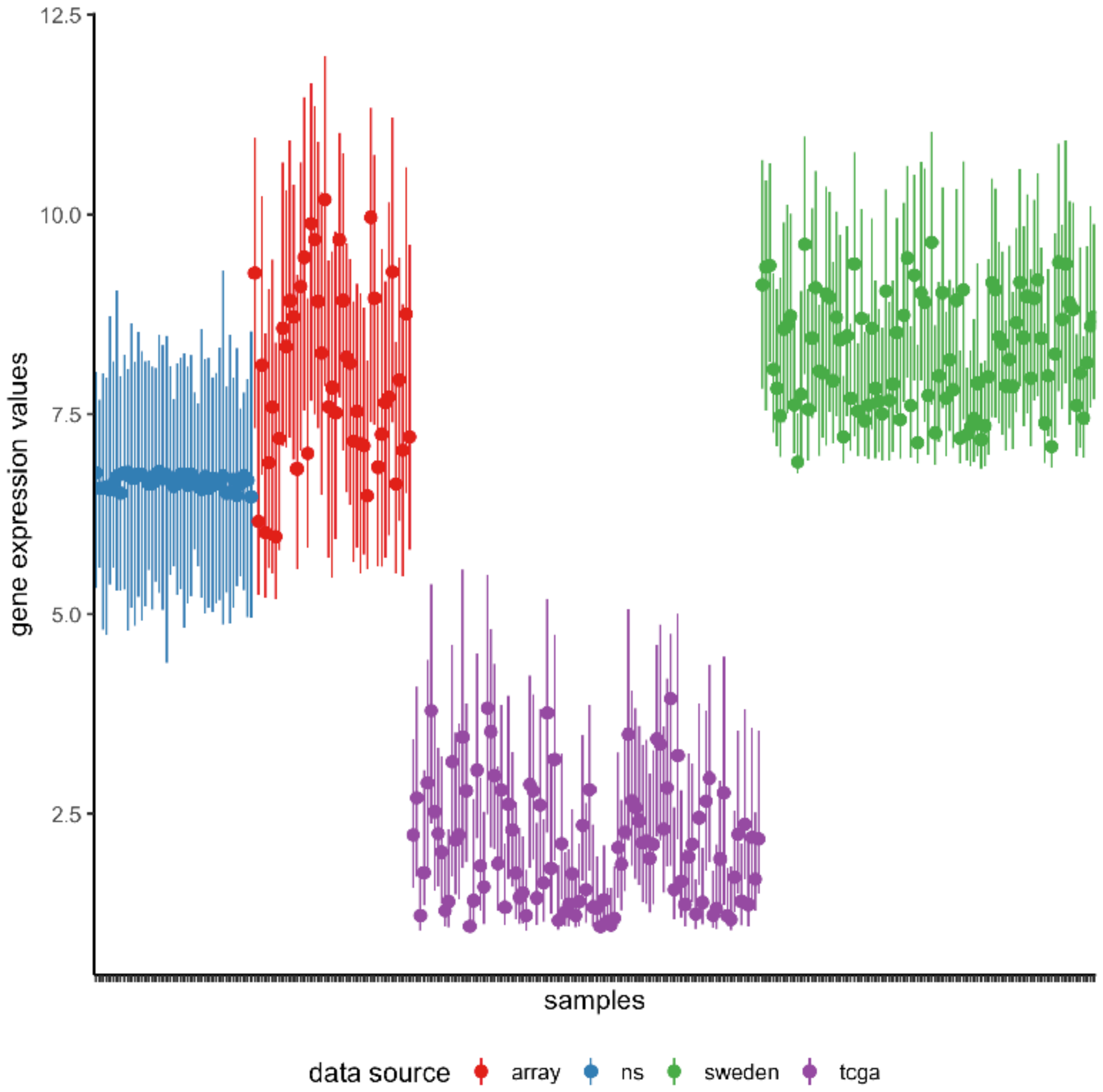


Omics-based clinical risk score: what is so difficult?

1. Practical: we cannot renormalise data in a clinical setting

	Sample 1	Sample 2	Sample 3
Gene 1	1.2	2.1	1.5
Gene 2	5.6	4.6	7.1
Gene 3	9.2	10.1	6.9
Gene 4	4.1	3.6	2.7

Sample 4
1.2
1.4
8.6
7.1



The flowchart of a clinical risk score

Data

Model

Prediction

(X_1, y_1)

(X_2, y_2)

$\hat{\beta}_1$

$\hat{y}_1 = X_1\hat{\beta}_1$

$\hat{y}_2 = X_2\hat{\beta}_1$

The flowchart of a clinical risk score

Data

$$(X_1, y_1)$$

$$(X_2, y_2)$$

No renormalisation

Model

$$\hat{\beta}_1$$

No model retraining

Prediction

$$\hat{y}_1 = X_1\hat{\beta}_1$$

$$\hat{y}_2 = X_2\hat{\beta}_1$$

Scale-equivalent prediction

Statistical challenges

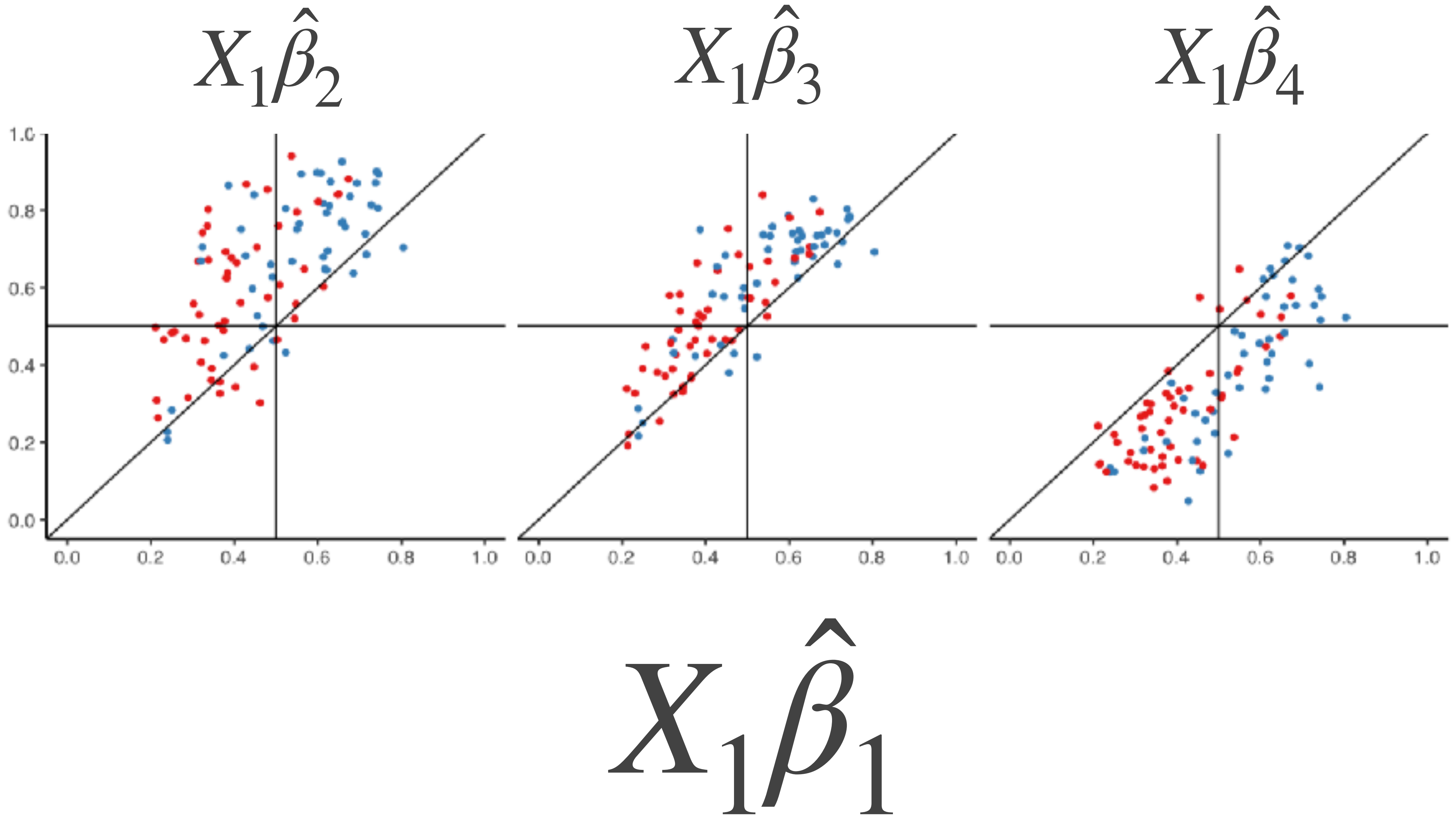
1. Concordance in gene features scaling across platforms

2. Concordance in feature selection and coefficient estimates

3. Single-patient prediction

Transferability

For the same samples,
the prediction from one gene expression platform
should be equivalent to another platform



First component of CPOP: feature engineering

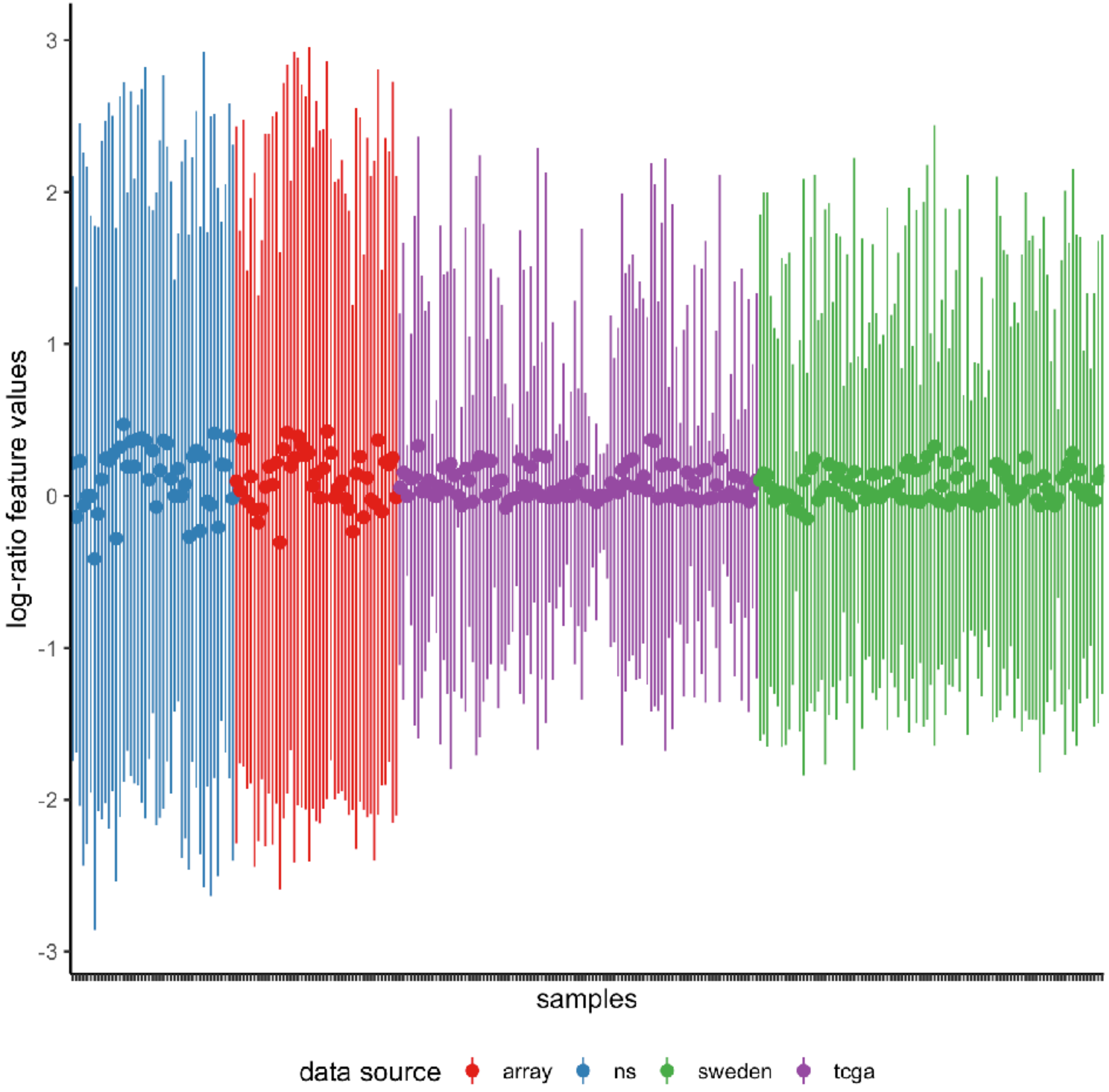
就让我 来次透彻心扉的痛
都拿走 让我再次两手空空
只有奄奄一息过
那个真正的我
他才能够诞生

Within-sample feature standardisation

Single-patient prediction prevents us from calculating any cross-sample statistics, so the natural solution is within-sample standardisation

Log-ratio

log(gene A) – log(gene B)



The solution is trivial?

1. Concordance in gene features scaling across platforms
2. Concordance in feature selection and coefficient estimates
3. Single-patient prediction

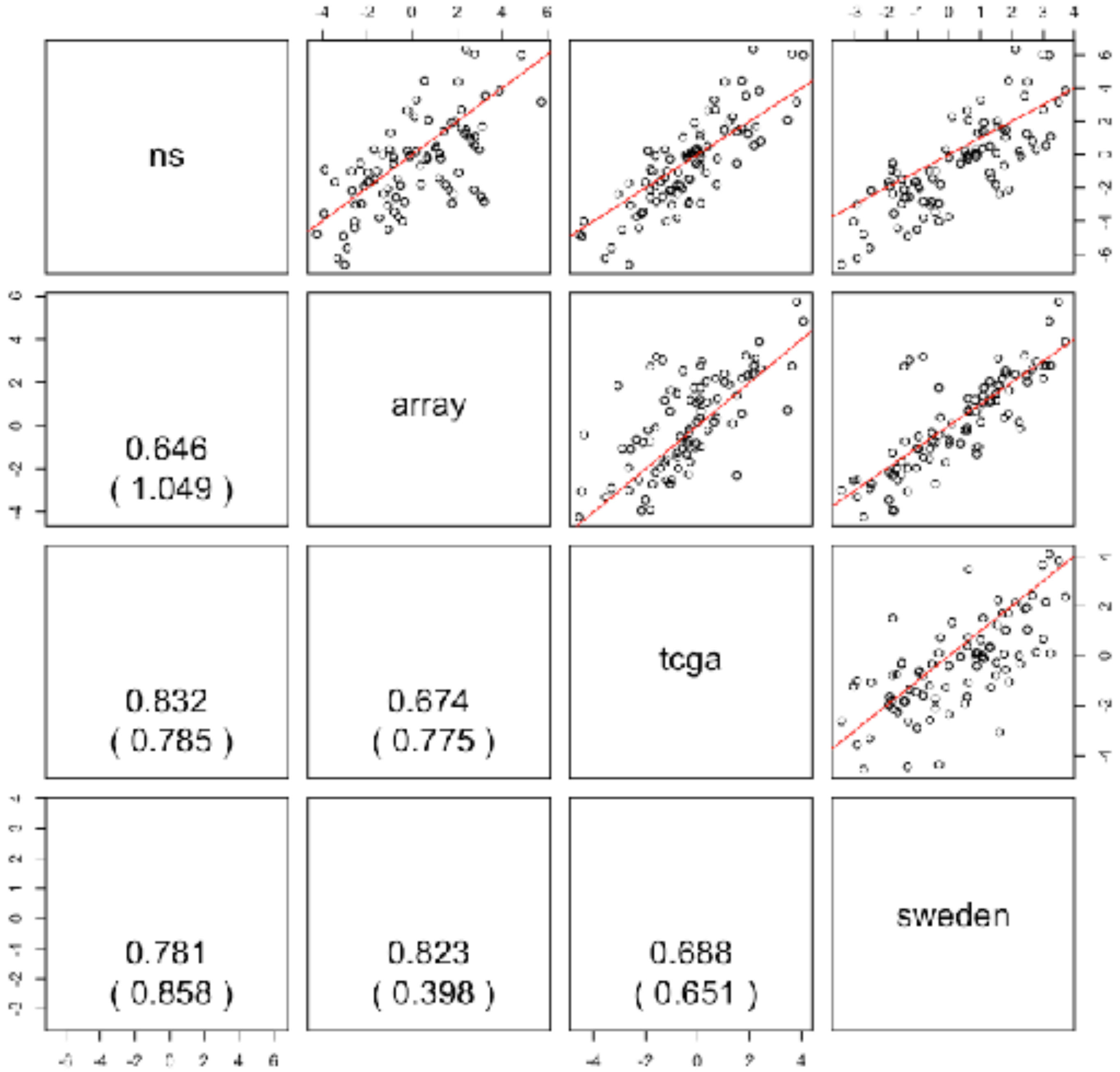
The solution is trivial?

1. Concordance in log-ratio features scaling across platforms

2. Concordance in feature selection and coefficient estimates

3. ~~Single-patient prediction~~

Concordance of log-ratios after Lasso selection

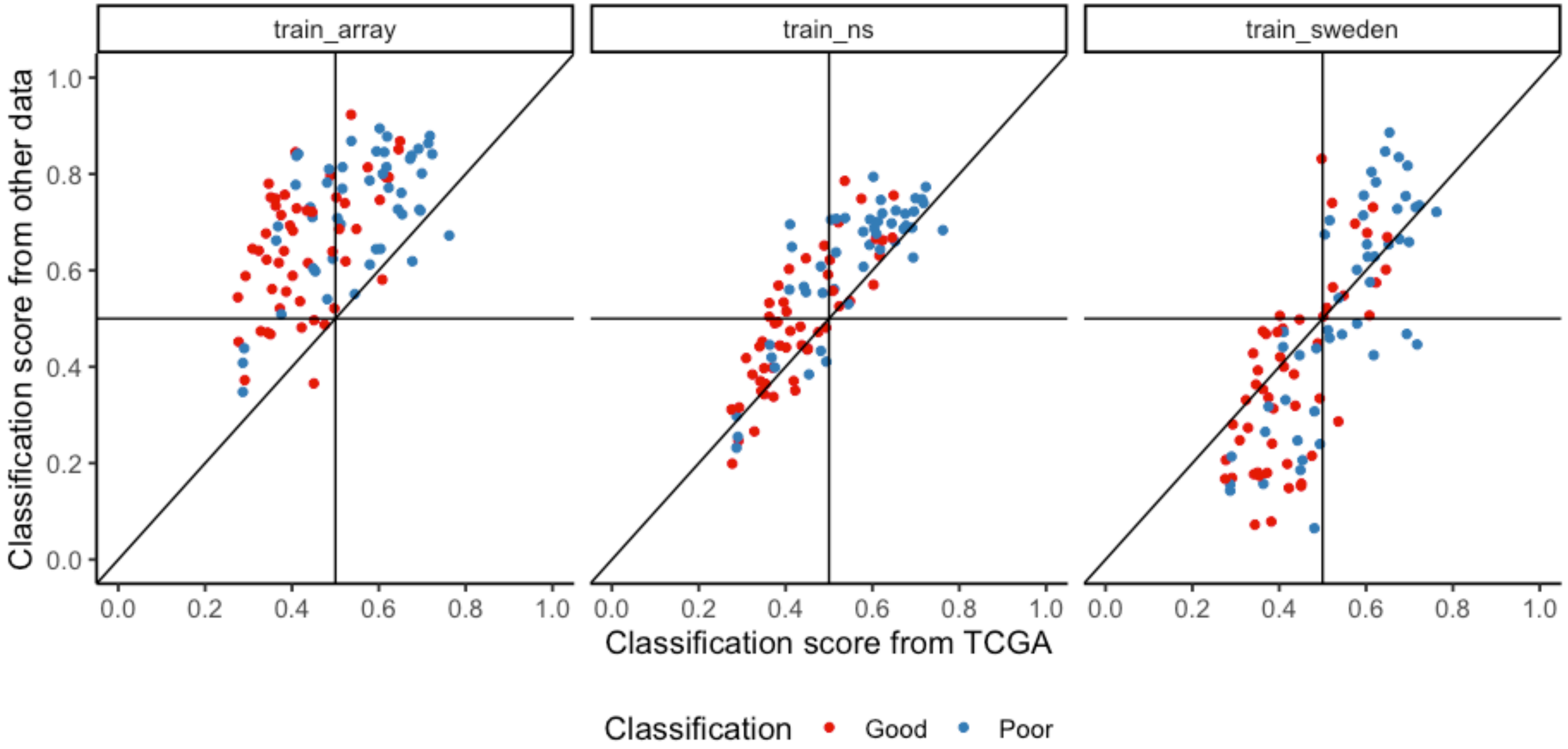


The solution is trivial?

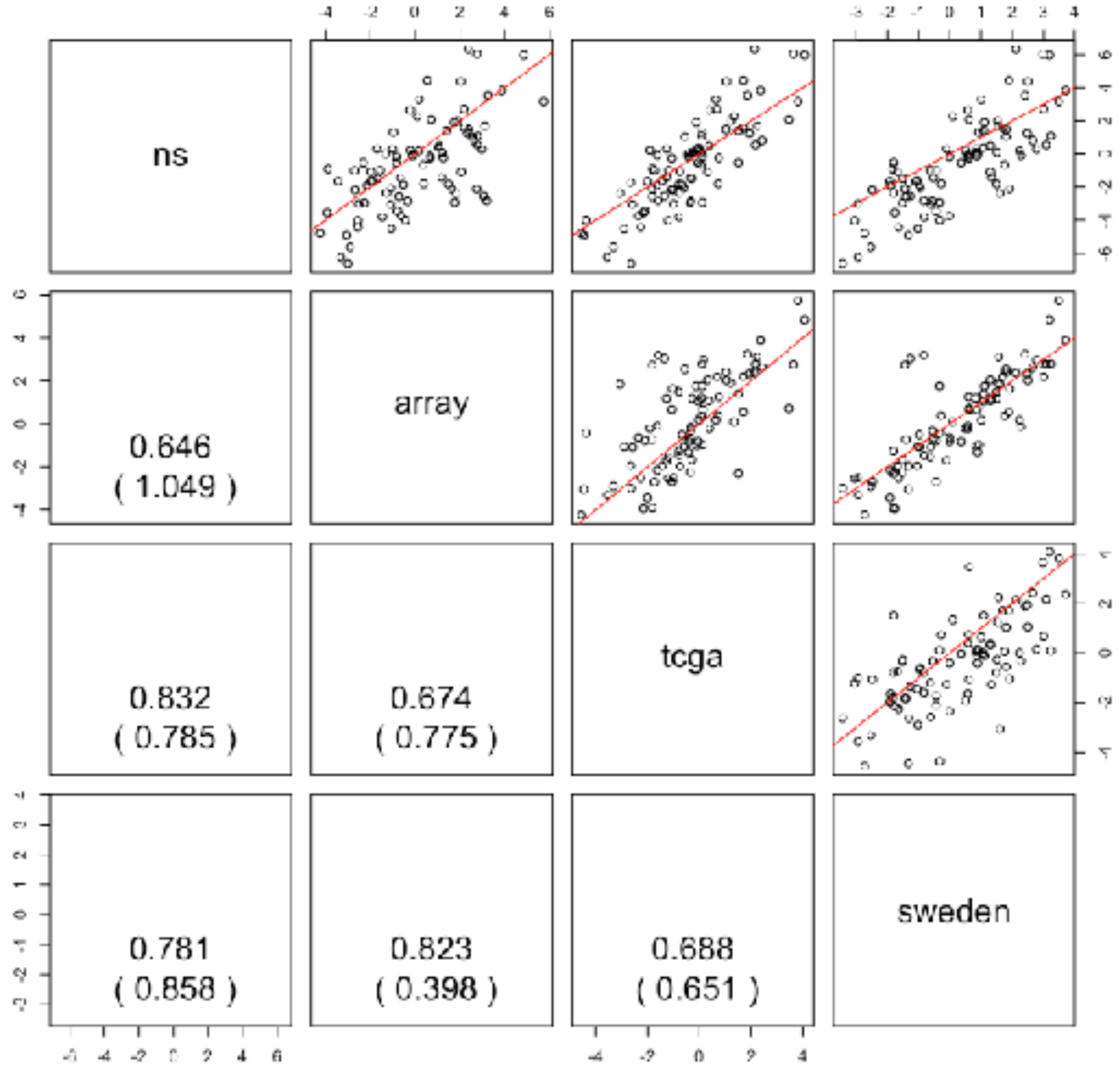
1. Concordance in log-ratio features scaling across platforms

2. Concordance in feature selection and coefficient estimates

~~3. Single-patient prediction~~



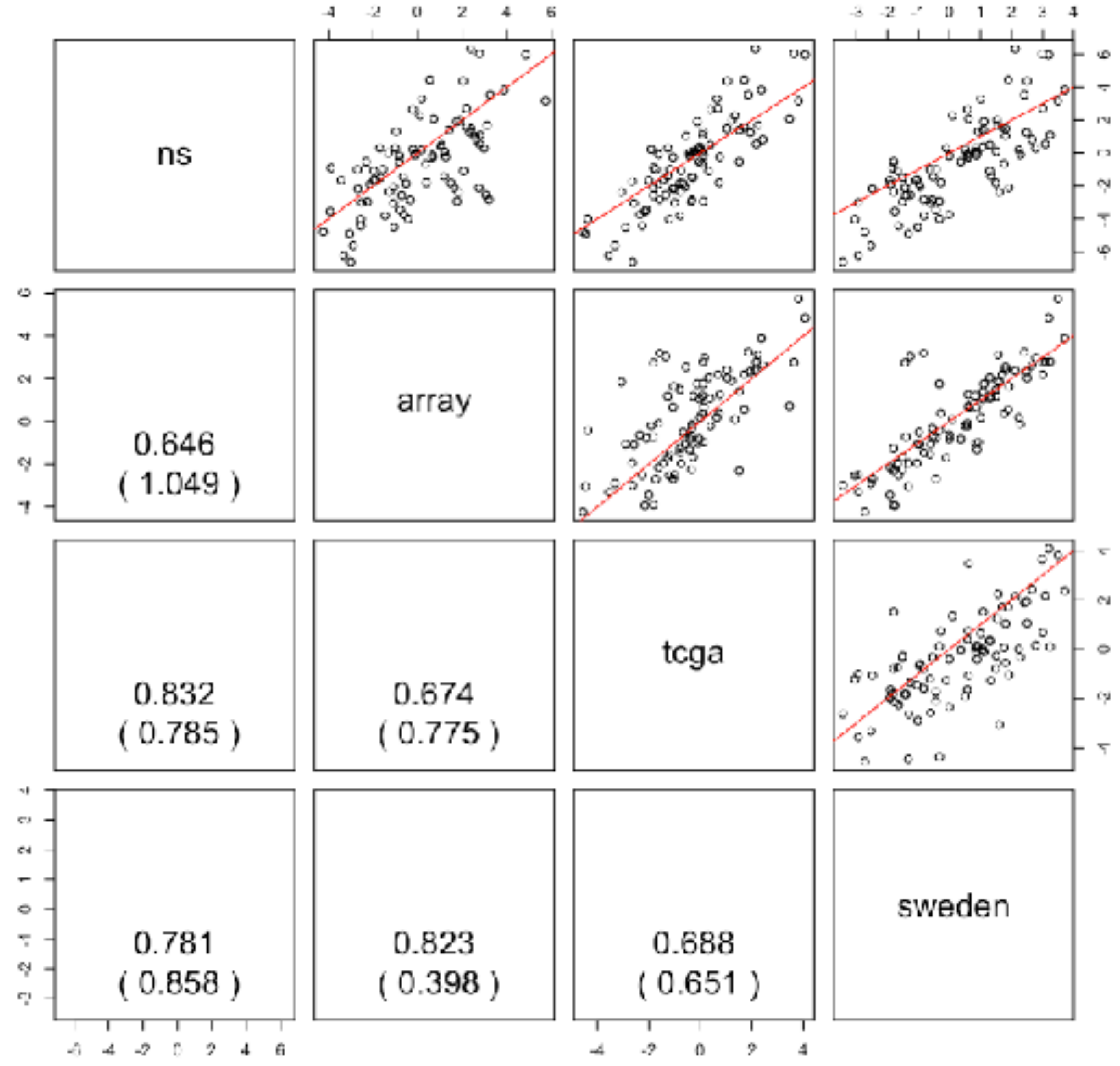
Concordance of log-ratios after Lasso selection

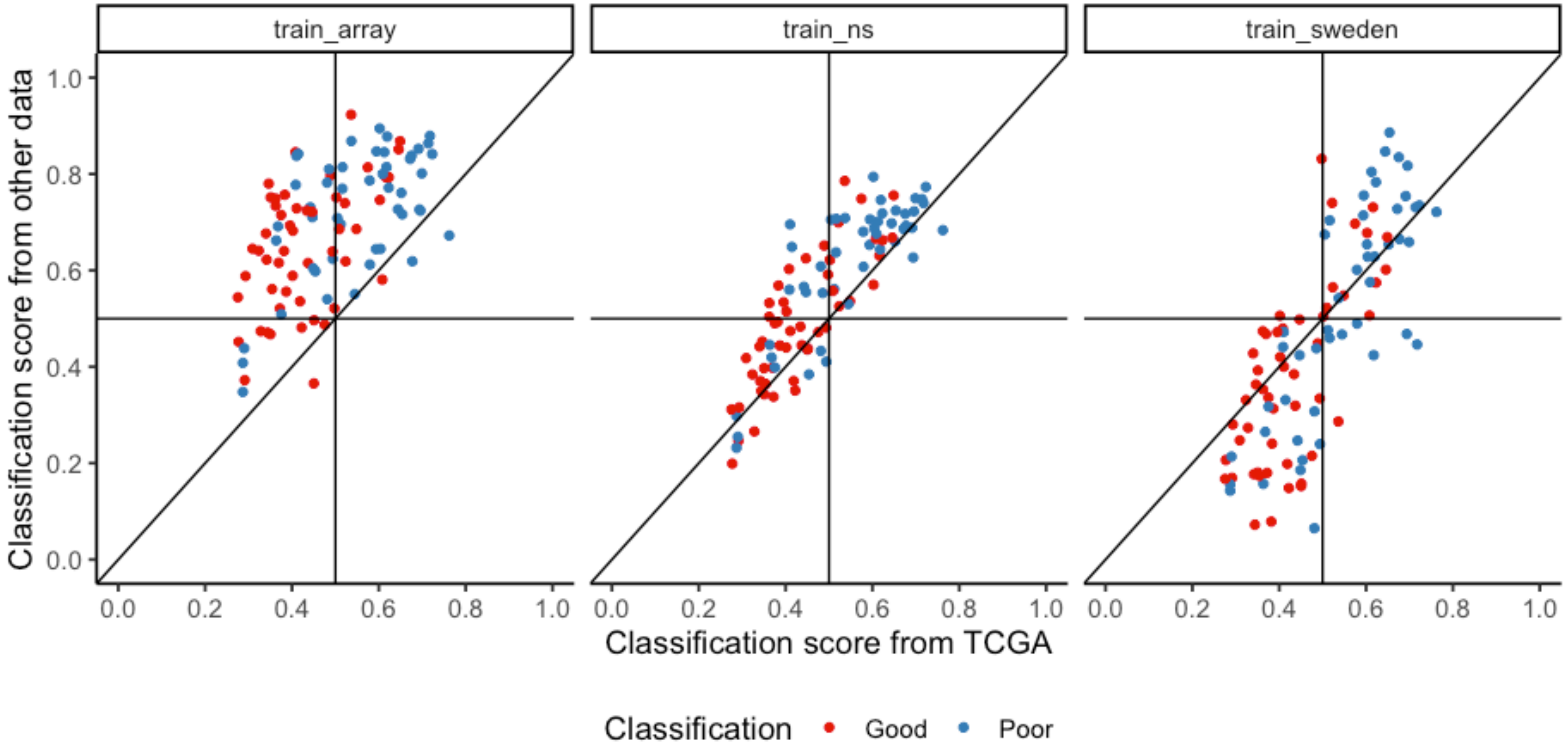


The solution is trivial?

1. Concordance in log-ratio features scaling across platforms
2. Concordance in feature selection and coefficient estimates
3. ~~Single-patient prediction~~

Concordance of log-ratios after Lasso selection





Lasso variable selection is NOT stable

The solution is not so trivial

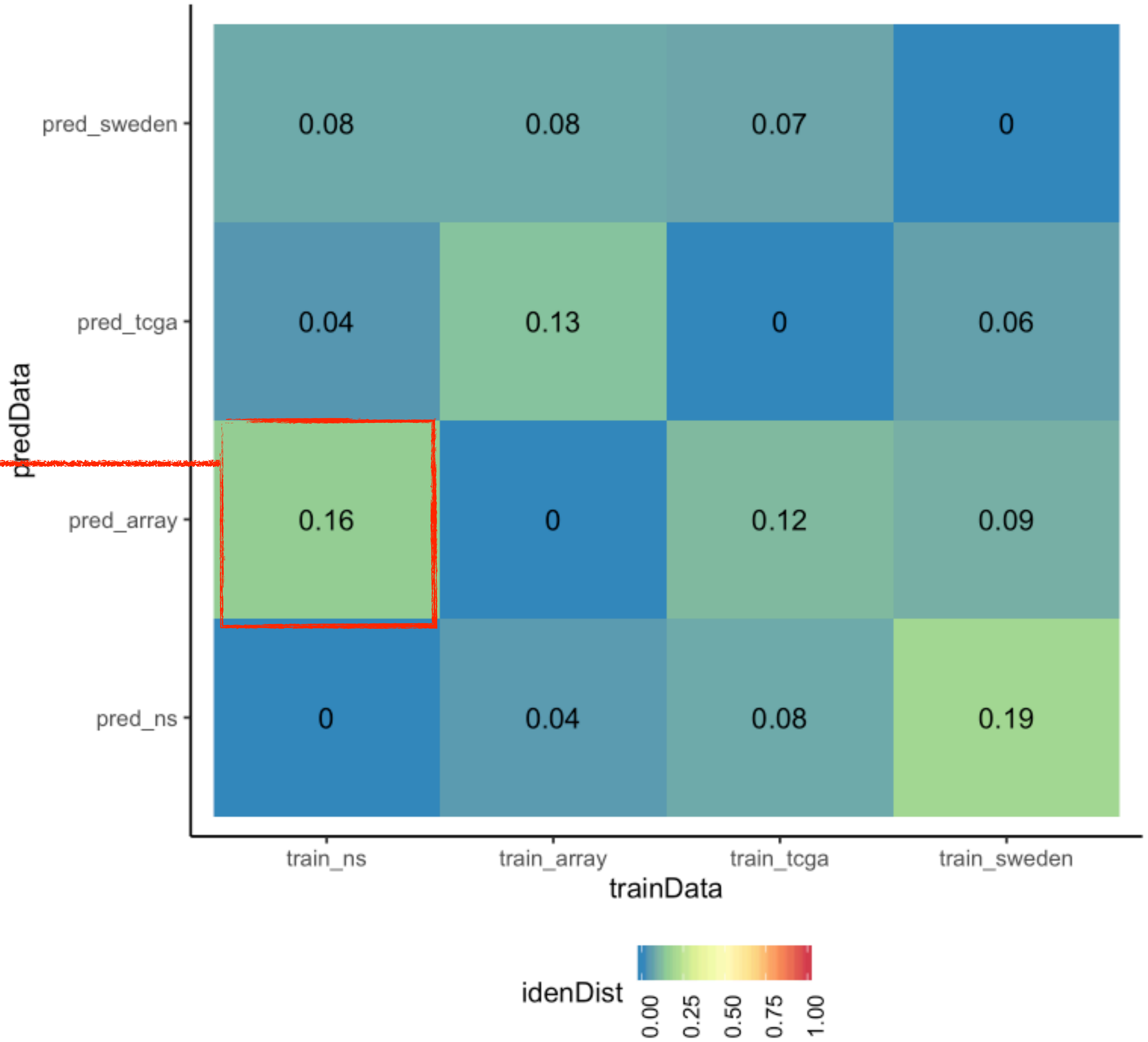
Estimated prognosis probabilities from

training data

vs

validation data

differ by 0.16 on average



Second component of CPOP:
feature selection and estimation stability

我曾经毁了我的一切
只想永远地离开
我曾经堕入无边黑暗
想挣扎无法自拔
我曾经 像你 像他 像那野草 野花
绝望着 也渴望着
也哭 也笑 也平凡着

Motivation for CPOP: one patient cohort, two gene expression data

$$Z_1\hat{\beta}_1 \approx Z_2\hat{\beta}_2$$

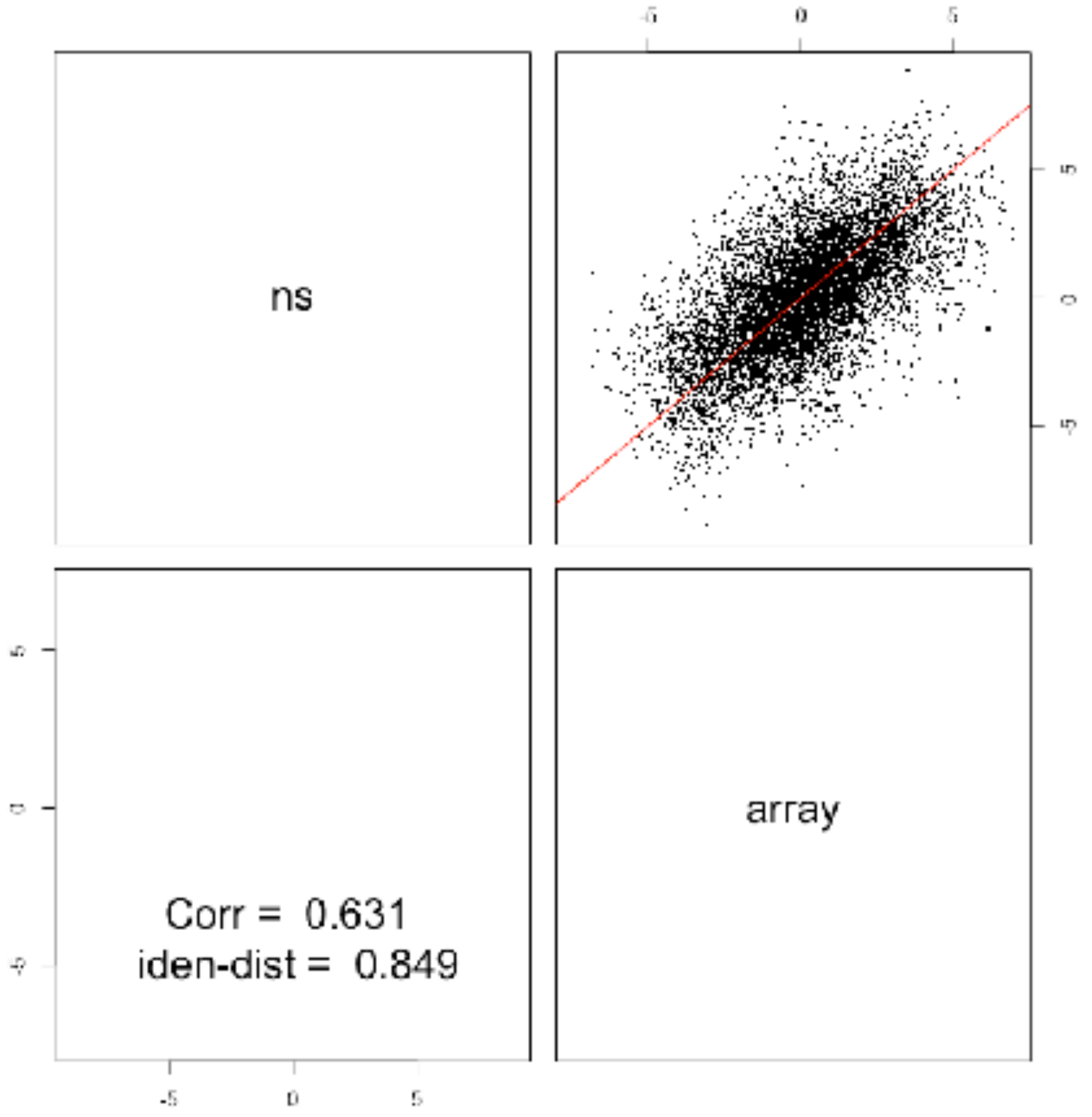
loosely translate to

$$Z_1 \approx Z_2$$ column-wise

$$\hat{\beta}_1 \approx \hat{\beta}_2$$ element-wise

CPOP weighted variable selection

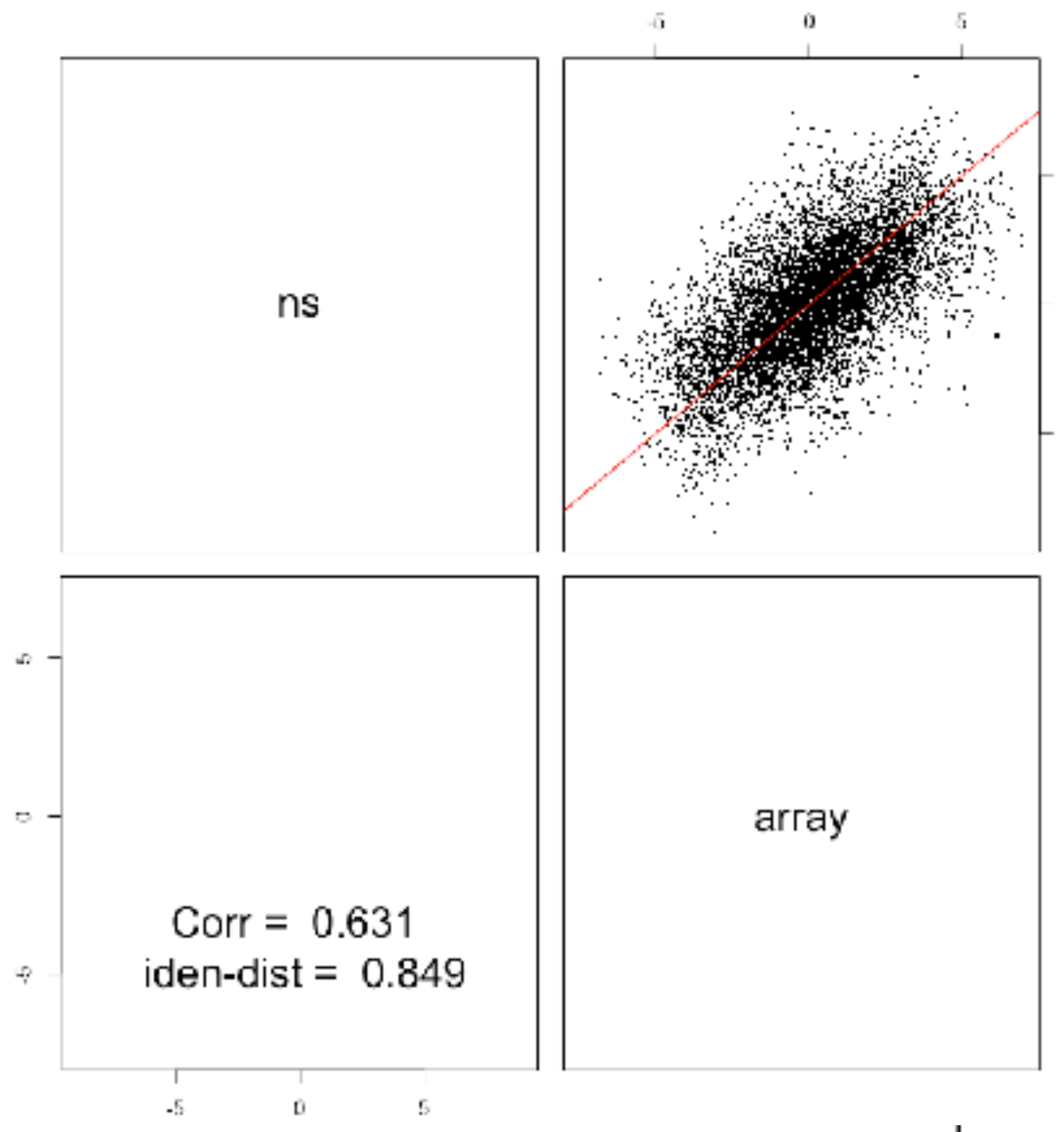
1. Perform a **weighted Lasso** by placing higher weights on features closer to the identity line



$$Z_1 \approx Z_2$$

CPOP weighted variable selection

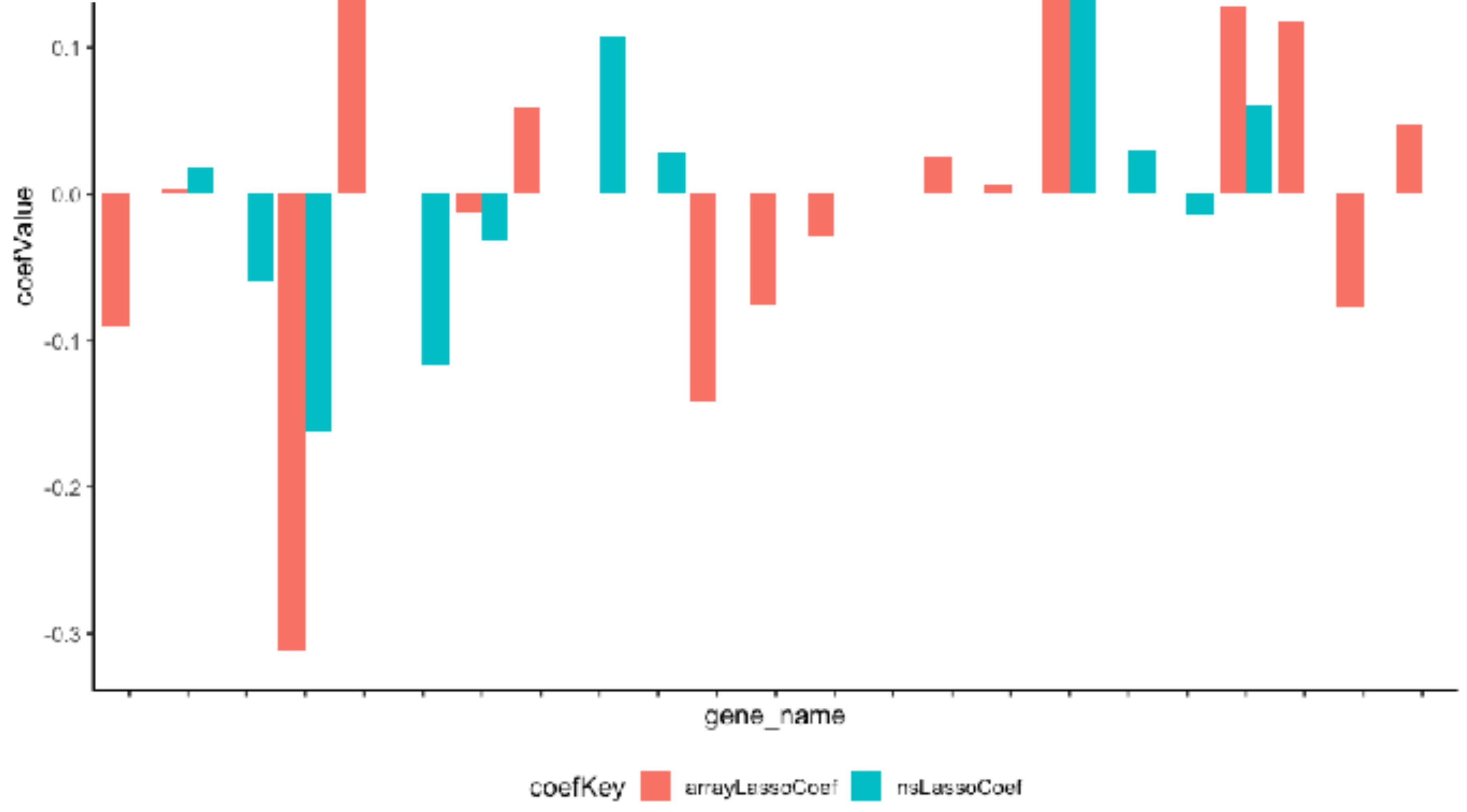
1. Perform a **weighted Lasso** by placing higher weights on features closer to the identity line



$$Z_1 \approx Z_2$$

2. Perform a **Ridge regression** and only retain those features with coefficients similar to each other

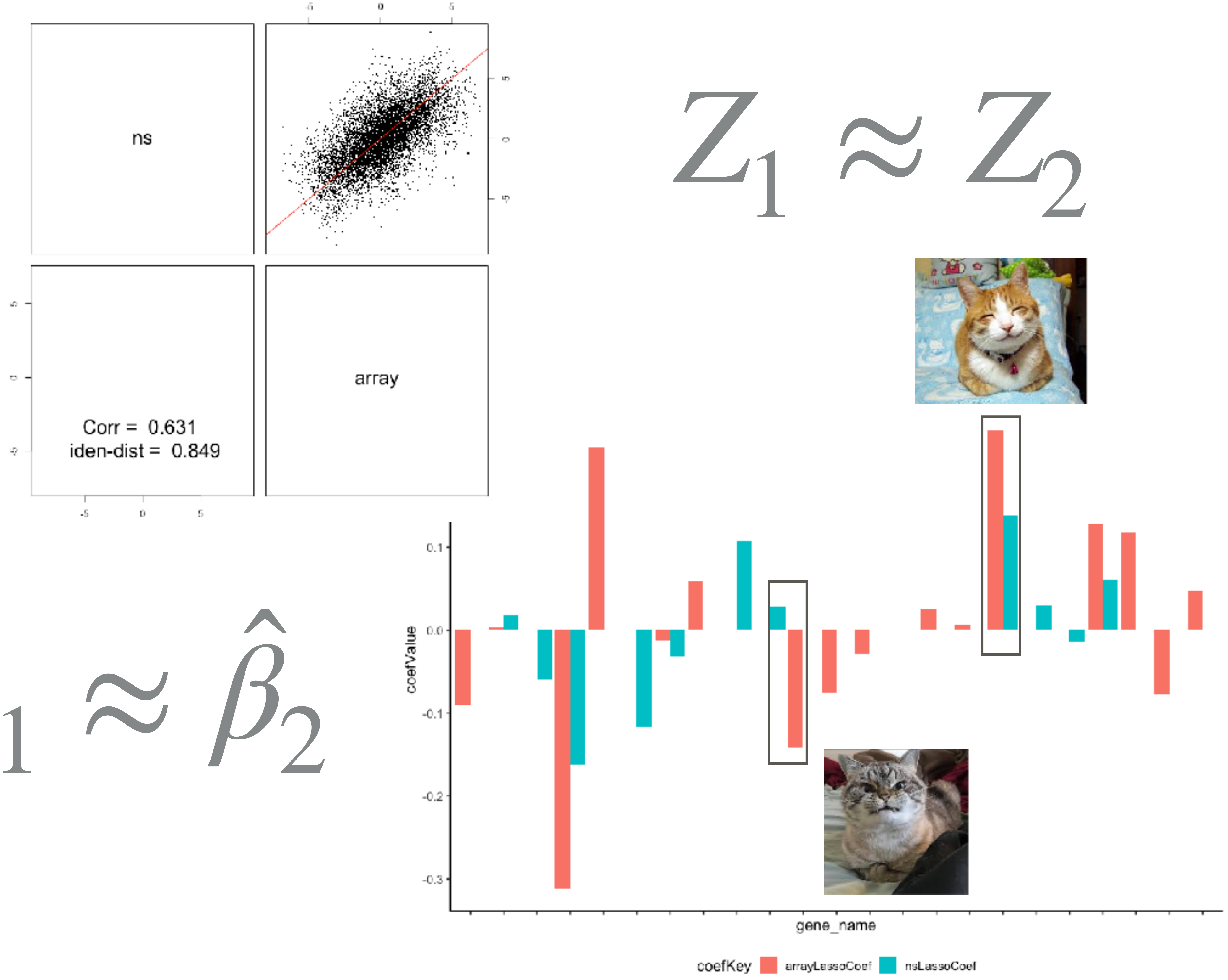
$$\hat{\beta}_1 \approx \hat{\beta}_2$$



CPOP weighted variable selection

1. Perform a **weighted Lasso** by placing higher weights on features closer to the identity line

2. Perform a **Ridge regression** and only retain those features with coefficients similar to each other



$$Z_1 \approx Z_2$$

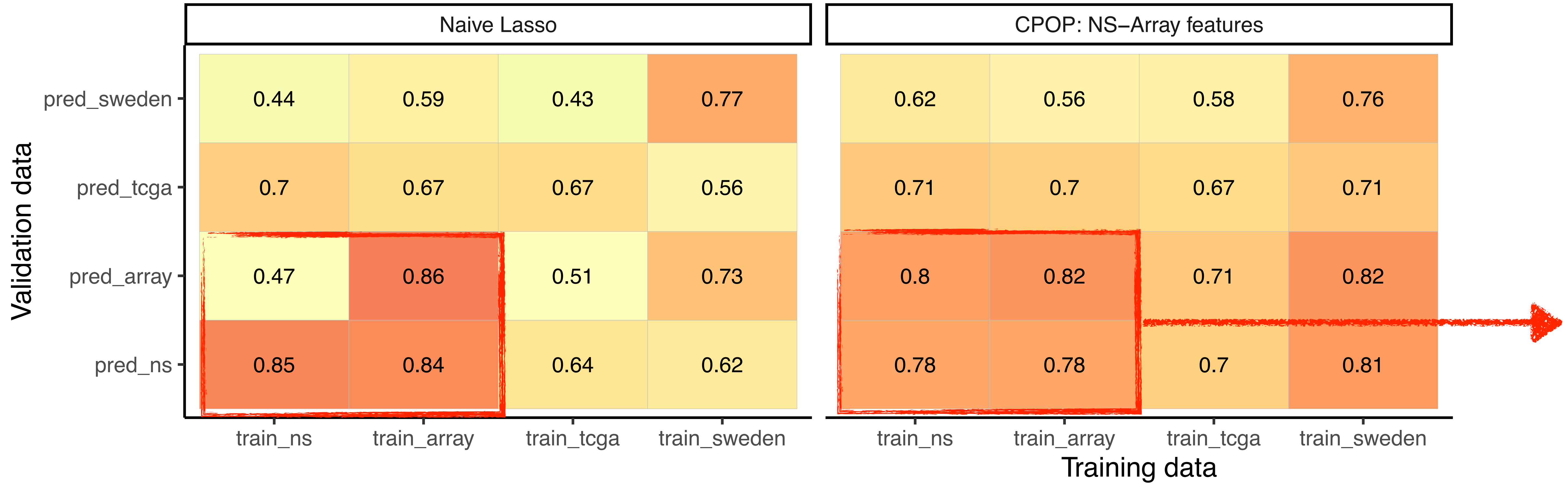
$$\hat{\beta}_1 \approx \hat{\beta}_2$$

CPOP results 1: four melanoma data

1. Predictive performance of CPOP matches that of re-substitution

2. Smaller identity distance between predicted values

F1 classification statistic under various models



Naive Lasso

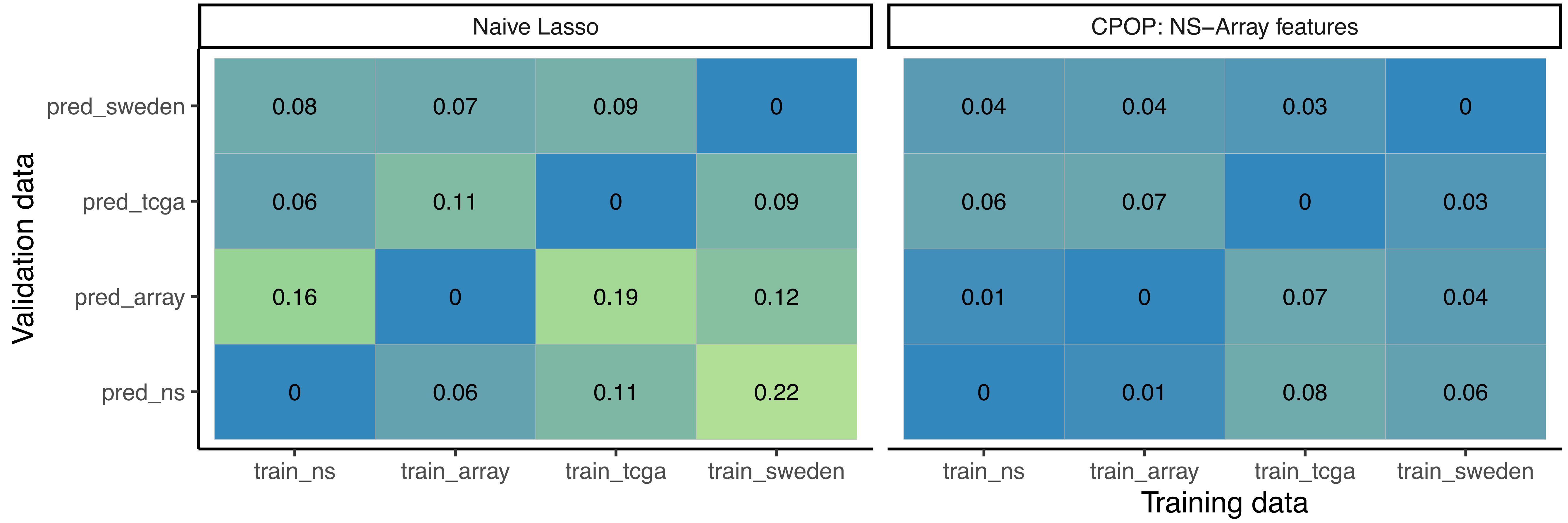
Validation data	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.44	0.59	0.43	0.77
pred_tcga	0.7	0.67	0.67	0.56
pred_array	0.47	0.86	0.51	0.73
pred_ns	0.85	0.84	0.64	0.62

CPOP: NS–Array features

Validation data	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.62	0.56	0.58	0.76
pred_tcga	0.71	0.7	0.67	0.71
pred_array	0.8	0.82	0.71	0.82
pred_ns	0.78	0.78	0.7	0.81

Datasets for feature selection

Identity distance between predicted values under various models



Naive Lasso

Validation data	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.08	0.07	0.09	0
pred_tcga	0.06	0.11	0	0.09
pred_array	0.16	0	0.19	0.12
pred_ns	0	0.06	0.11	0.22

CPOP: NS–Array features

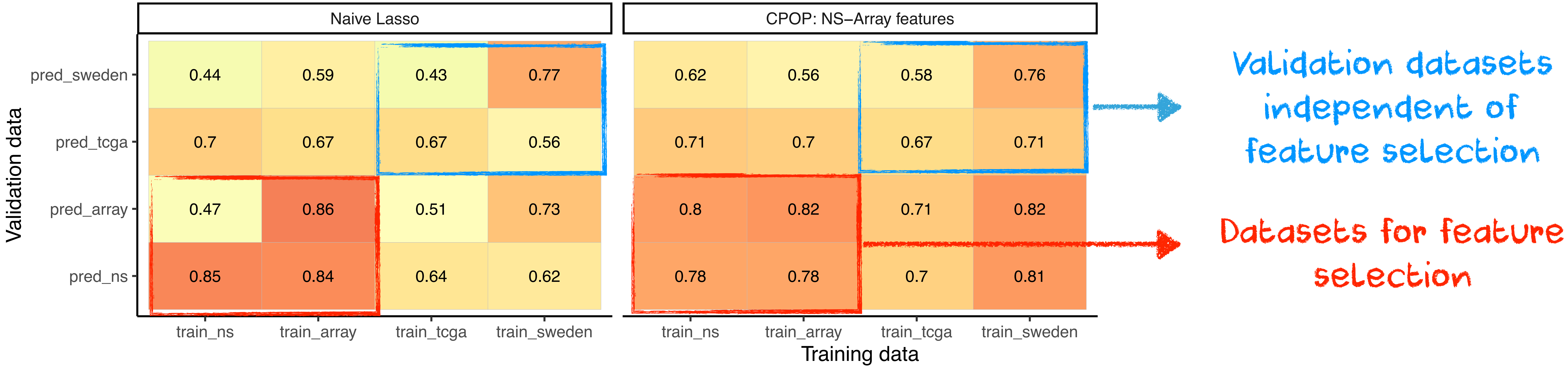
Validation data	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.04	0.04	0.03	0
pred_tcga	0.06	0.07	0	0.03
pred_array	0.01	0	0.07	0.04
pred_ns	0	0.01	0.08	0.06

CPOP results 1: four melanoma data

1. Predictive performance of CPOP matches that of re-substitution

2. Smaller identity distance between predicted values

F1 classification statistic under various models



Naive Lasso

Validation data	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.44	0.59	0.43	0.77
pred_tcga	0.7	0.67	0.67	0.56
pred_array	0.47	0.86	0.51	0.73
pred_ns	0.85	0.84	0.64	0.62

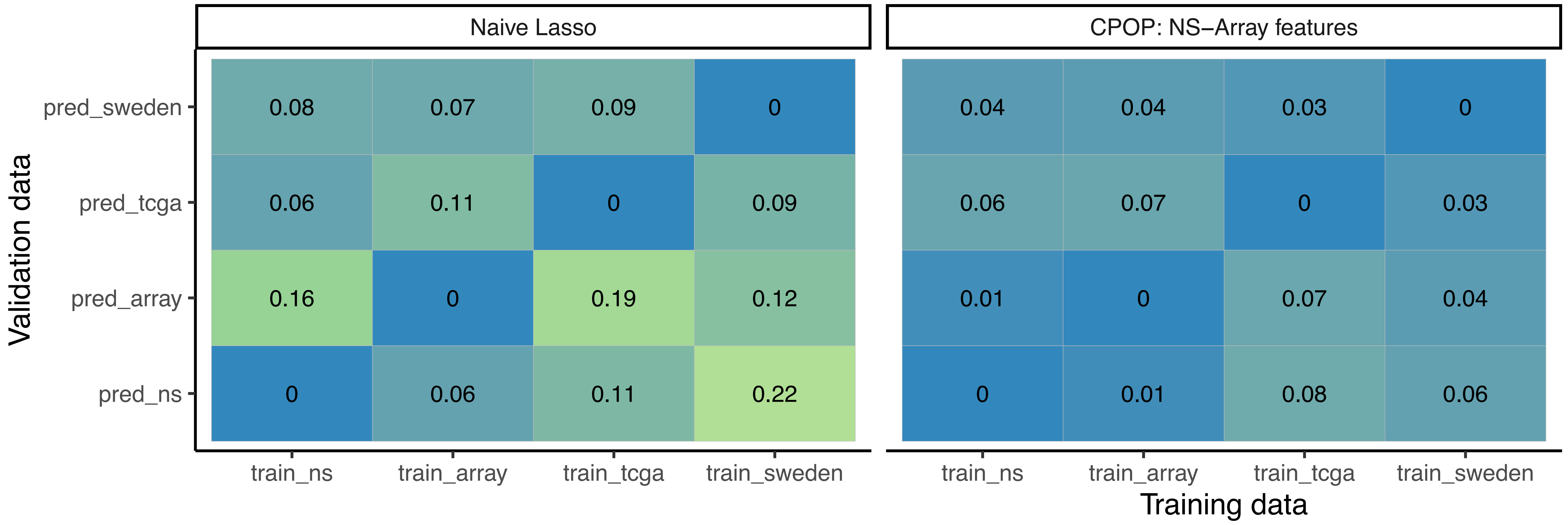
CPOP: NS–Array features

Validation data	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.62	0.56	0.58	0.76
pred_tcga	0.71	0.7	0.67	0.71
pred_array	0.8	0.82	0.71	0.82
pred_ns	0.78	0.78	0.7	0.81

Validation datasets independent of feature selection

Datasets for feature selection

Identity distance between predicted values under various models



Naive Lasso

Validation data	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.08	0.07	0.09	0
pred_tcga	0.06	0.11	0	0.09
pred_array	0.16	0	0.19	0.12
pred_ns	0	0.06	0.11	0.22

CPOP: NS–Array features

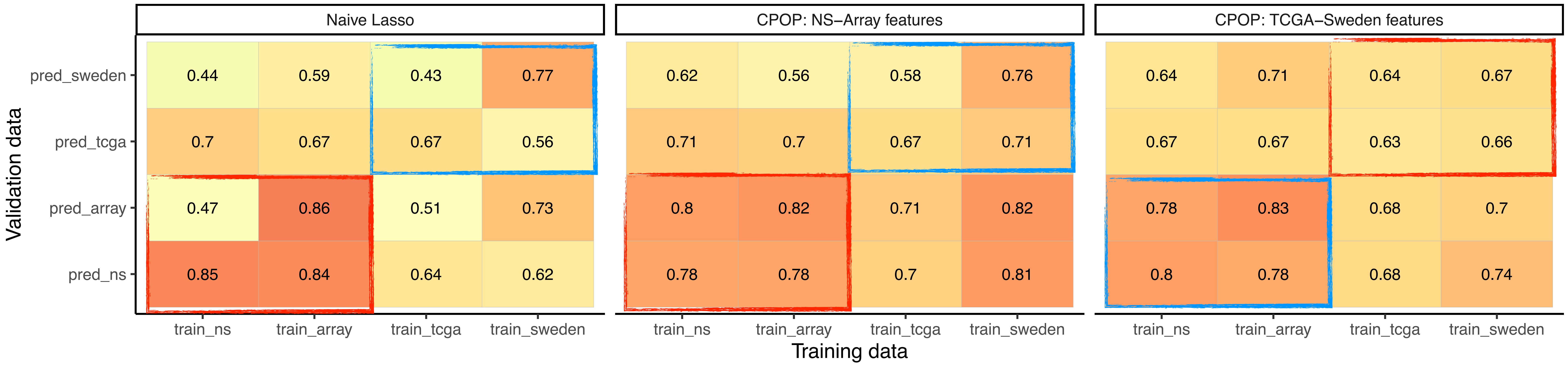
Validation data	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.04	0.04	0.03	0
pred_tcga	0.06	0.07	0	0.03
pred_array	0.01	0	0.07	0.04
pred_ns	0	0.01	0.08	0.06

CPOP results 1: four melanoma data

1. Predictive performance of CPOP matches that of re-substitution

2. Smaller identity distance between predicted values

F1 classification statistic under various models



Naive Lasso

	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.44	0.59	0.43	0.77
pred_tcga	0.7	0.67	0.67	0.56
pred_array	0.47	0.86	0.51	0.73
pred_ns	0.85	0.84	0.64	0.62

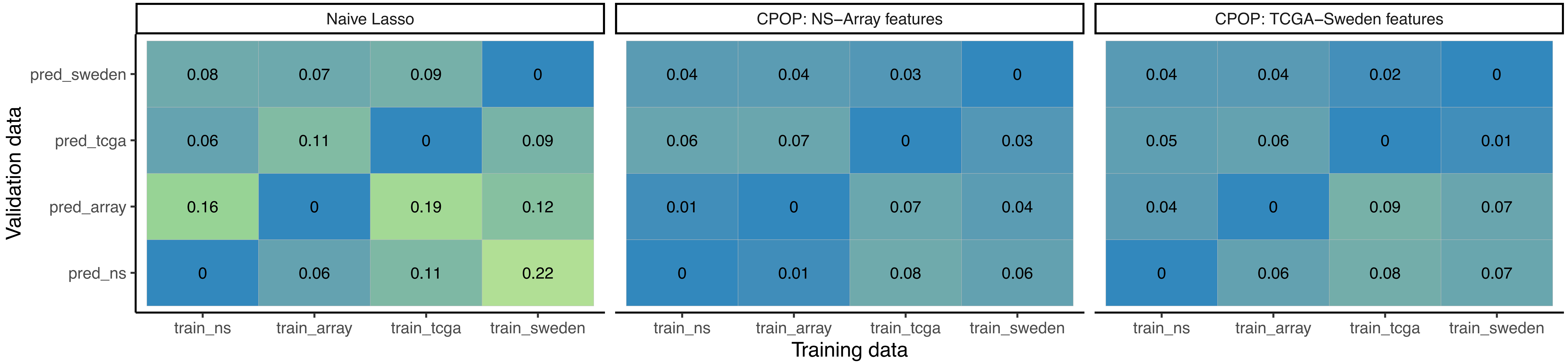
CPOP: NS–Array features

	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.62	0.56	0.58	0.76
pred_tcga	0.71	0.7	0.67	0.71
pred_array	0.8	0.82	0.71	0.82
pred_ns	0.78	0.78	0.7	0.81

CPOP: TCGA–Sweden features

	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.64	0.71	0.64	0.67
pred_tcga	0.67	0.67	0.63	0.66
pred_array	0.78	0.83	0.68	0.7
pred_ns	0.8	0.78	0.68	0.74

Identity distance between predicted values under various models



Naive Lasso

	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.08	0.07	0.09	0
pred_tcga	0.06	0.11	0	0.09
pred_array	0.16	0	0.19	0.12
pred_ns	0	0.06	0.11	0.22

CPOP: NS–Array features

	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.04	0.04	0.03	0
pred_tcga	0.06	0.07	0	0.03
pred_array	0.01	0	0.07	0.04
pred_ns	0	0.01	0.08	0.06

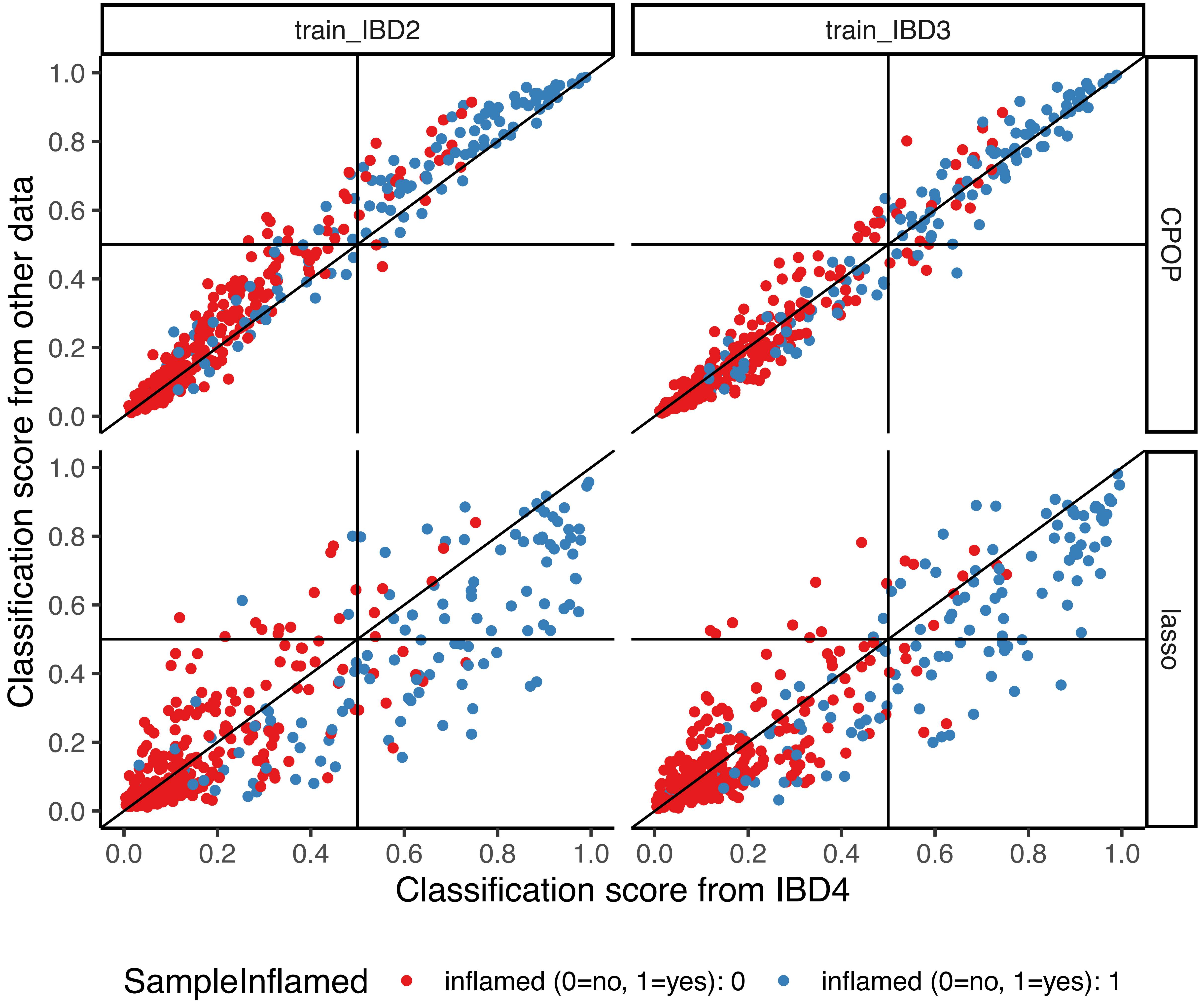
CPOP: TCGA–Sweden features

	train_ns	train_array	train_tcga	train_sweden
pred_sweden	0.04	0.04	0.02	0
pred_tcga	0.05	0.06	0	0.01
pred_array	0.04	0	0.09	0.07
pred_ns	0	0.06	0.08	0.07

CPOP results 2: prospective prediction

- CPOP on IBD NanoString data demonstrated improvements on stability

- We are planning to exploring other data of higher relevance to precision medicine (e.g. drug sensitivity)


train_IBD2
train_IBD3
CPOP
lasso
Classification score from other data
Classification score from IBD4
SampleInflamed
● inflamed (0=no, 1=yes): 0
● inflamed (0=no, 1=yes): 1