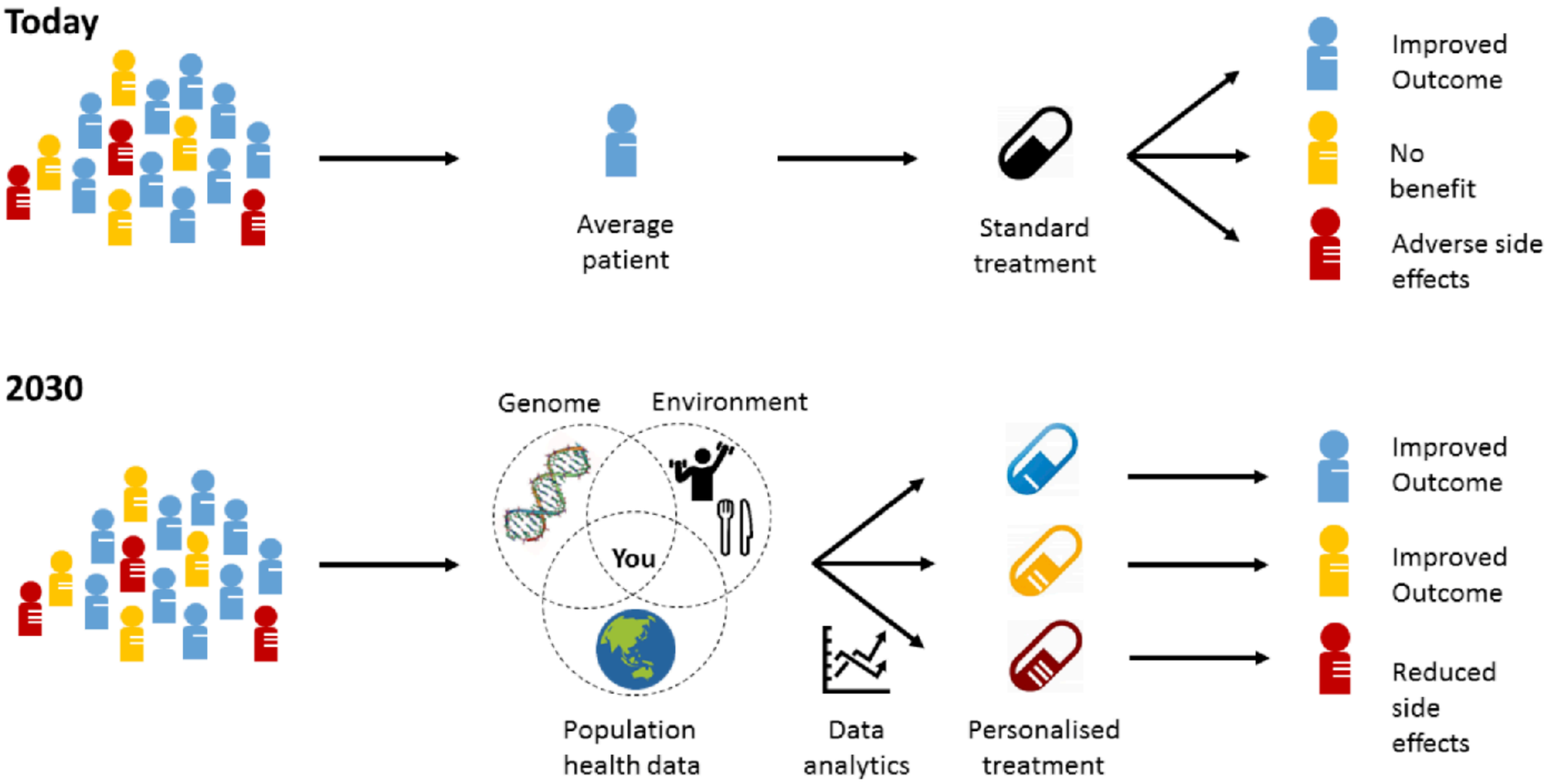
Kevin Wang

Cross-Platform Omics Prediction

Precision medicine: predicting best cause of action using omics data

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

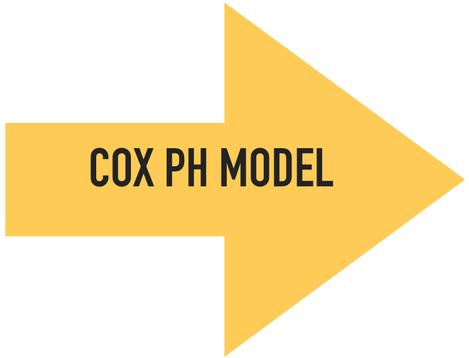
Precision medicine: predicting best cause of action using omics data



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

Stratification and classification using a risk score

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



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

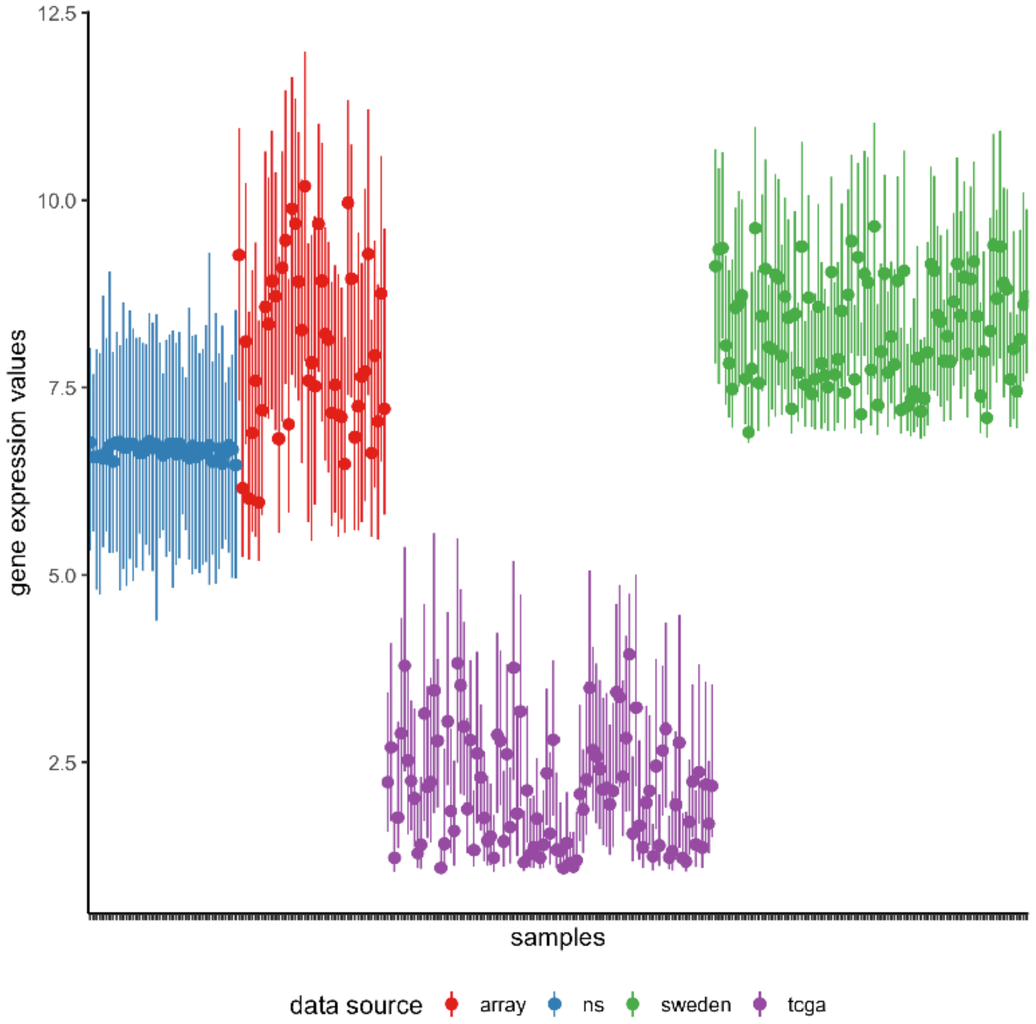
20 points model

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

Omics features are typically on a relative scale and unitless

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

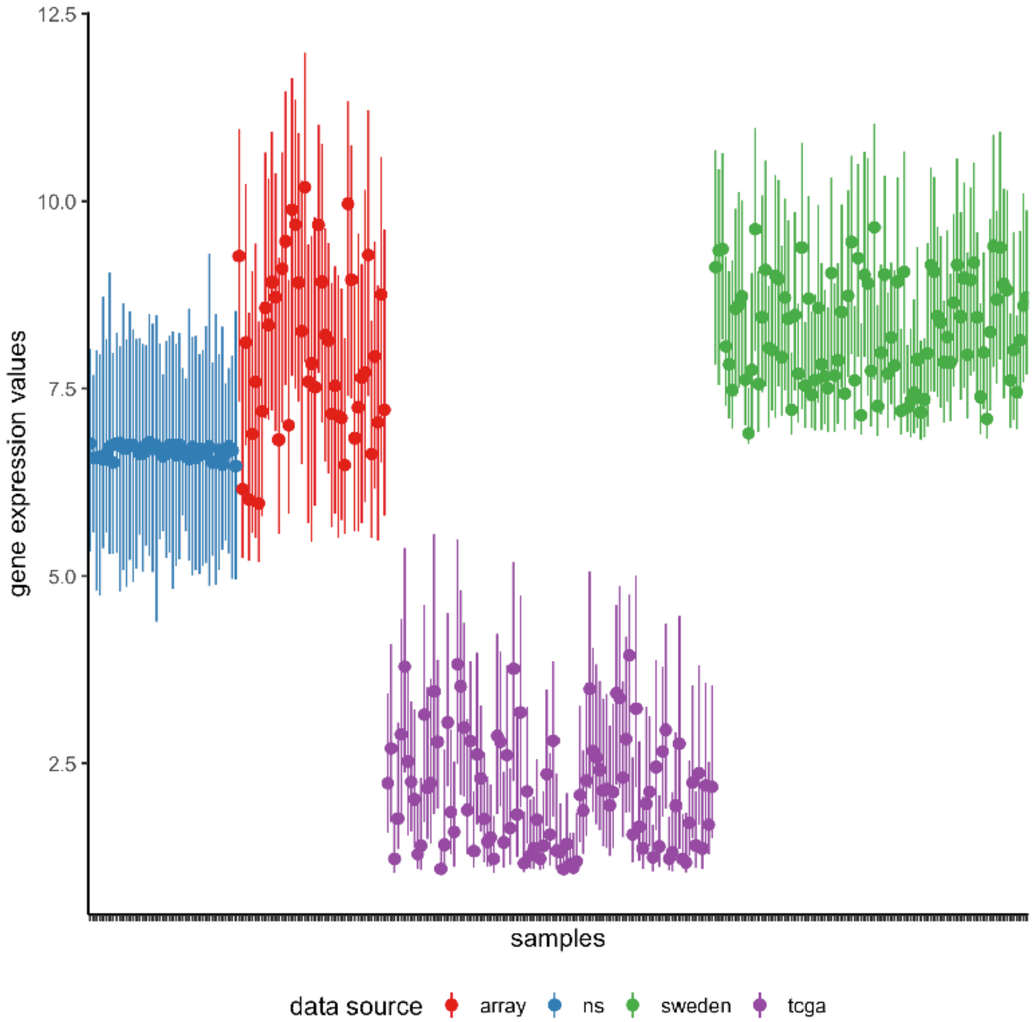
$$\hat{y}_2 = X_2\hat{\beta}_1 = (X_1 + \mathbf{1})\hat{\beta}_1$$



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

We cannot renormalise data in a clinical setting

	Sample 1	Sample 2	Sample 3	Sample 4
Gene 1	1.2	2.1	1.5	1.2
Gene 2	5.6	4.6	7.1	1.4
Gene 3	9.2	10.1	6.9	8.6
Gene 4	4.1	3.6	2.7	7.1



The flowchart of a clinical risk score

Data

$$(X_1, y_1)$$

$$(X_2, y_2)$$

Model

$$\hat{\beta}_1$$

Prediction

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

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

The flowchart of a clinical risk score

Data

$$(X_1, y_1)$$
$$(X_2, y_2)$$

Model

$$\hat{\beta}_1$$

Prediction

$$\hat{y}_1 = X_1\hat{\beta}_1$$
$$\hat{y}_2 = X_2\hat{\beta}_1$$

No renormalisation

No model retraining

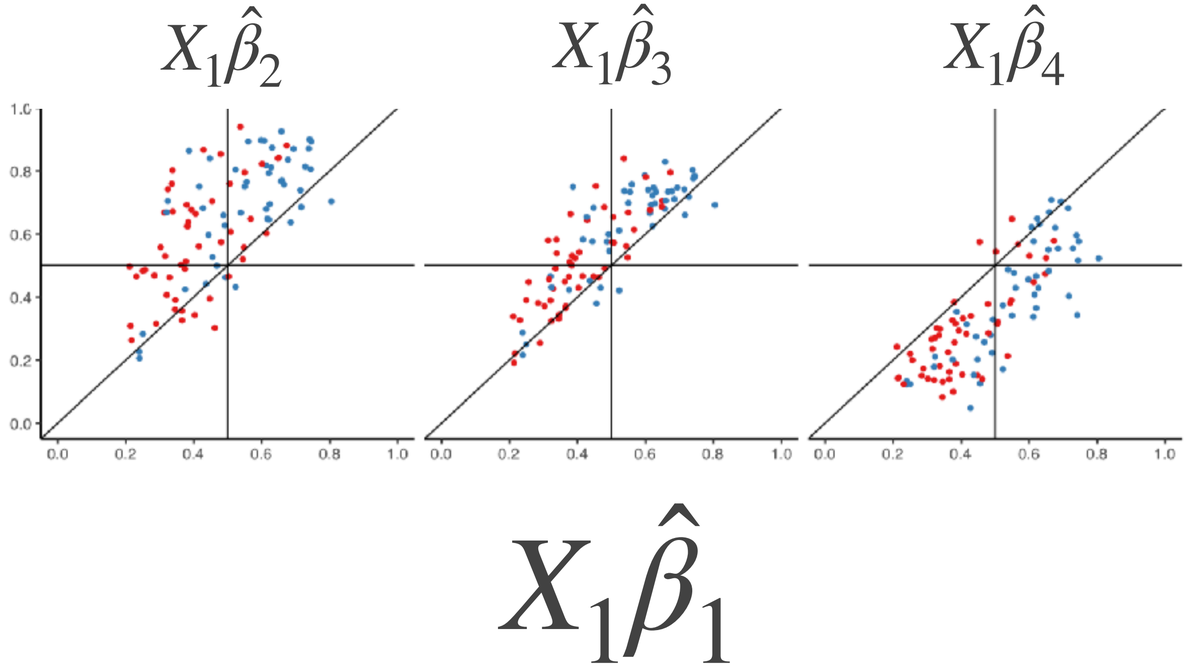
Scale-equivalent prediction

Statistical challenges

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

Transferability

The prediction on one gene expression platform should be equivalent to another platform



First component of CPOP: feature enginnering

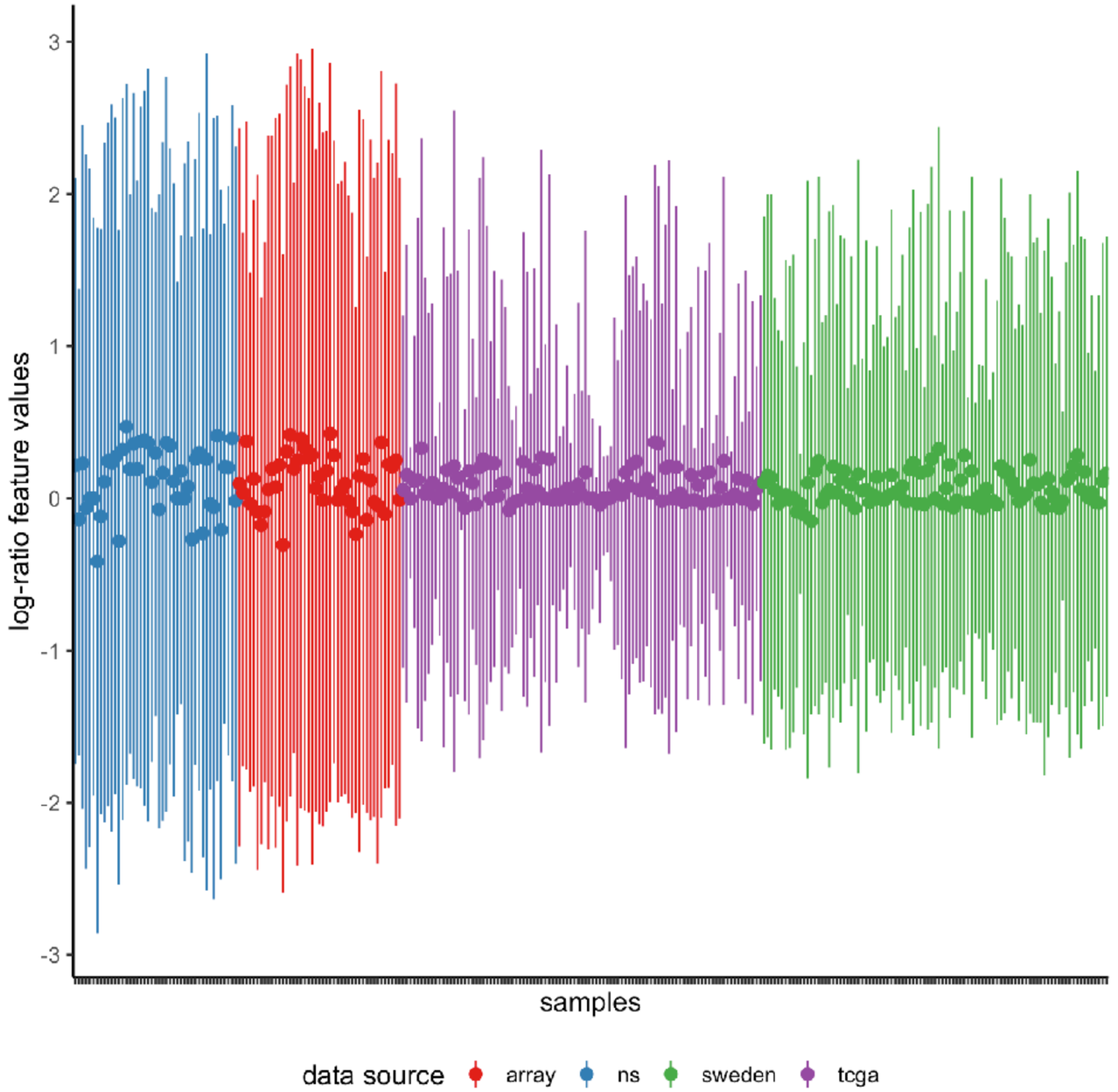
就让我 来次透彻心扉的痛
都拿走 让我再次两手空空
只有奋奋一息过
那个真正的我
他才能够诞生

Within-sample feature standardisation

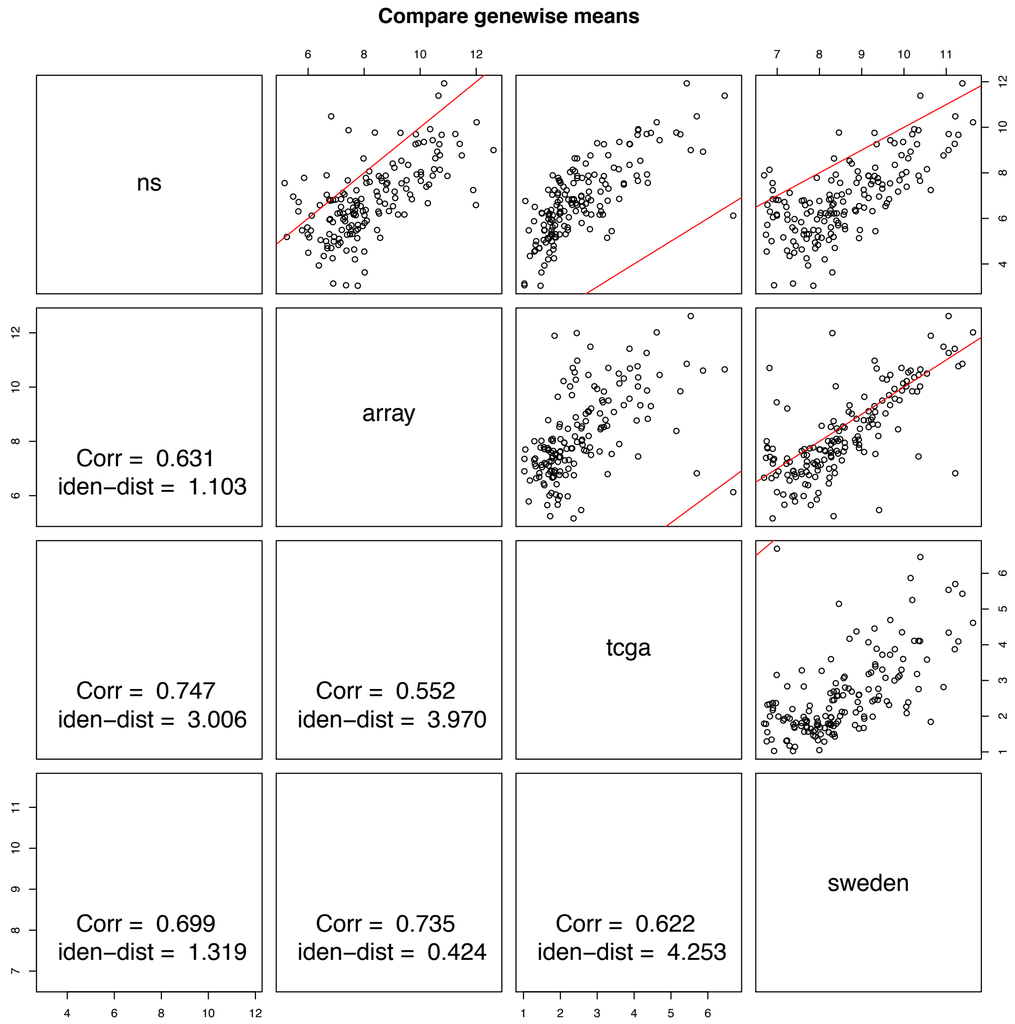
Single-patient prediction prevents us from calculating any cross-sample statistics

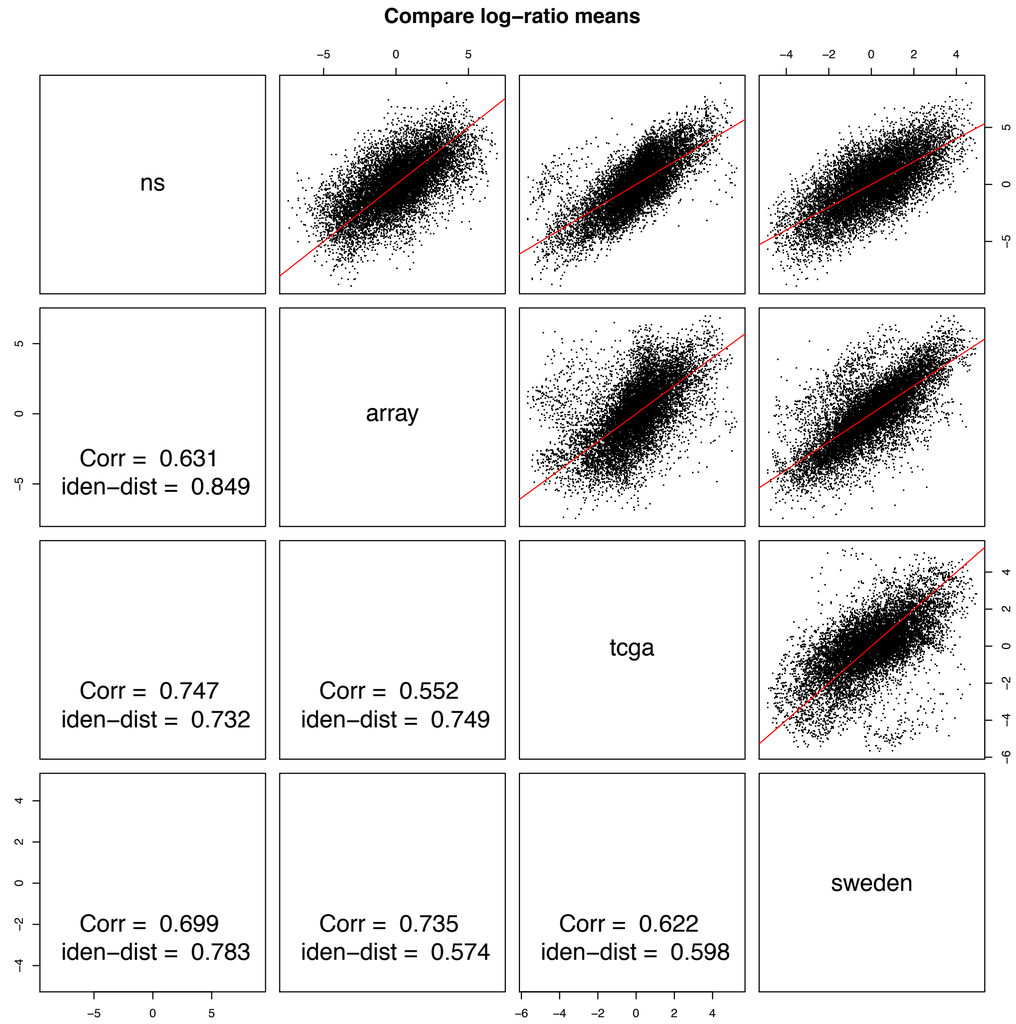
Log-ratio

log(gene A) – log(gene B)



Within-sample feature standardisation





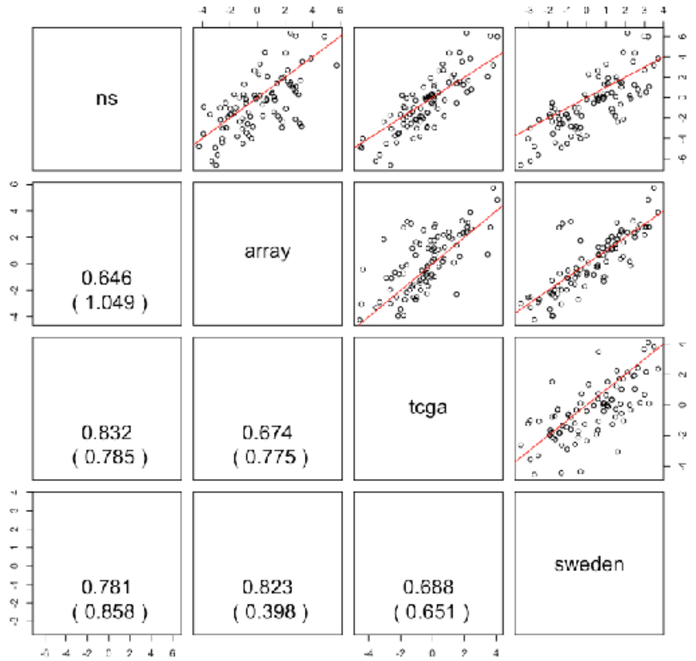
The solution is trivial?

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~~

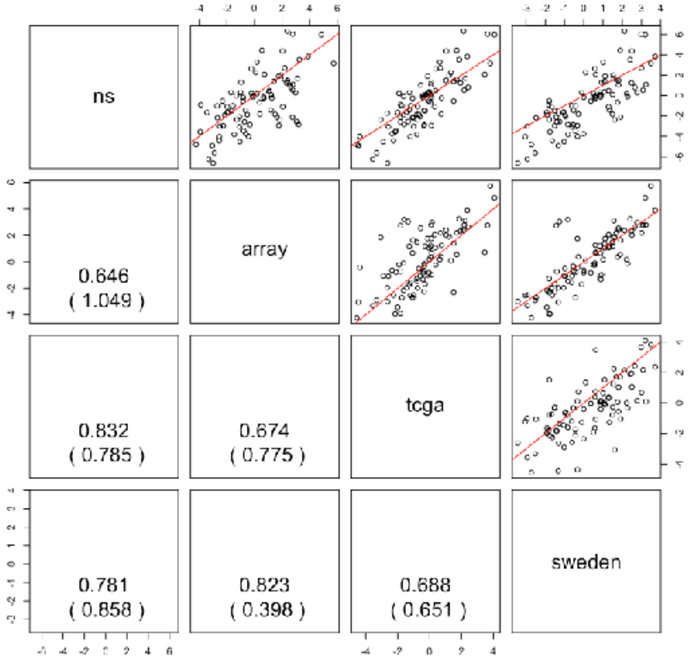
Concordance of log-ratios after Lasso selection

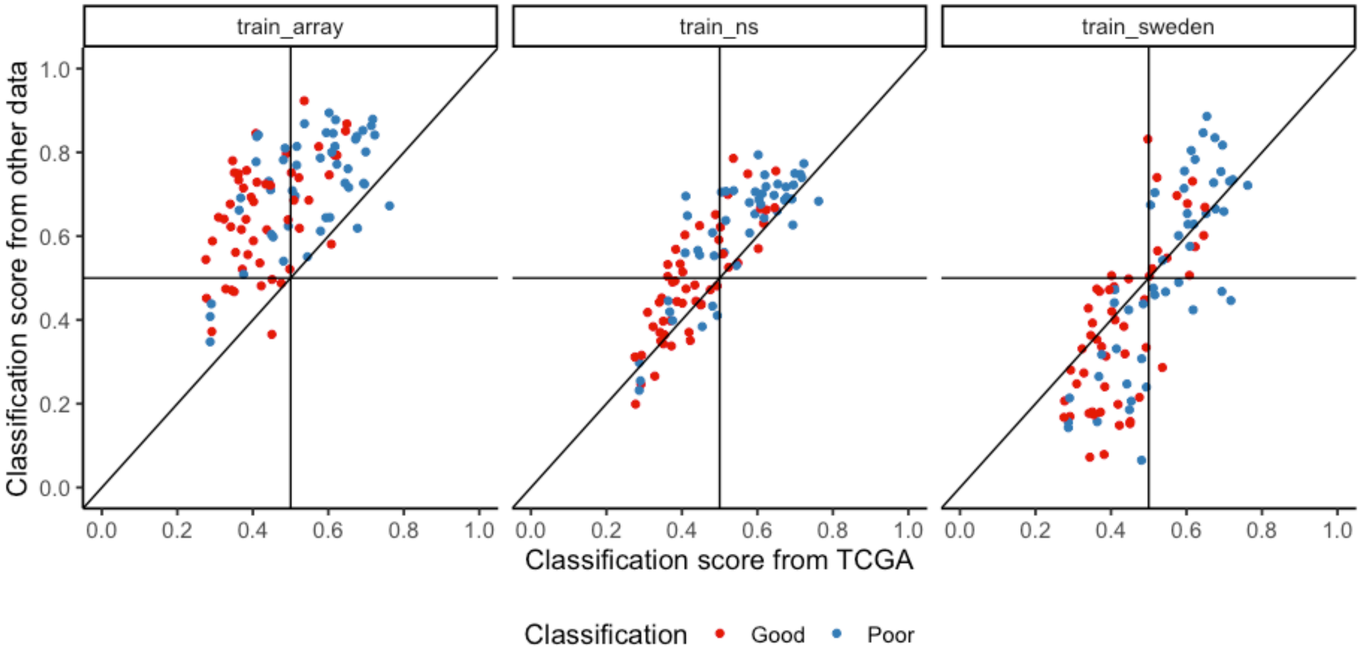


The solution is trivial?

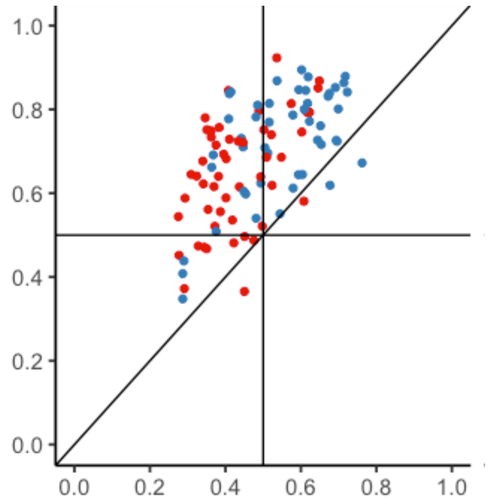
1. Concordance in log-ratio features scaling across platforms
2. Concordance in feature selection and coefficient estimates
3. ~~Single-patient prediction~~

Concordance of log-ratios after Lasso selection

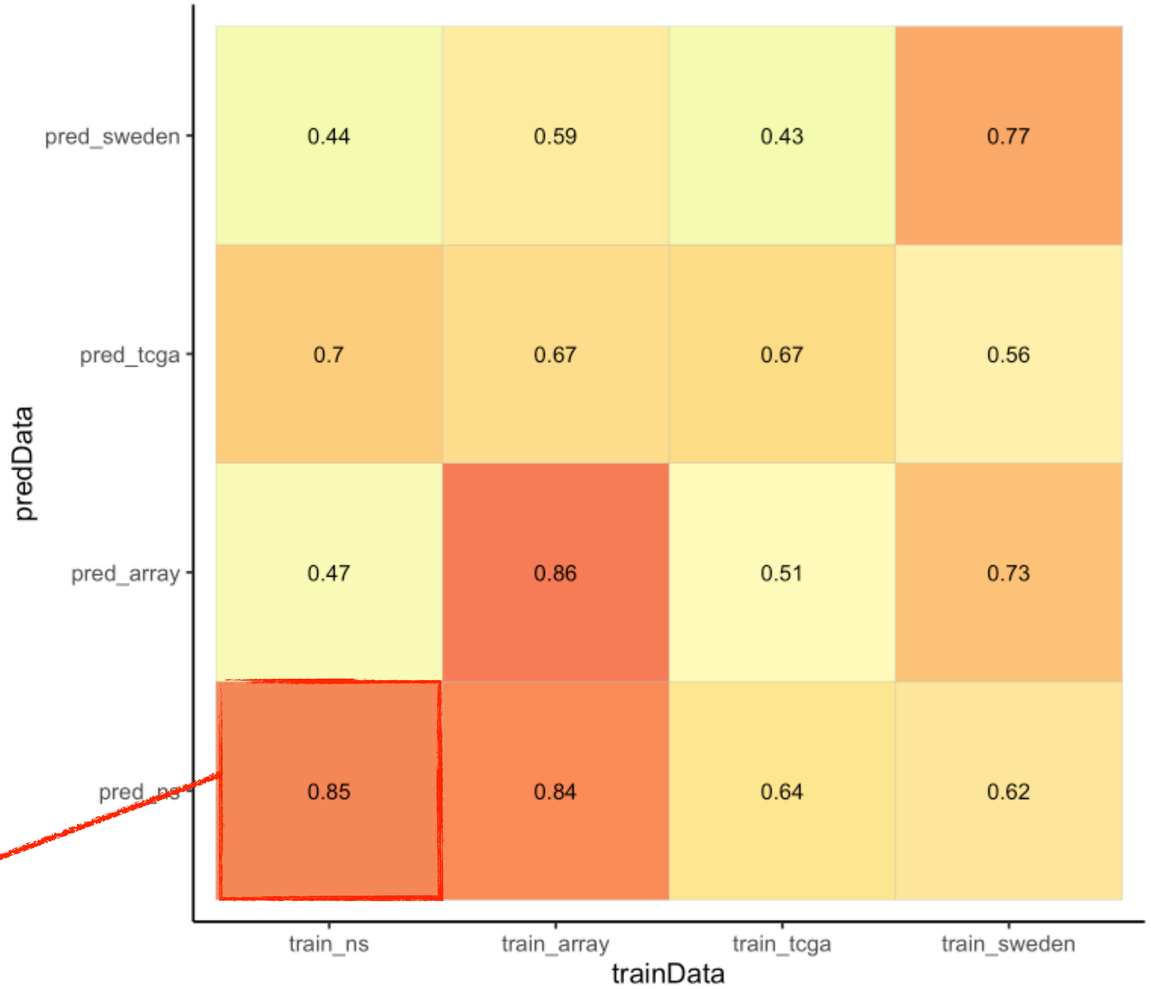




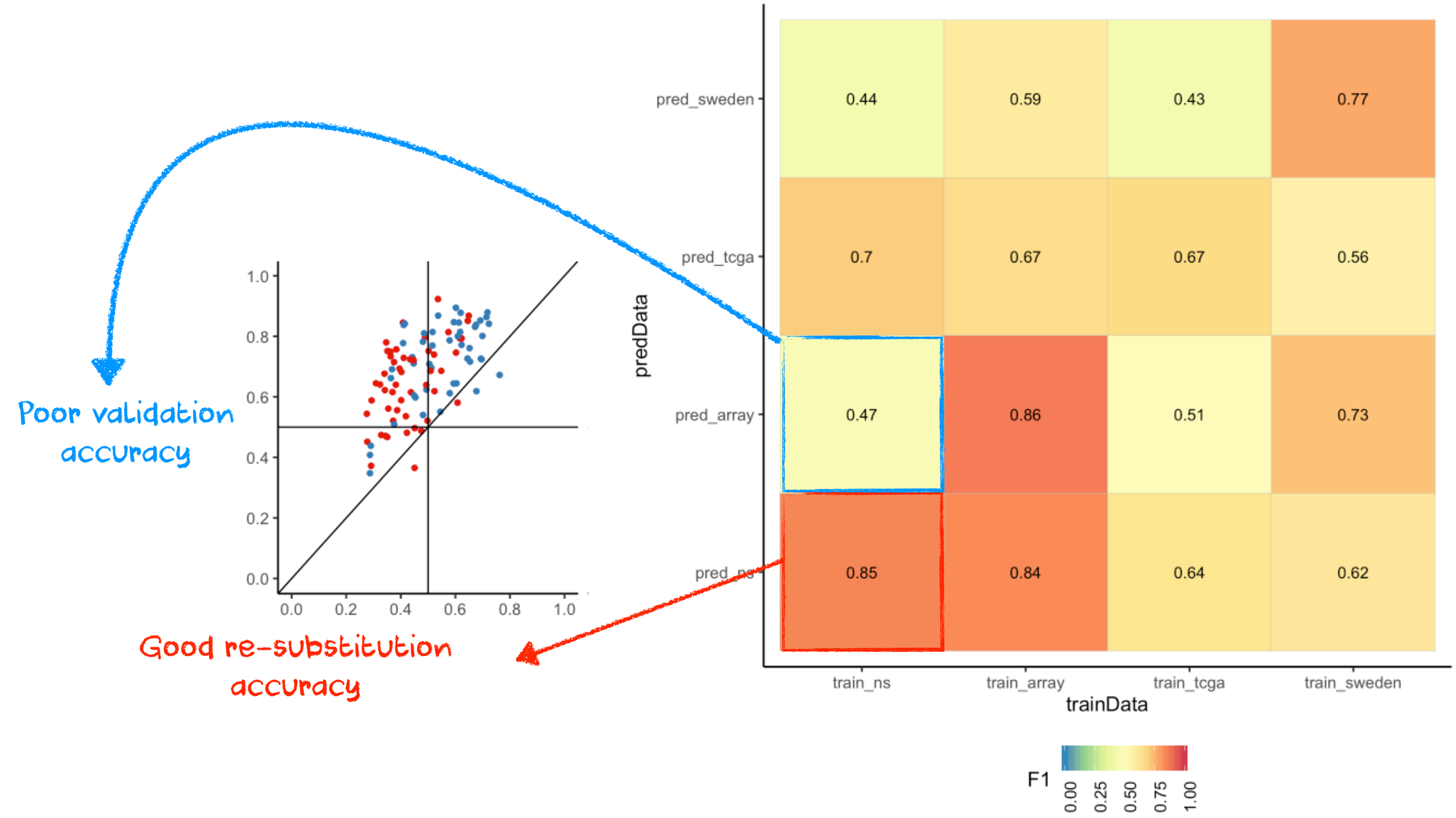
The solution is not so trivial



Good re-substitution accuracy



The solution is not so trivial



The solution is not so trivial

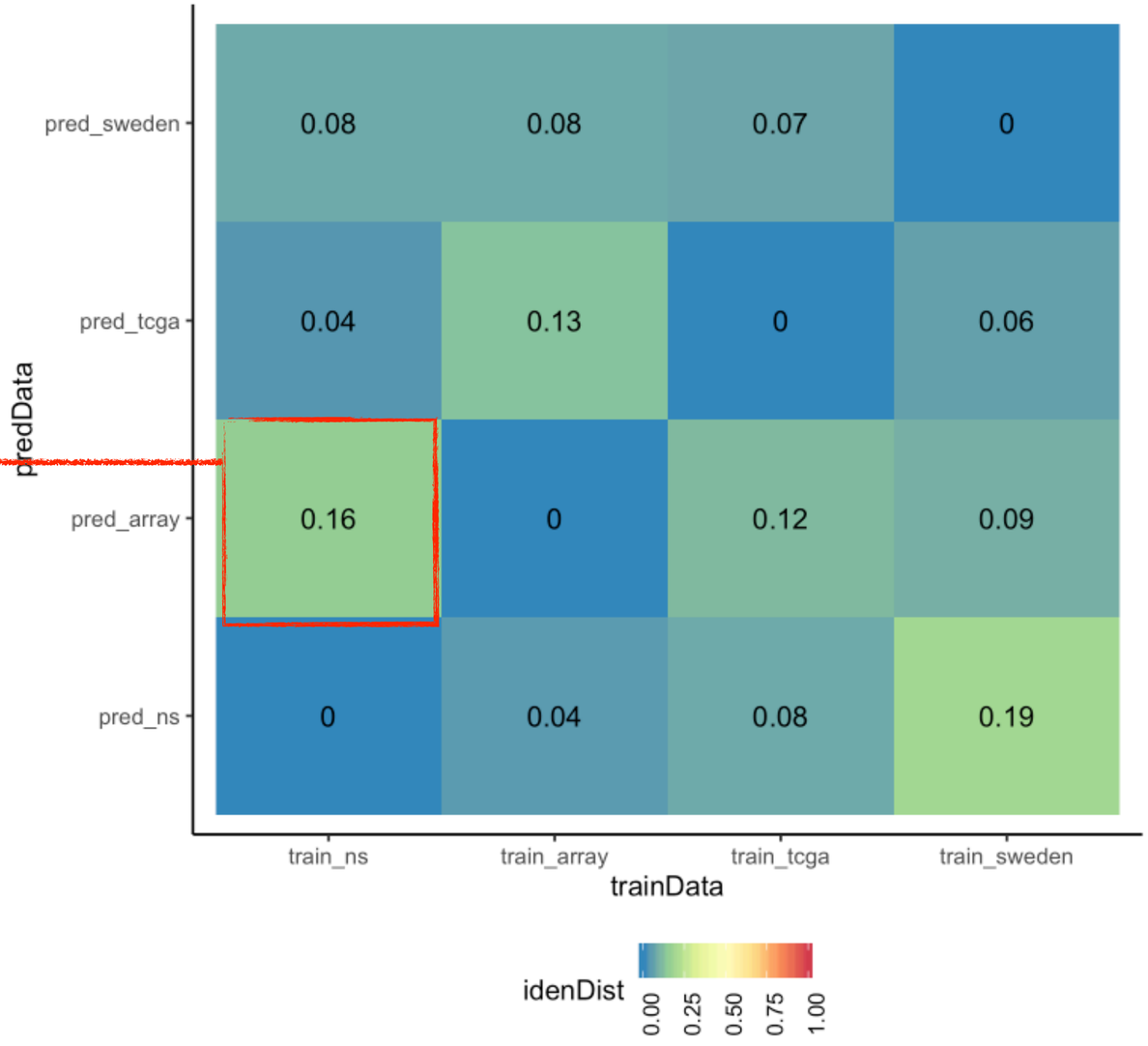
Estimated prognosis probabilities from

training data

vs

validation data

differ by 0.16 on average



And as you know, I always put jokes in my slides.

The past 9 months when I had to mentally deal with this problem was the most stressful time I had.

And during that time, for every presentation that I did, I inserted subtle references to my favourite Chinese music so that I can still mentally do my day-to-day jobs.

In this particular slide, this is a song written by one of my favourite musicians on the subject of overcoming his own depression for over a decade.

So now you will know this easter egg that I hide in all my presentations.

Second component of CPOP: feature selection

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

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

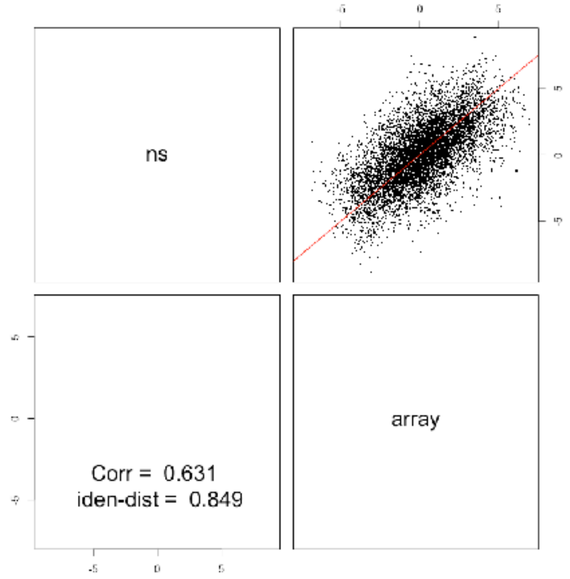
means

$Z_1 \approx Z_2$ column-wise

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

CPOP weighted variable selection

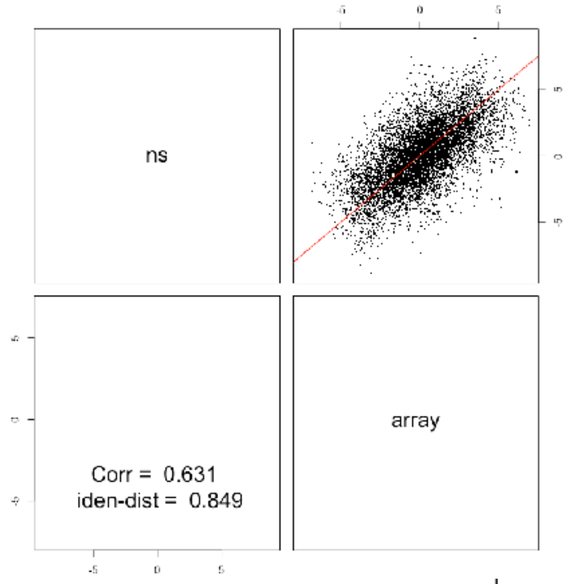
1. Perform a **weighted Lasso** by placing higher weights on features closer to the identity line



$$Z_1 \approx Z_2$$

CPOP weighted variable selection

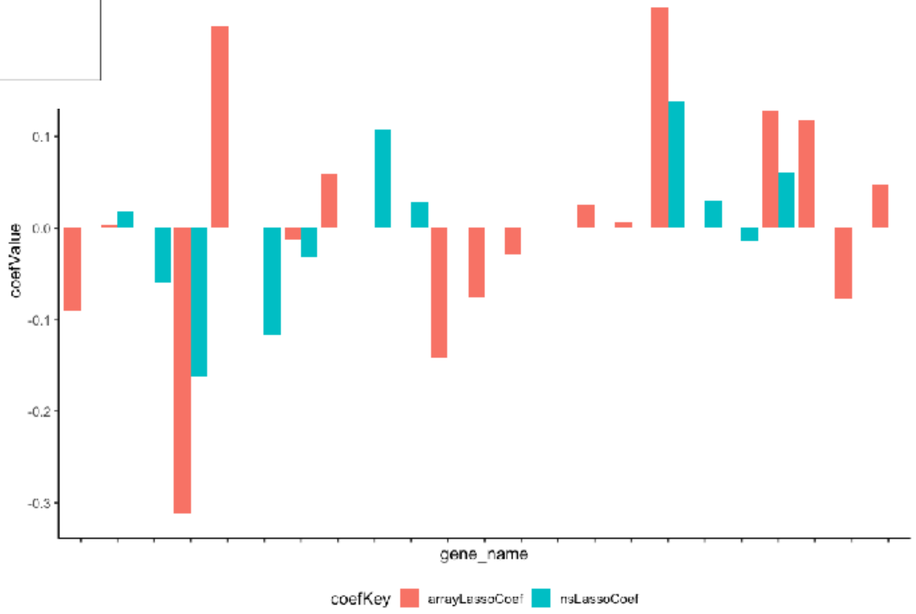
1. Perform a **weighted Lasso** by placing higher weights on features closer to the identity line



$$Z_1 \approx Z_2$$

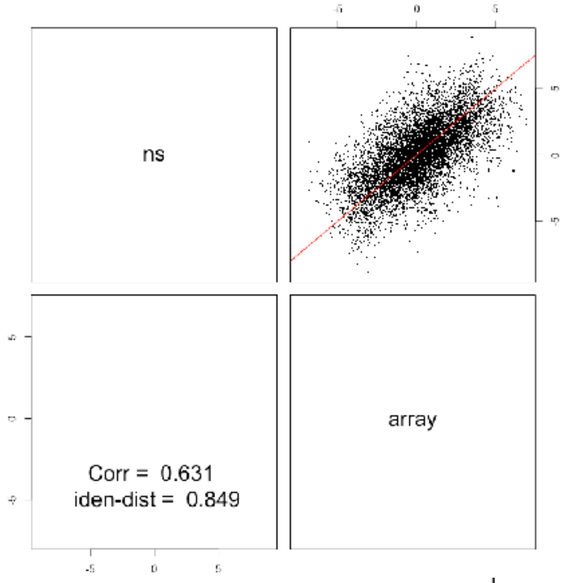
2. Perform a **Ridge regression** and only retain those features with coefficients similar to each other

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



CPOP weighted variable selection

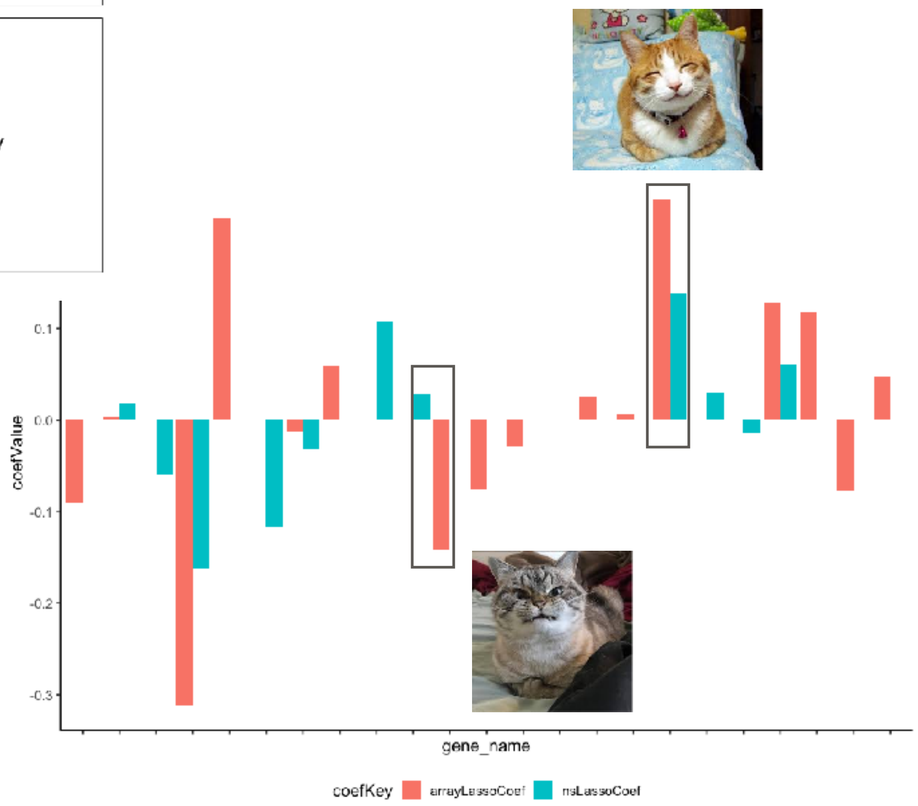
1. Perform a **weighted Lasso** by placing higher weights on features closer to the identity line



$$Z_1 \approx Z_2$$

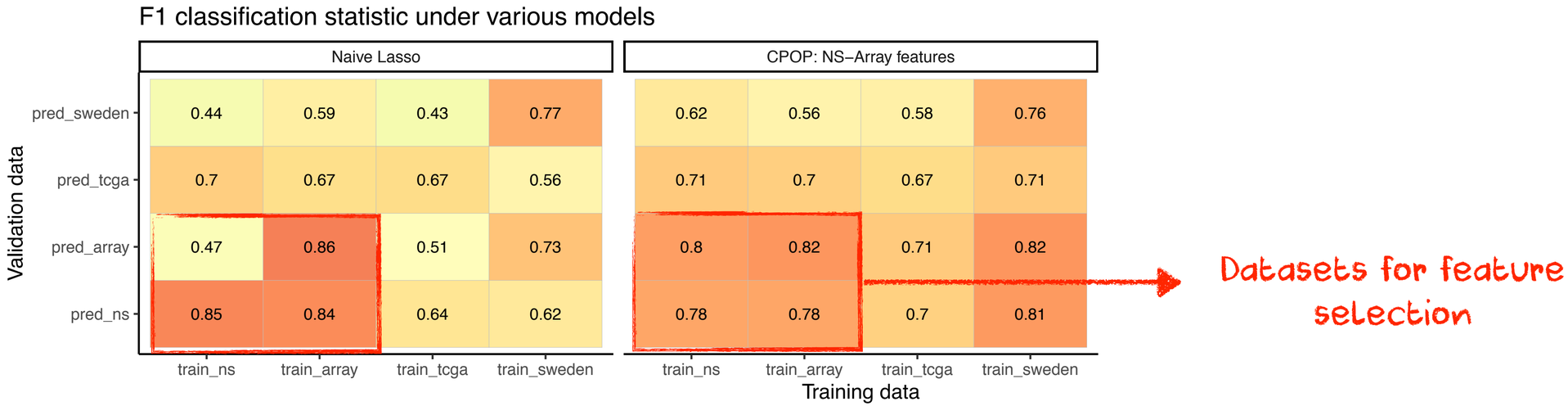
2. Perform a **Ridge regression** and only retain those features with coefficients similar to each other

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

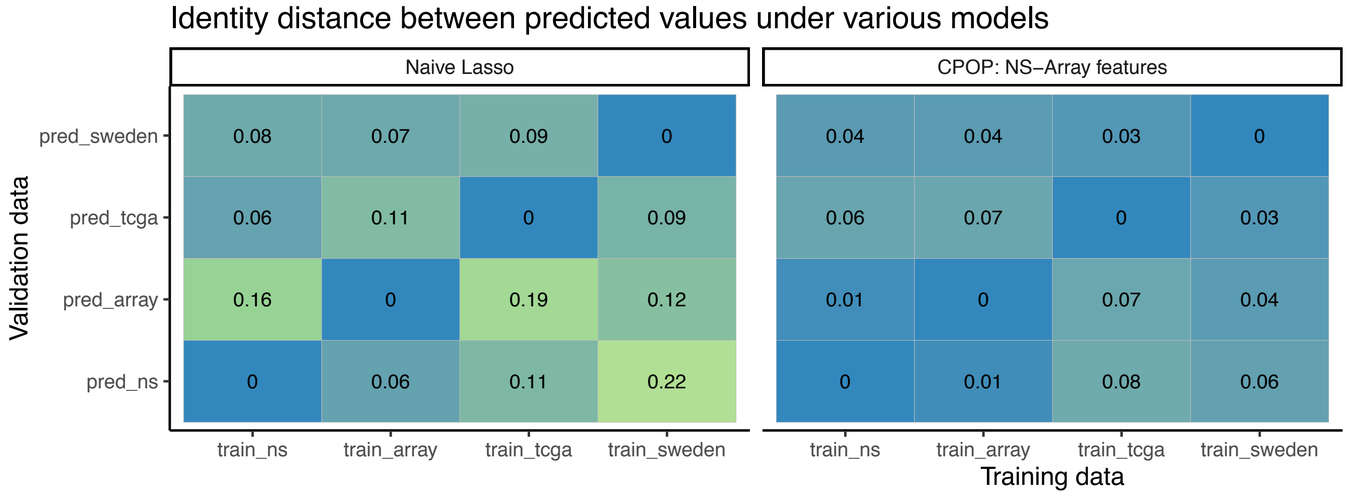


CPOP results 1: four melanoma data

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

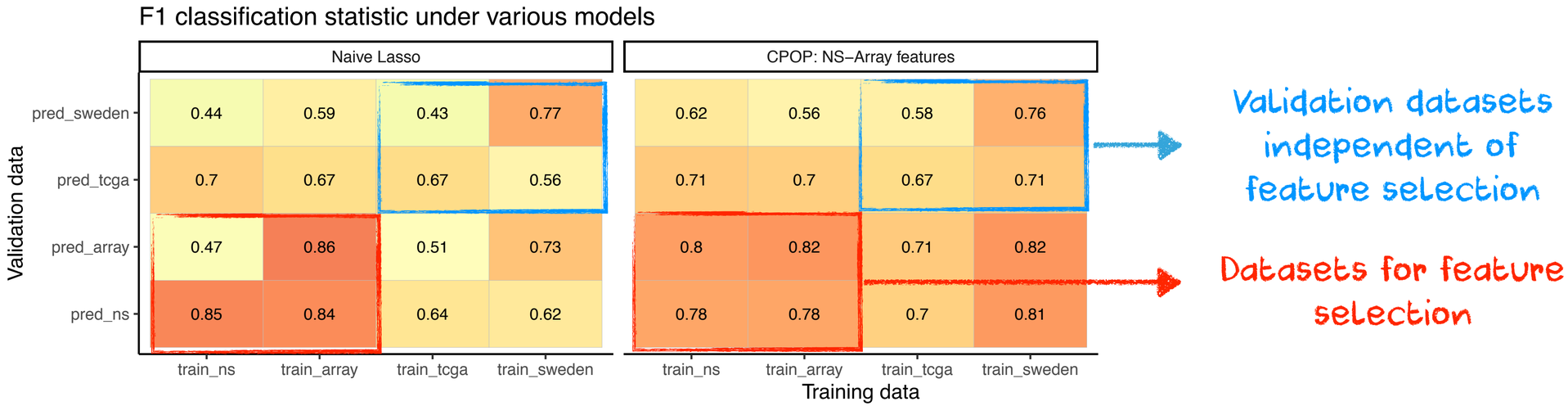


2. Smaller identity distance between predicted values

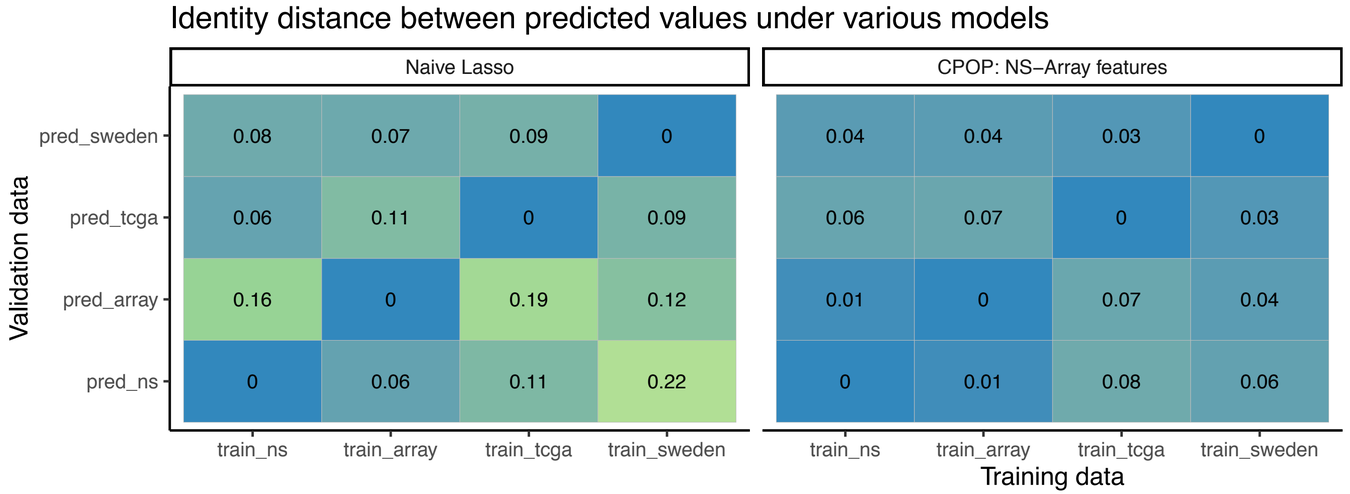


CPOP results 1: four melanoma data

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

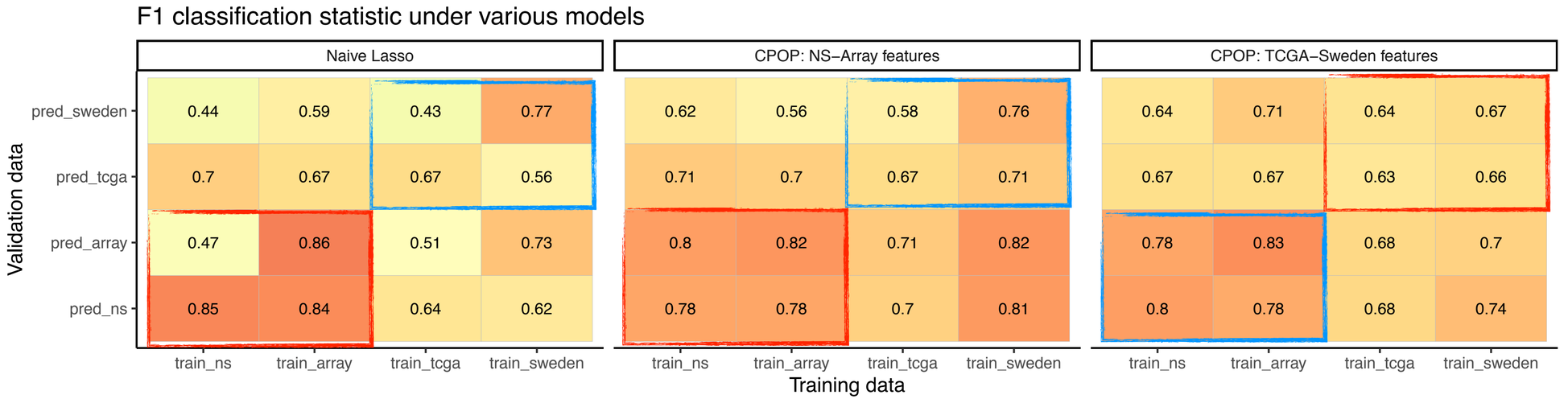


2. Smaller identity distance between predicted values



CPOP results 1: four melanoma data

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

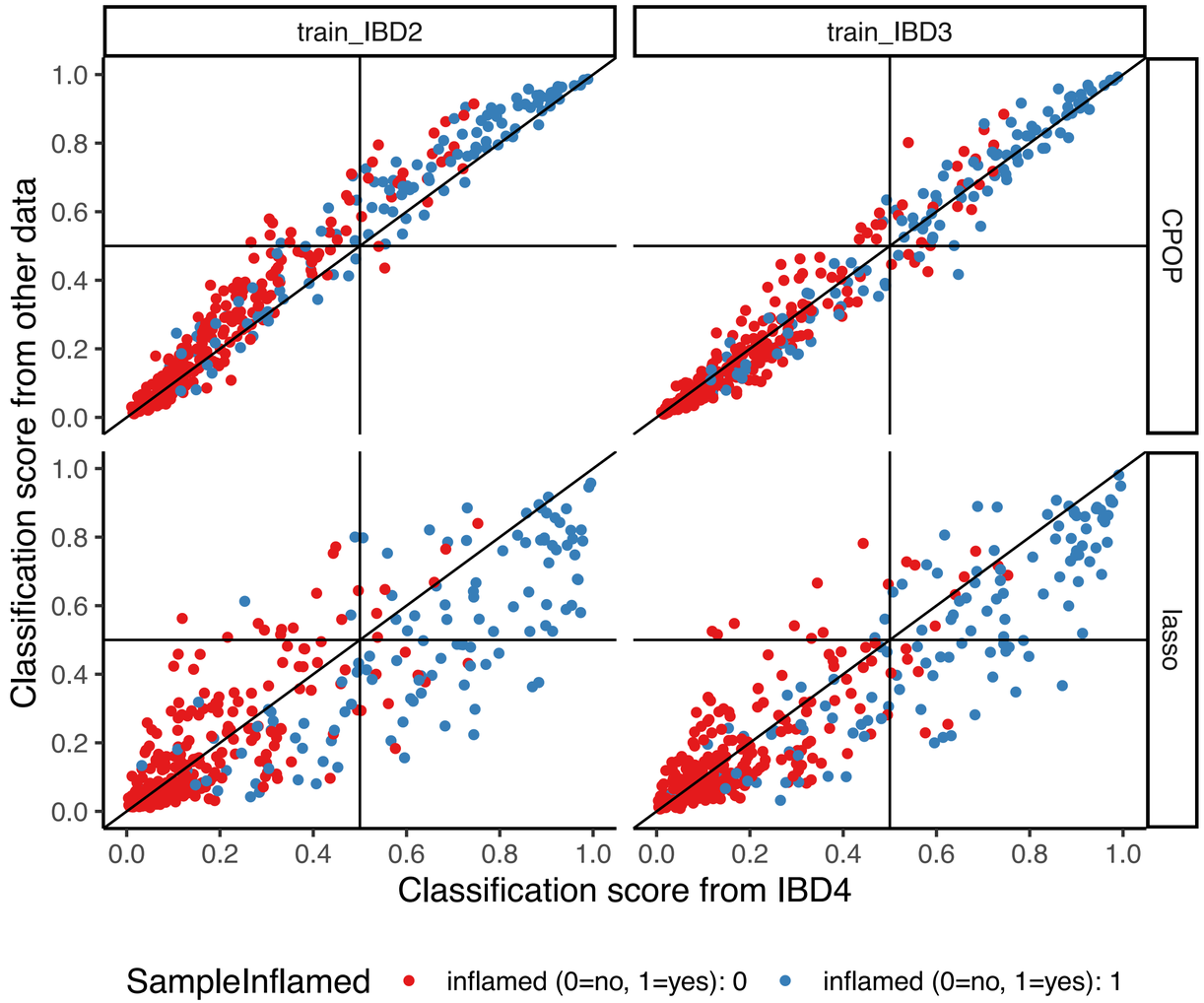


2. Smaller identity distance between predicted values

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)



Concluding remarks

- CPOP is a flexible procedure that allows for:
  - cross-platform omics prediction
  - stable single-patient prediction
- Not everyone can smooth-sailing through the PhD process, find your own way to deal with it (e.g. write your frustrations in white text in thesis)

But what about breast cancer?

Name	Predictors	Targets	Prediction	Technology	Legit?
Oncotype DX	21 genes	ER +	Score	qRT-PCR	ASCO, NCCN
Prosigna	50 genes	Hormone receptor +	Score	NanoString	FDA 510k
MammaPrint	70 genes	Any ER status	Binary	DNA microarray	FDA

- Alvarado et. al. (2015) reported poor concordance in the prediction scores
- Hyeon et. al. (2017) considered NanoString as a viable alternative to RT-PCR

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

1. Omics features are typically on a relative scale and unitless

	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 solution is trivial.

- We have “standardised features” within every patient to build models

Log-ratio

log(gene A) – log(gene B)