Kevin Wang

JB Douglas Award - USYD internal talk

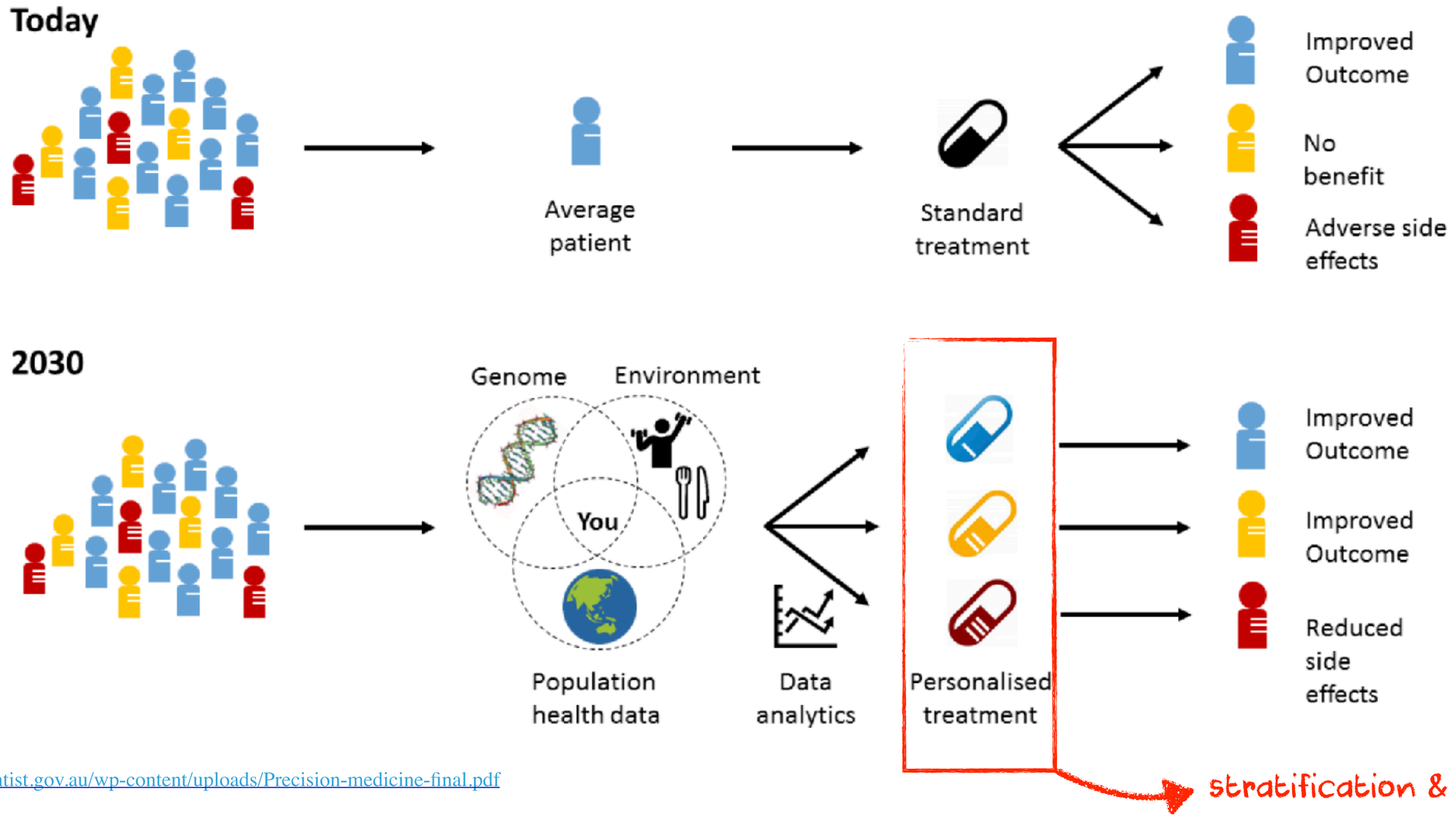
Cross-Platform Omics Prediction: a step towards precision medicine

21 Oct 2019

Precision medicine: predicting best cause of action using omics data

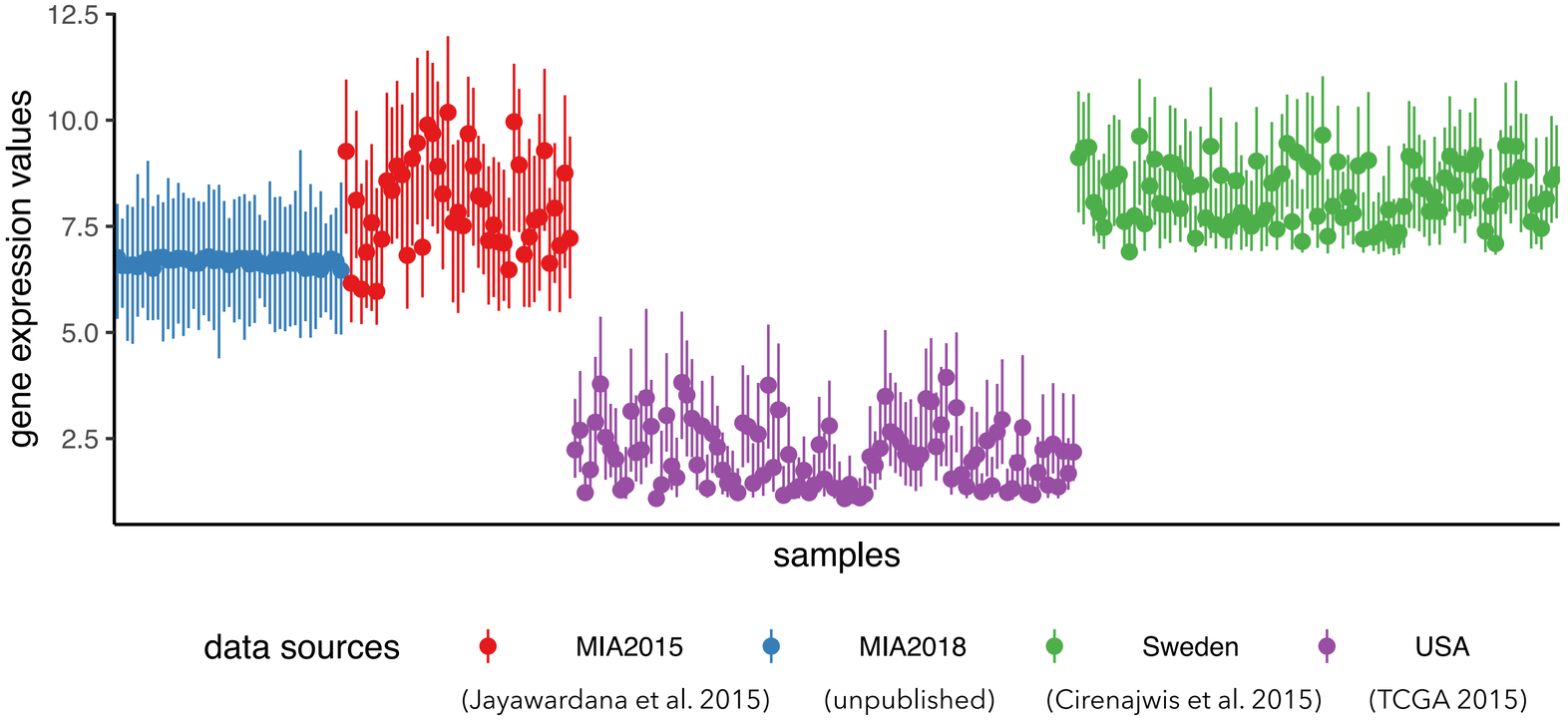
https://www.chiefscientist.gov.au/wp-content/uploads/Precision-medicine-final.pdf

Precision medicine: predicting best cause of action using omics data



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

Melanoma Institute Australia



- Risk score using gene expression

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

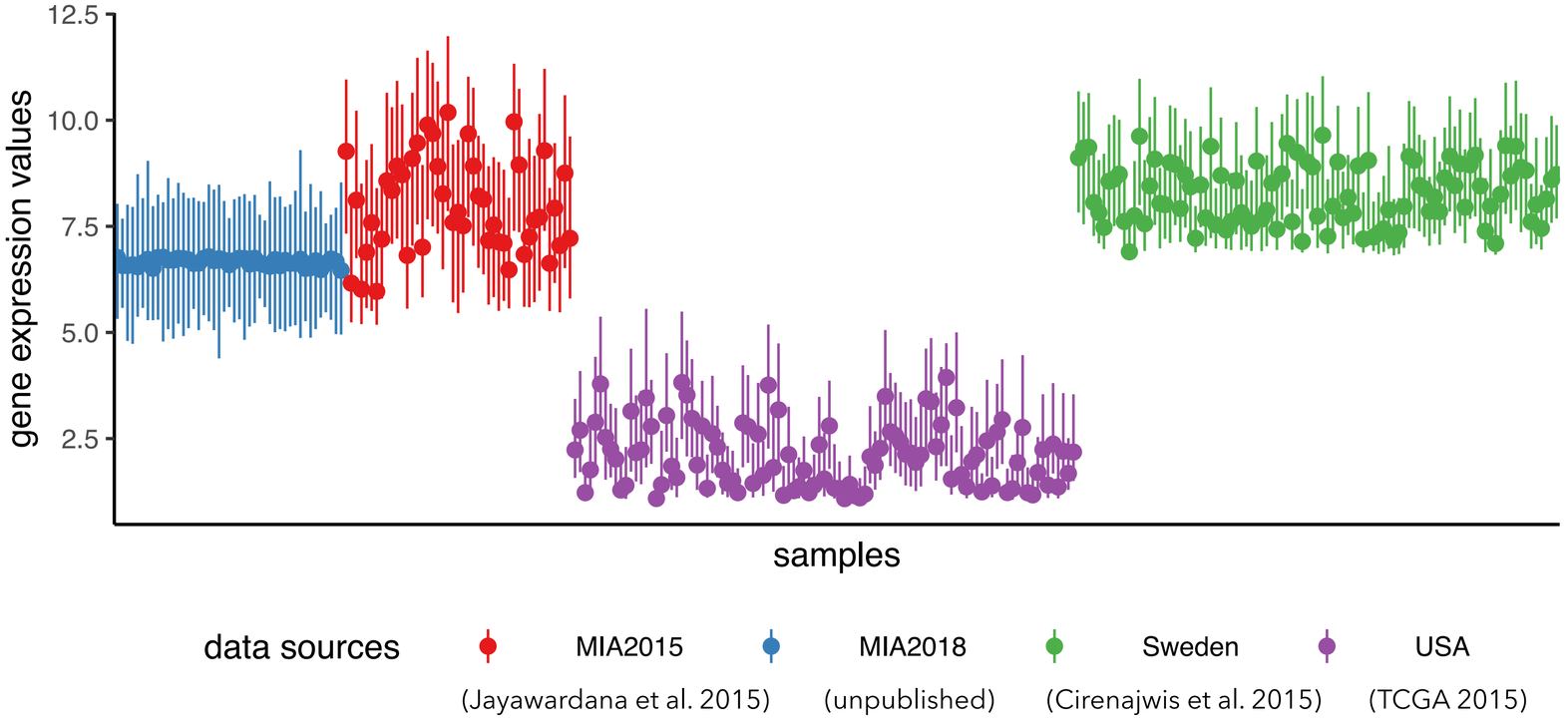
- CRE grant will support implementation
  - prospective
  - **multi-centres**

Melanoma Institute Australia

Prof. Graham Mann

https://researchdata.ands.org.au/centre-research-excellence-outcomes-melanoma/1318231

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



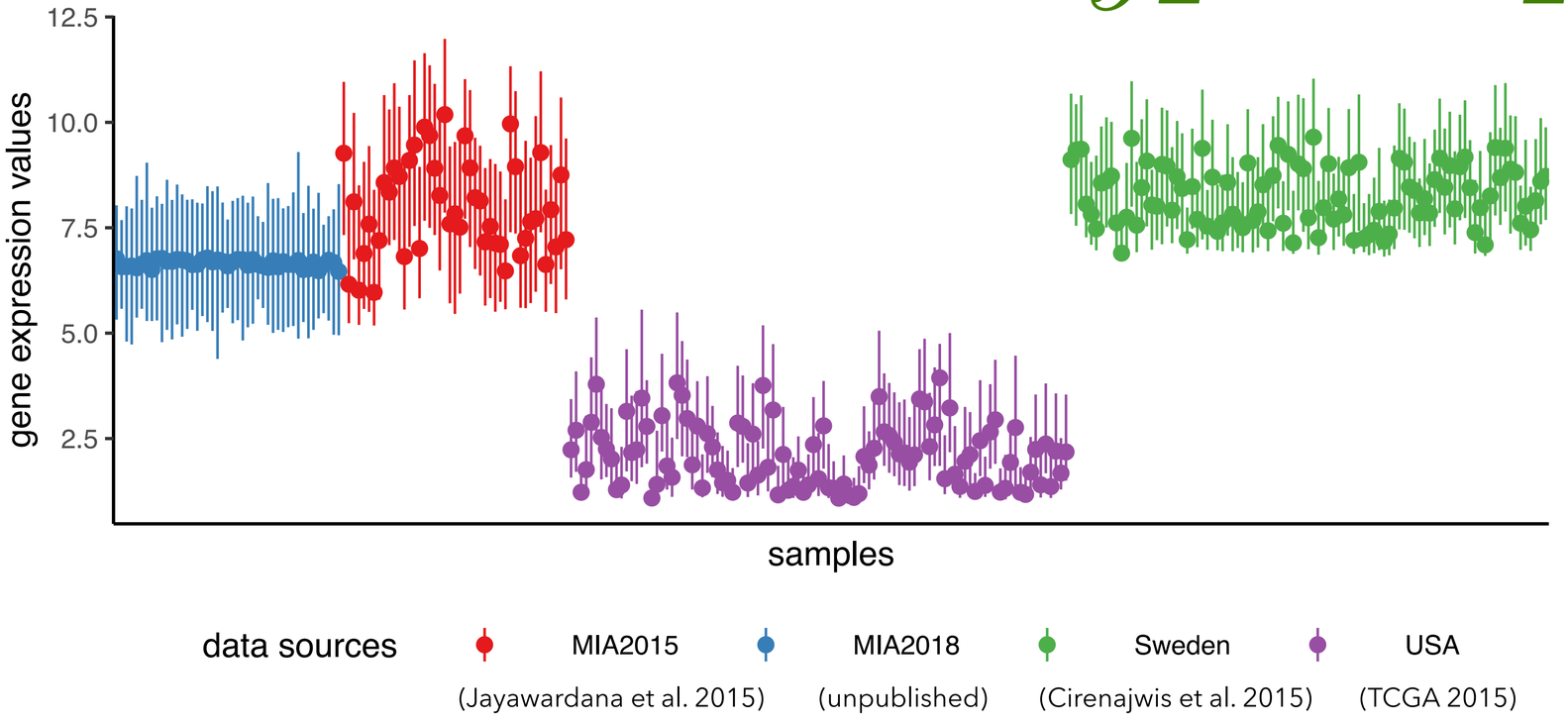
Gene expression features on a unit-less scale.

A typical value in one data can be an impossible value on another.

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

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

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



Gene expression features on a unit-less scale.

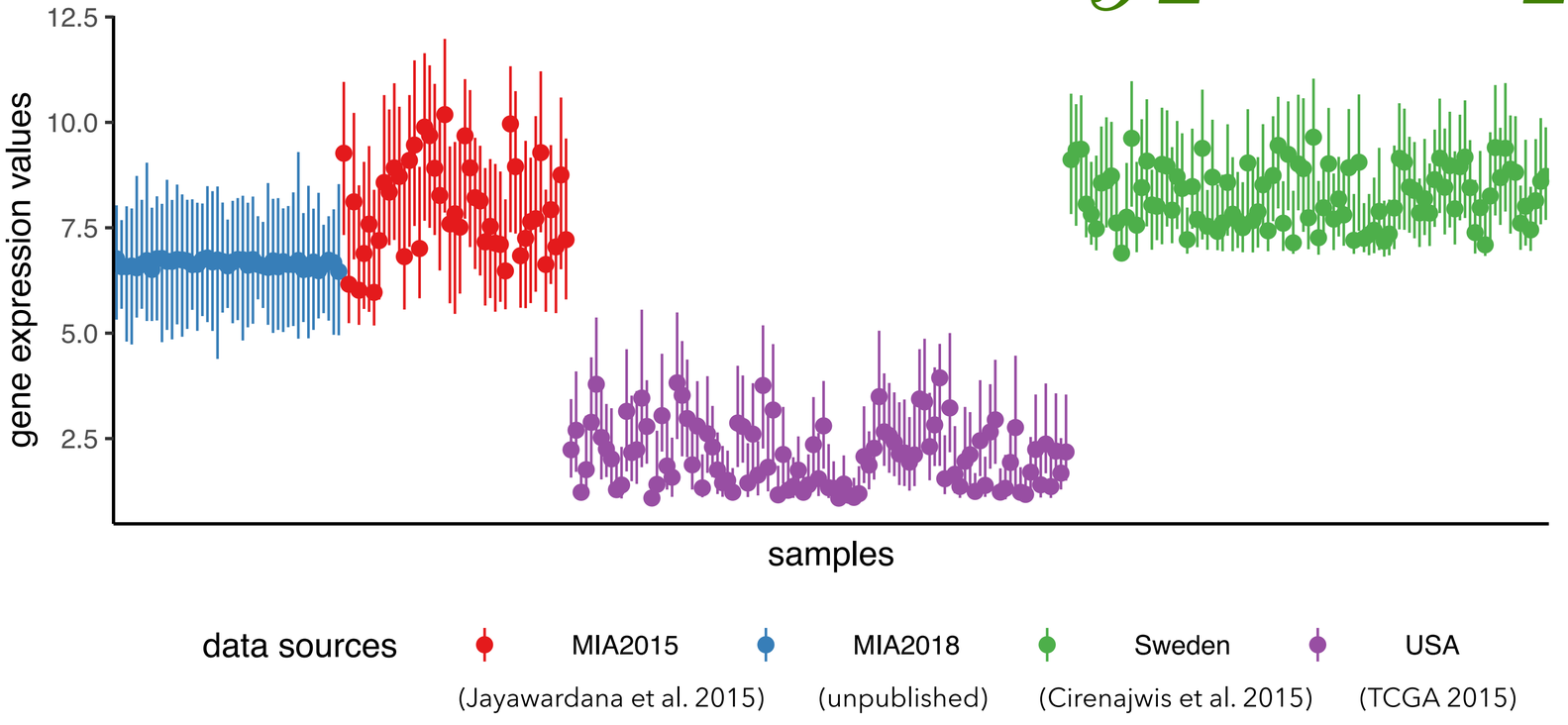
A typical value in one data can be an impossible value on another.

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

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

Assuming a noiseless shift

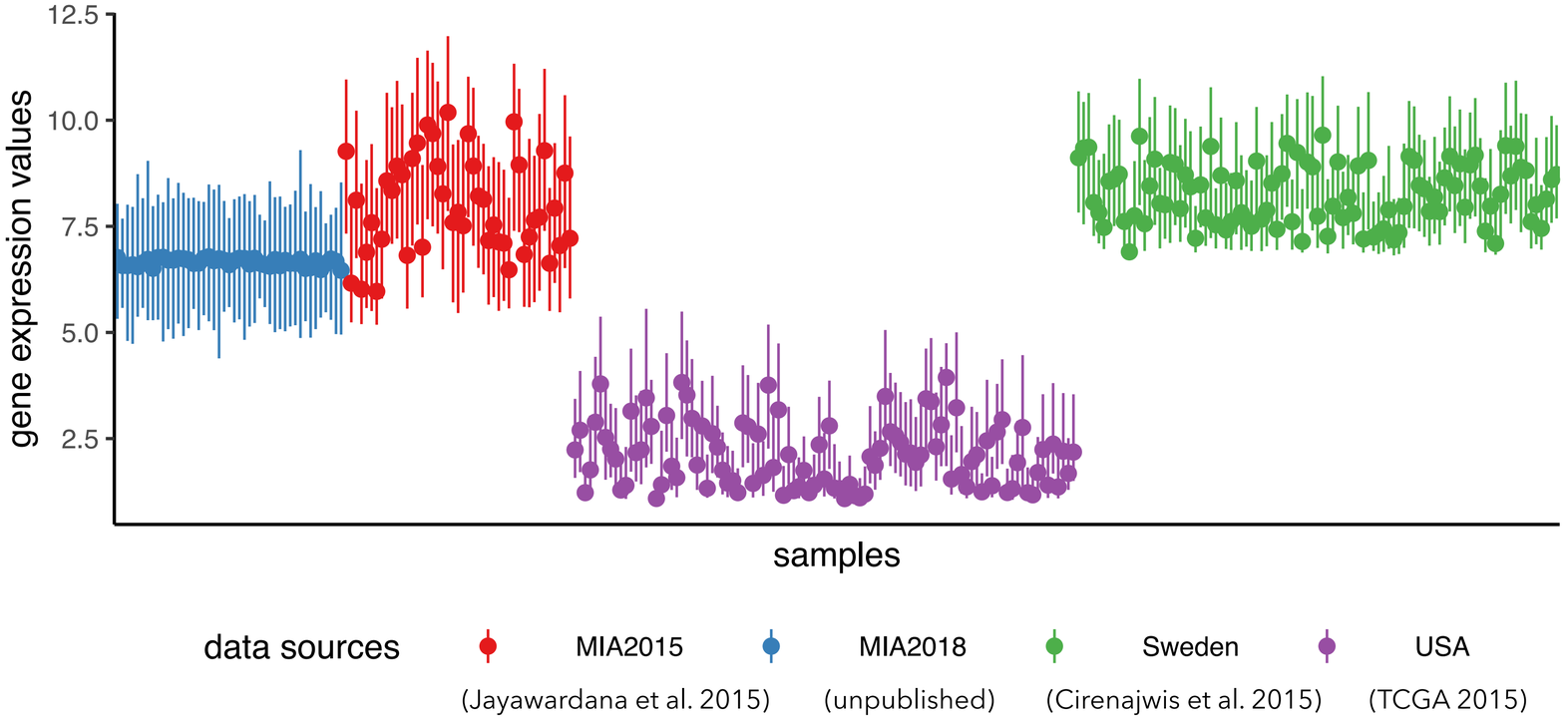
$$\hat{y}_2 = X_2 \hat{\beta}_1 = (X_1 + \mathbf{1}) \hat{\beta}_1$$



Gene expression features on a unit-less scale.

A typical value in one data can be an impossible value on another.

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



Transferability

A model trained from one data and should be predictive on another

Existing approaches

Components of a risk score

Data

$$(X_1, y_1)$$

$$(X_2, y_2)$$

Model

$$\hat{\beta}_1$$

Prediction

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

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

Data-harmonisation

Data	Model	Prediction
(X_1, y_1)	$\hat{\beta}_1$	$\hat{y}_1 = X_1\hat{\beta}_1$
(X_2, y_2)		$\hat{y}_2 = X_2\hat{\beta}_1$
Harmonisation	**Classical estimation methods**	**Classical model prediction**

Data-harmonisation: normalisation or standardisation

Data

(X_1, y_1)

(X_2, y_2)

Harmonisation

1. Prospective: re-normalisation upon new single-samples
2. Multi-centres: re-training of model upon new populations

Prevented by ethics, privacy and data security

Clinical constraints in implementation

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 re-normalisation

No model re-training

Scale-equivalent prediction

CPOP: predictions that respect clinical constraints

CPOP flowchart

Data

$$(X_1, y_1) \rightarrow (Z_1, y_1)$$

$$(X_2, y_2) \rightarrow (Z_2, y_2)$$

Feature transform

Model

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

Stable estimation

Prediction

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

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

Stable prediction

First component of CPOP: feature transform

Log-ratio transformation: modelling relative gene expression

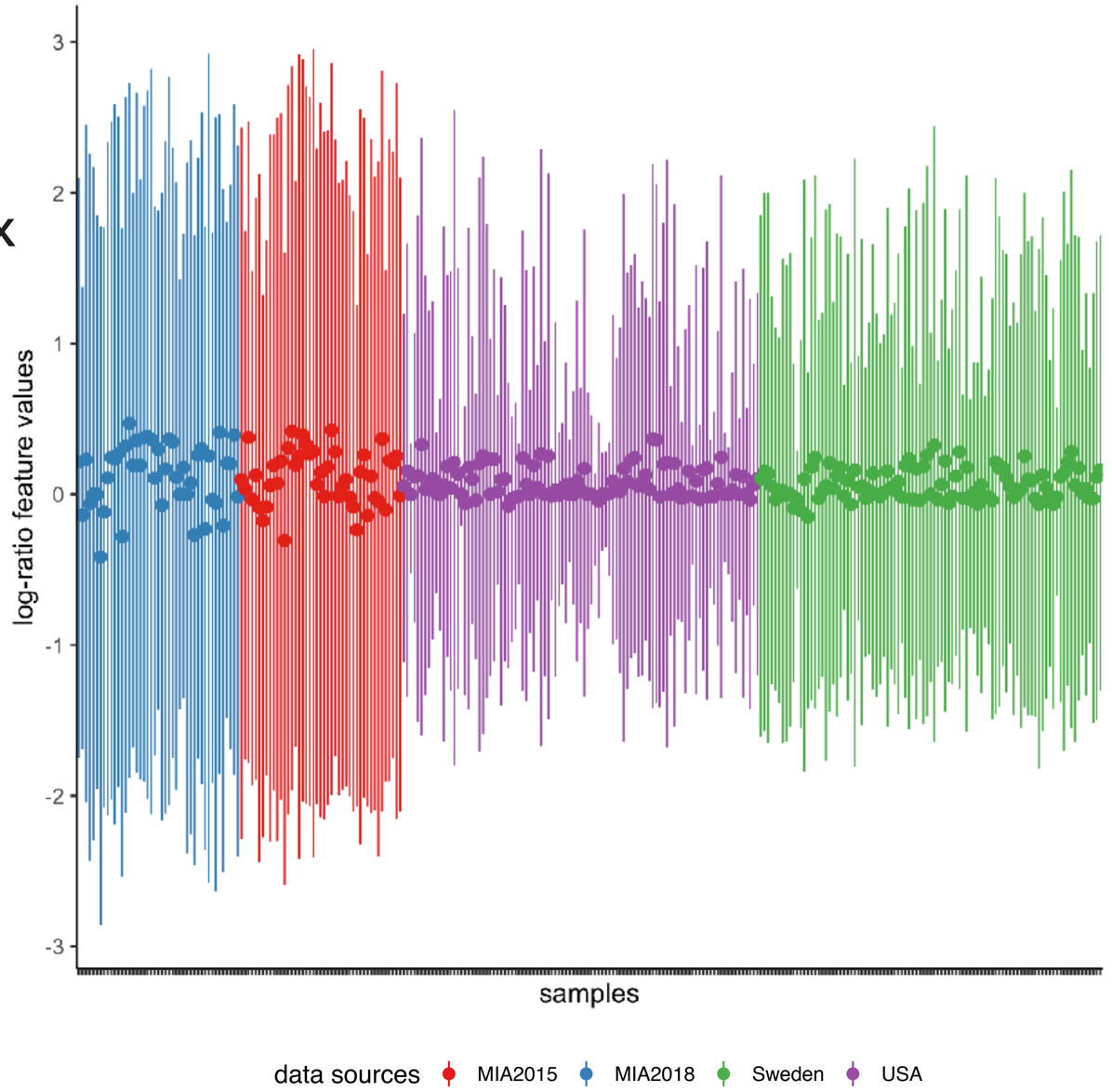
For each column in the gene expression matrix

$$X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p\} \in \mathbb{R}^{n \times p}$$

Construct $Z \in \mathbb{R}^{n \times \binom{p}{2}}$ column-wise:

$$\boldsymbol{z}_j = \log\left(\frac{\boldsymbol{x}_l}{\boldsymbol{x}_m}\right)$$

$$\text{for } 1 \leq l < m \leq p$$

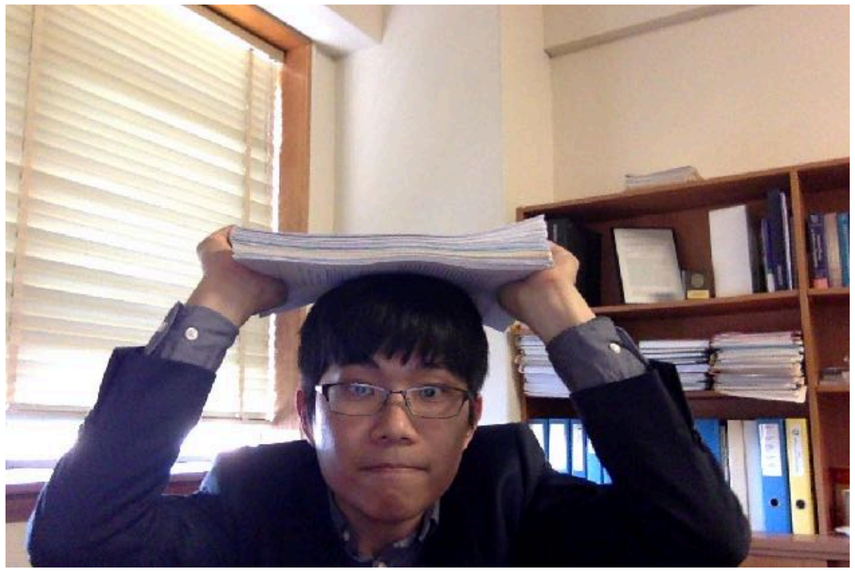


Why log-ratios?

- Within-sample standardisation avoids re-normalisation and model re-training

Why log-ratios?

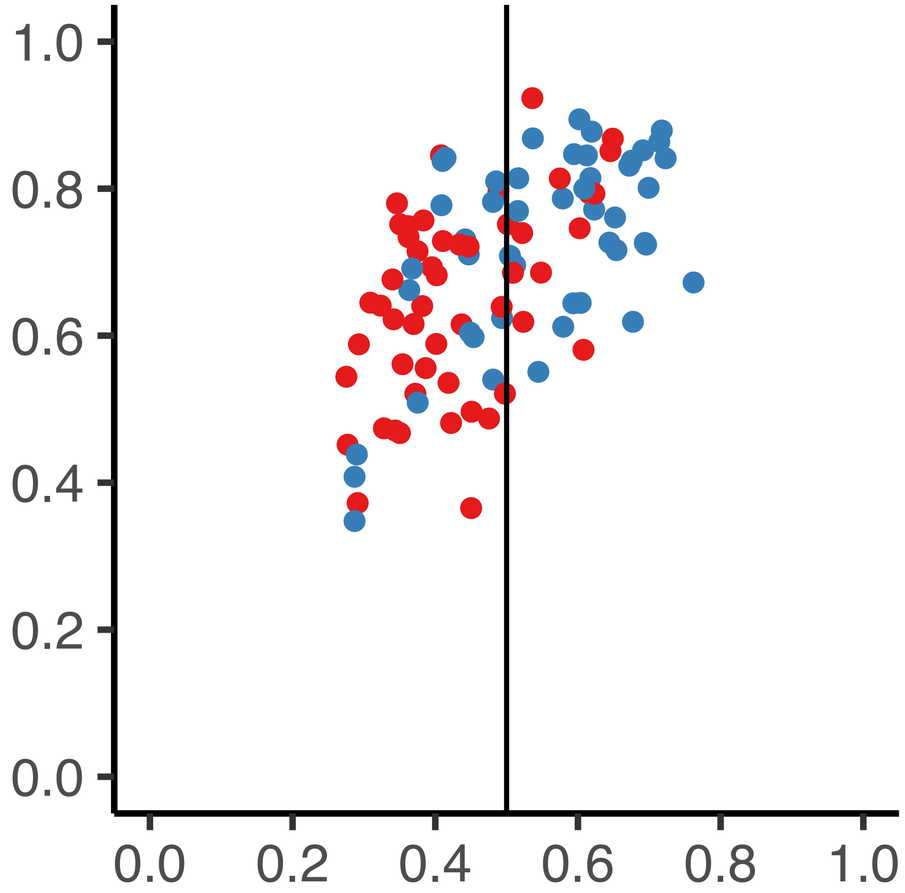
- Within-sample standardisation avoids re-normalisation and model re-training
- Under-used in clinical implementation
- Potential for further method developments



Papers that use gene expression

Papers that use ratios

Is log-ratio enough?

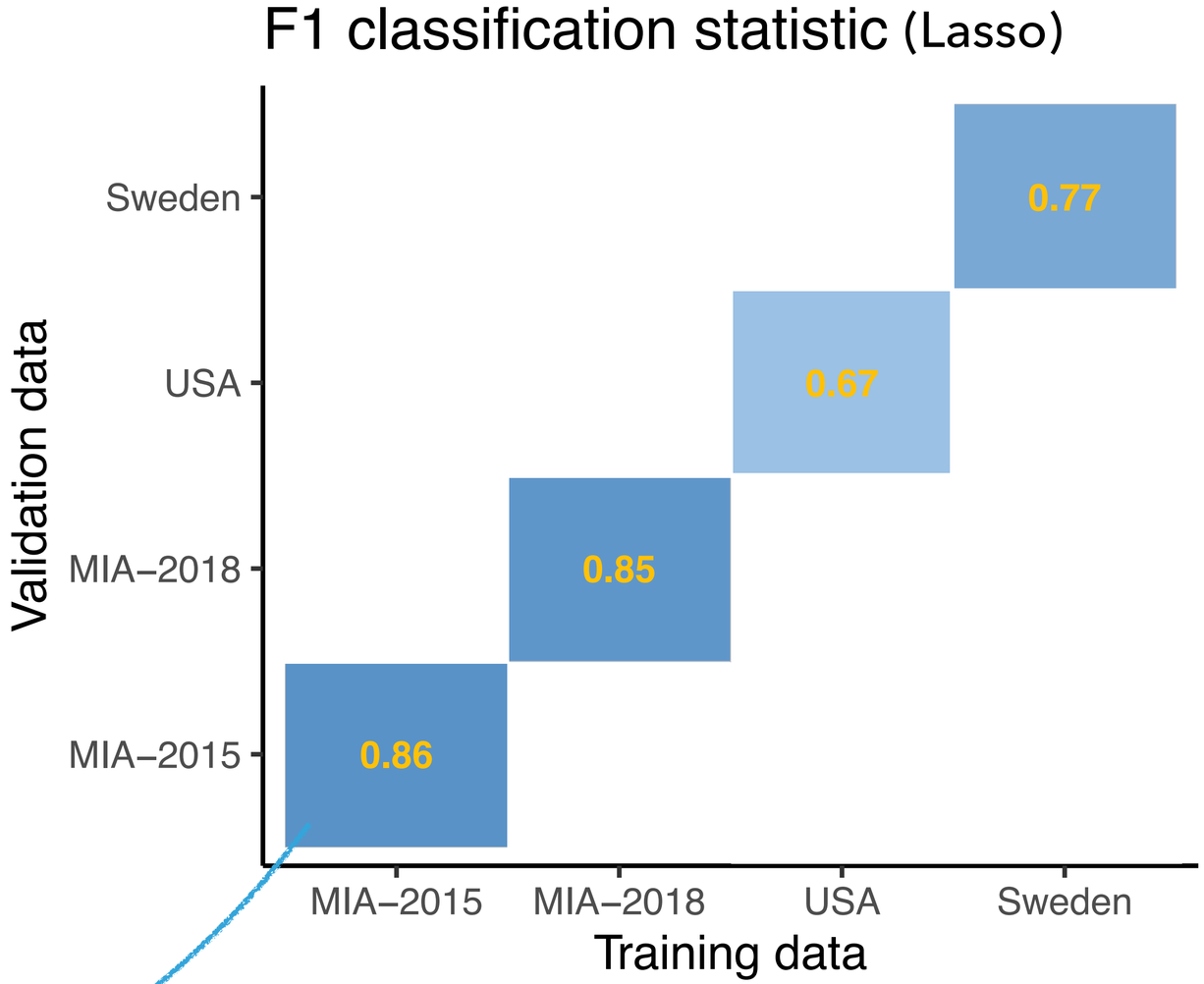


Sample class ● Good ● Poor

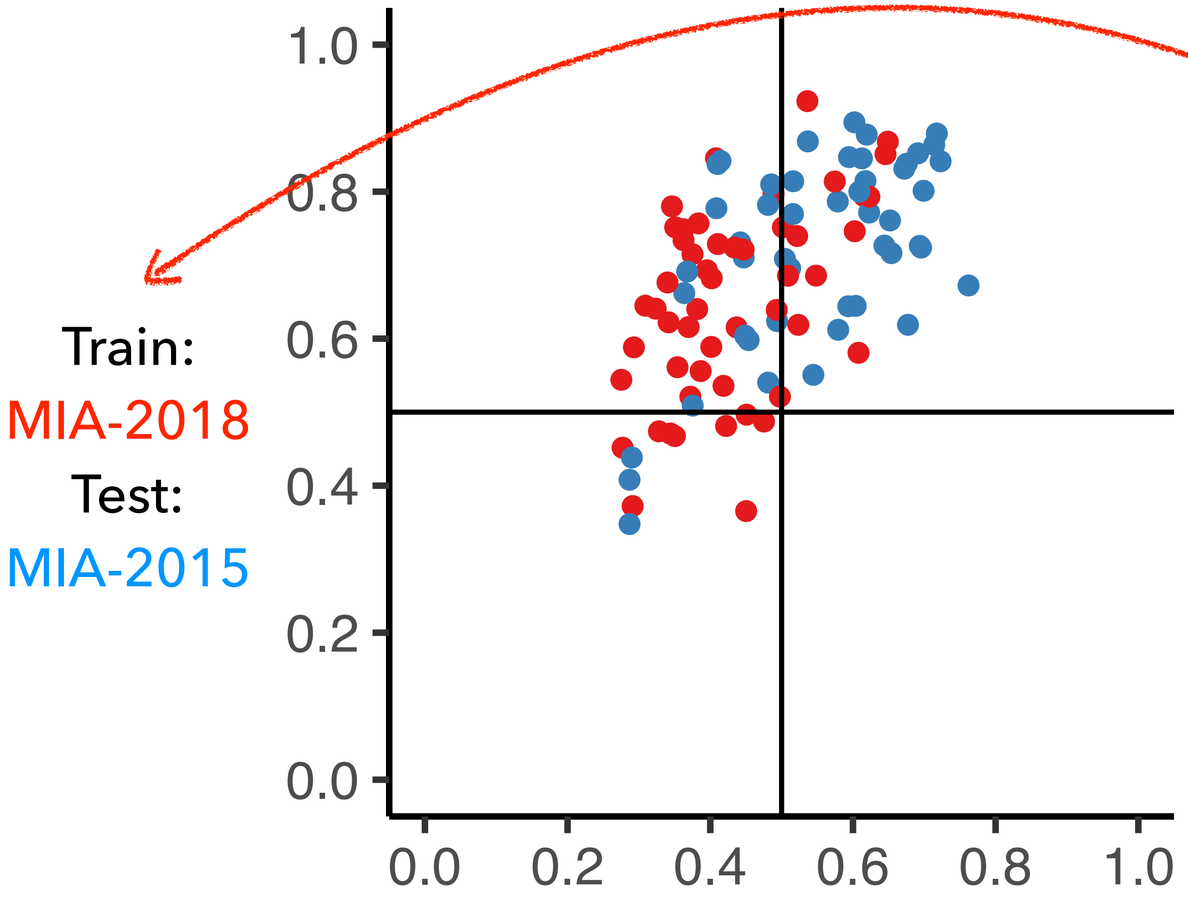
*Predicted values are shown as classification probability

Train: MIA-2015

Test: MIA-2015



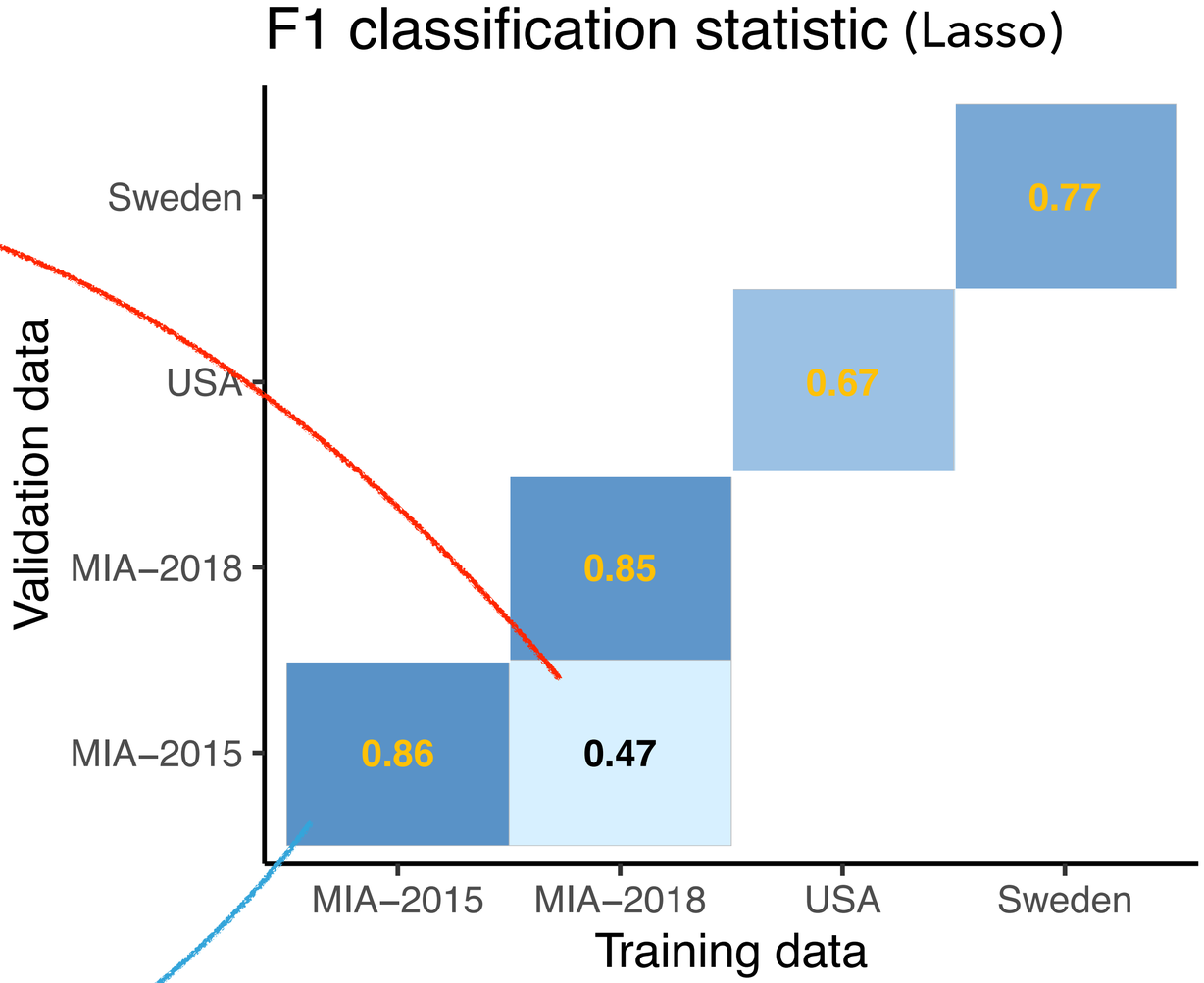
Is log-ratio enough?



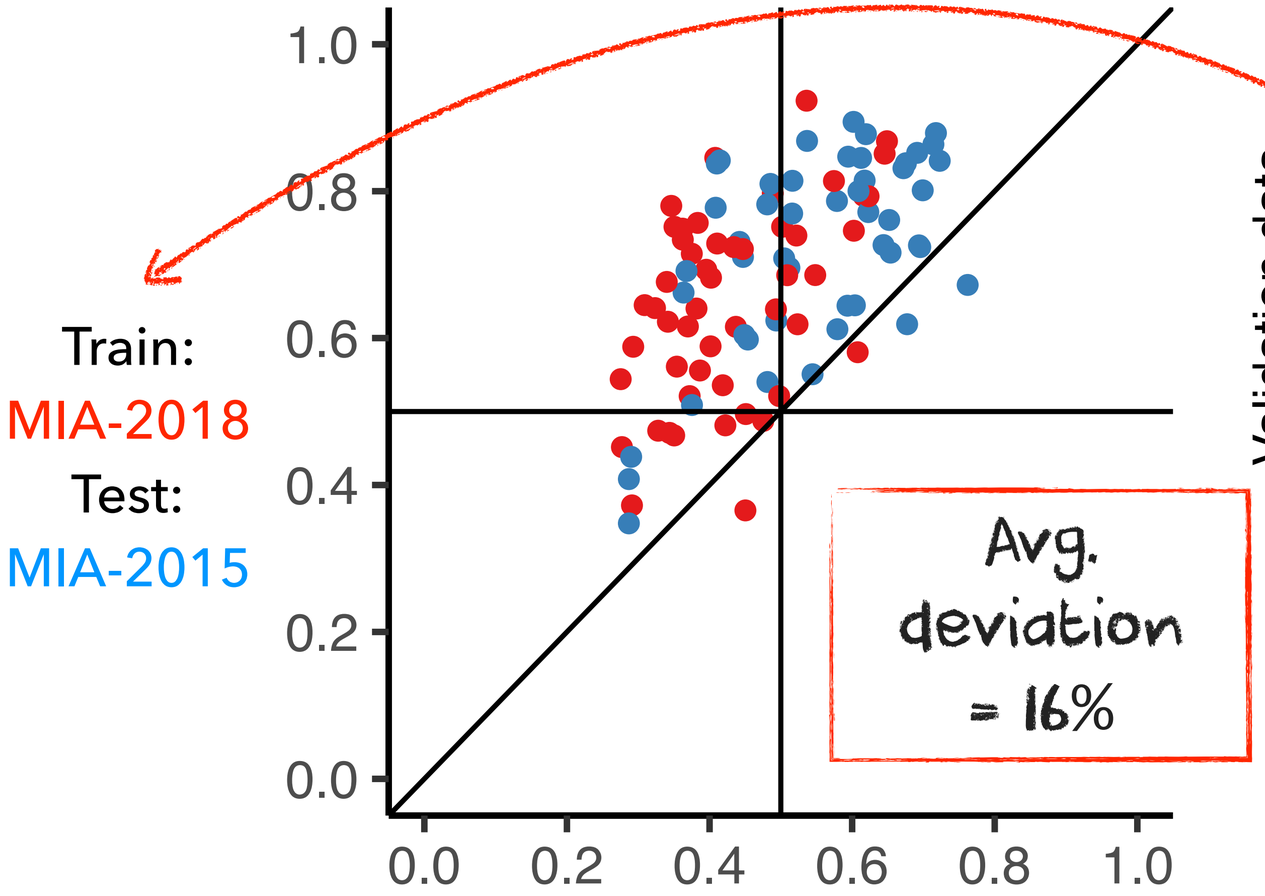
Train: MIA-2015
Test: MIA-2015

Sample class ● Good ● Poor

*Predicted values are shown as classification probability

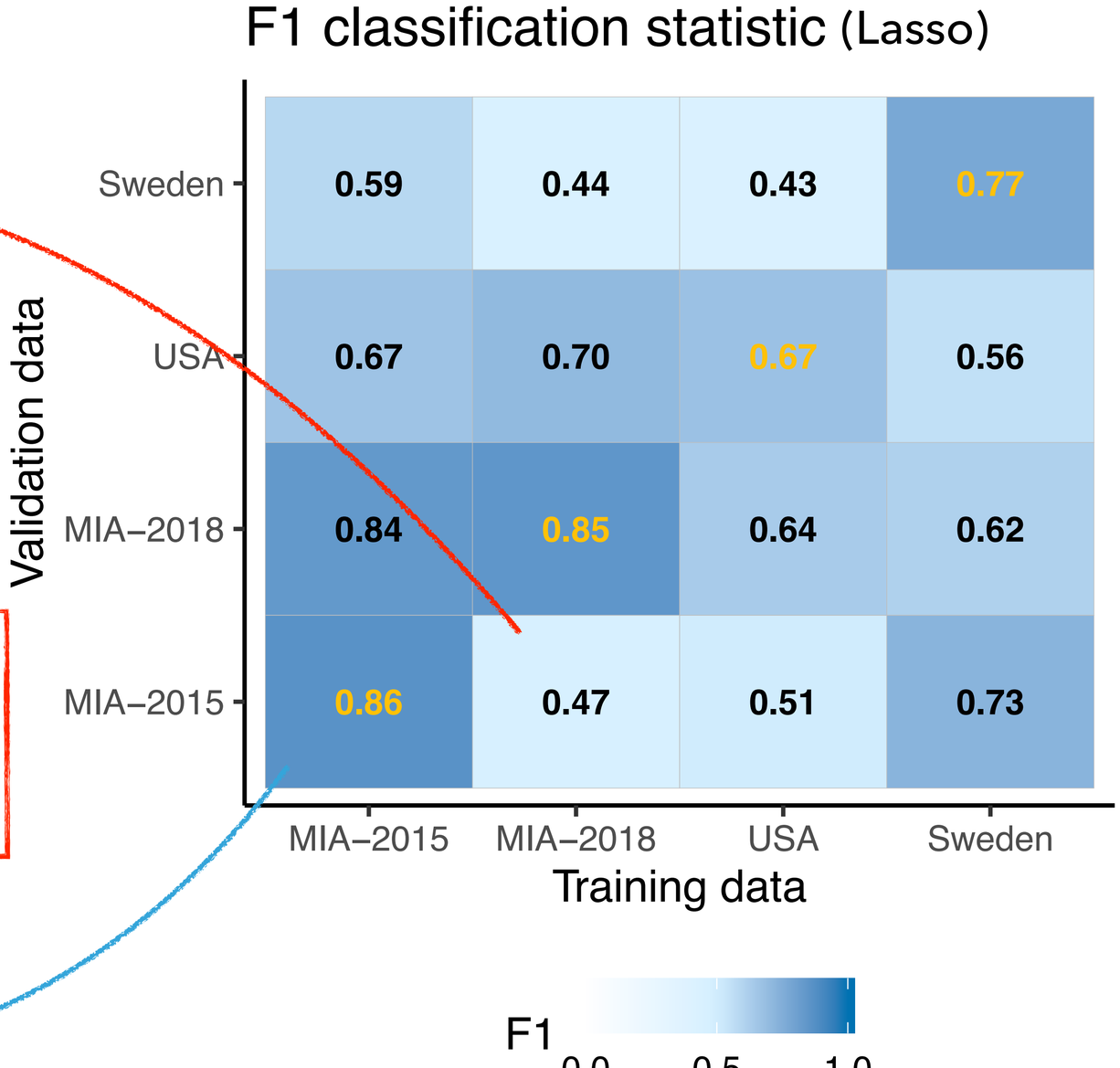


Is log-ratio enough?



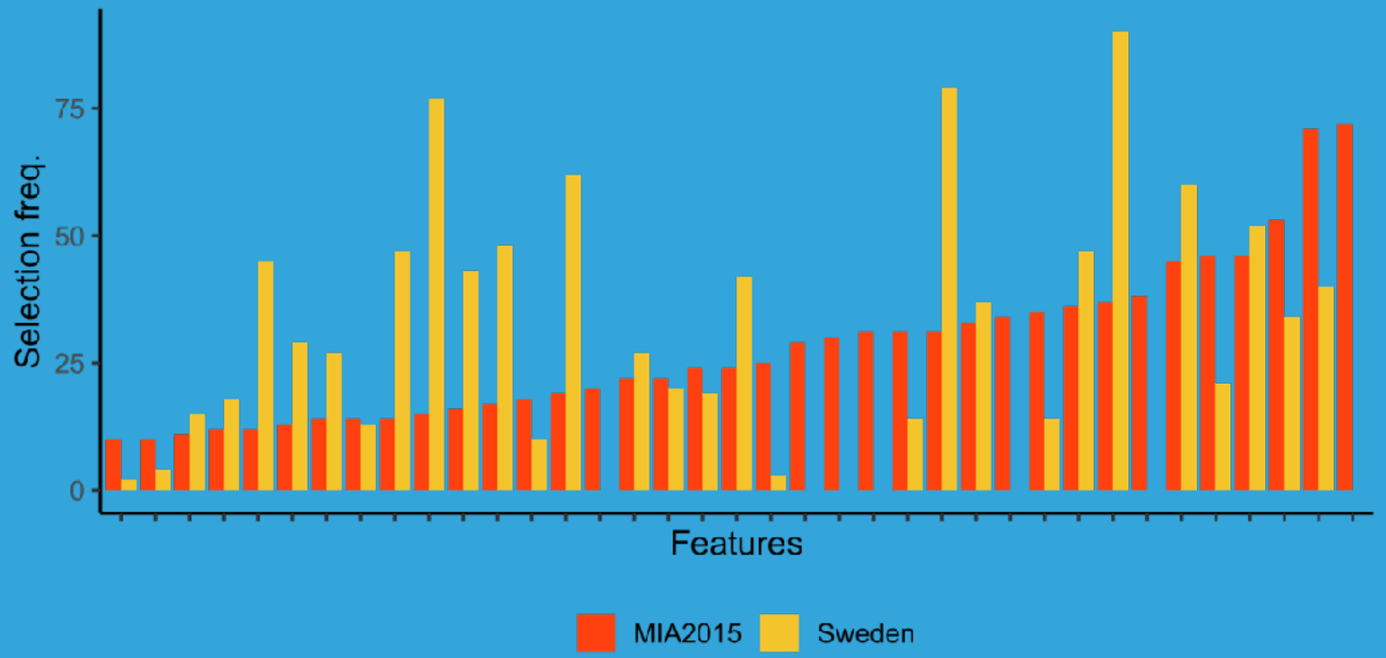
Sample class ● Good ● Poor

*Predicted values are shown as classification probability



Transferability implies that predicted values should be evenly scattered around the identity line!

Lasso variable selection is not stable!



Bach 2008, Meinshausen & Bühlmann 2010, Lim & Yu 2016

Second component of CPOP: stable feature selection and estimation

CPOP flowchart

Data

$$(X_1, y_1) \rightarrow (Z_1, y_1)$$

$$(X_2, y_2) \rightarrow (Z_2, y_2)$$

Feature transform

Model

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

Stable estimation

Prediction

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

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

Stable prediction

Weighted Elastic Net

logistic loss function

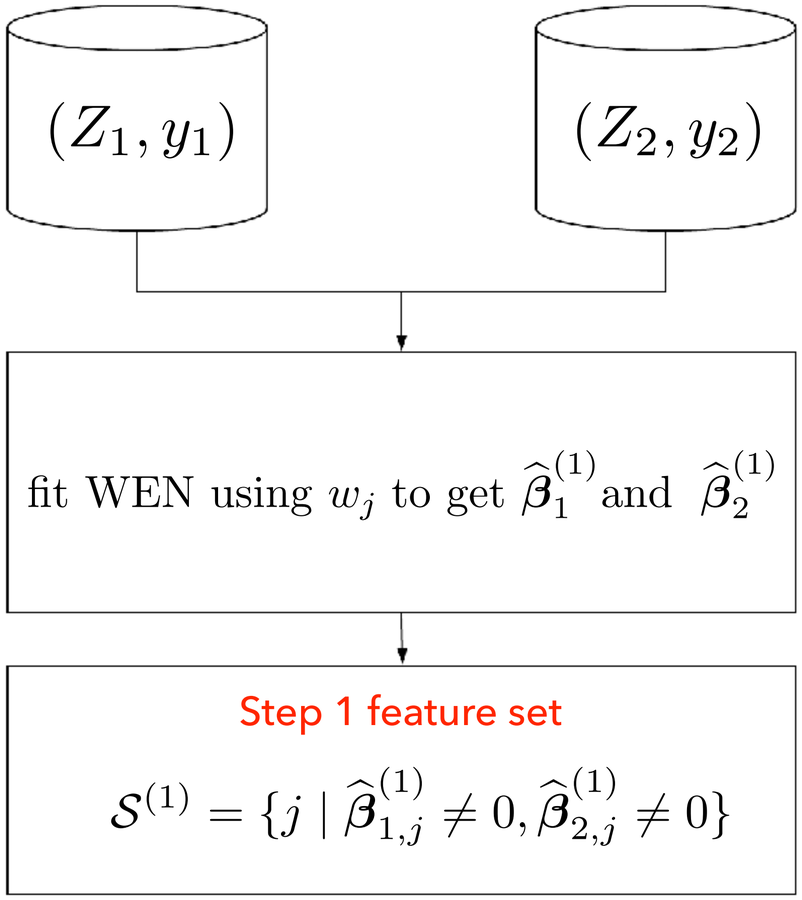
A mix of L1 and L2 penalties

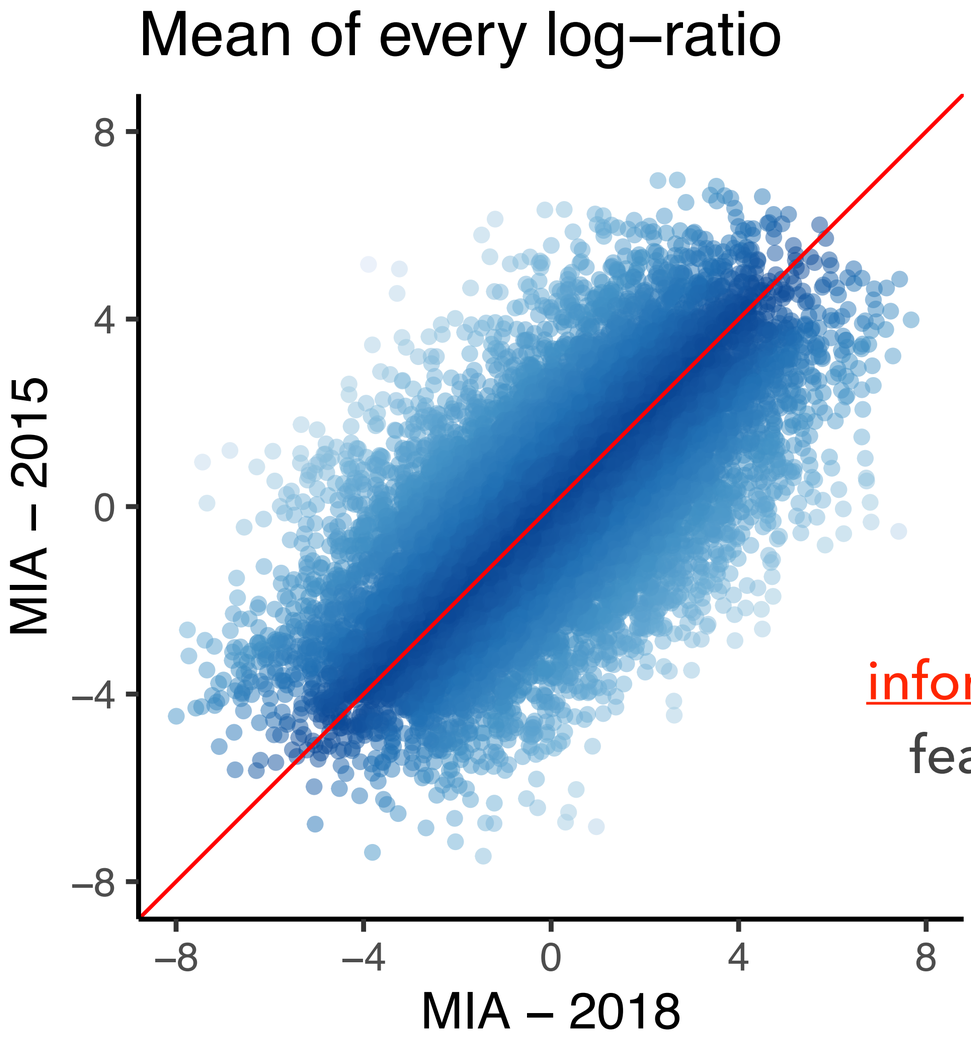
$$\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{Z}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{\binom{p}{2}}}{\operatorname{argmin}} \sum_{i=1}^{n} l\left(y_i, \boldsymbol{z}_i^{\top} \boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{q} w_j \left[\frac{(1-\alpha)}{2} \beta_j^2 + \alpha |\beta_j| \right]$$

Modelling on log-ratio features

Weights on each feature, proportional to the stability of features

Step 1: feature selection stability





informative and stable features agreed by both data

Step 2: feature estimation stability

Step 1 feature set

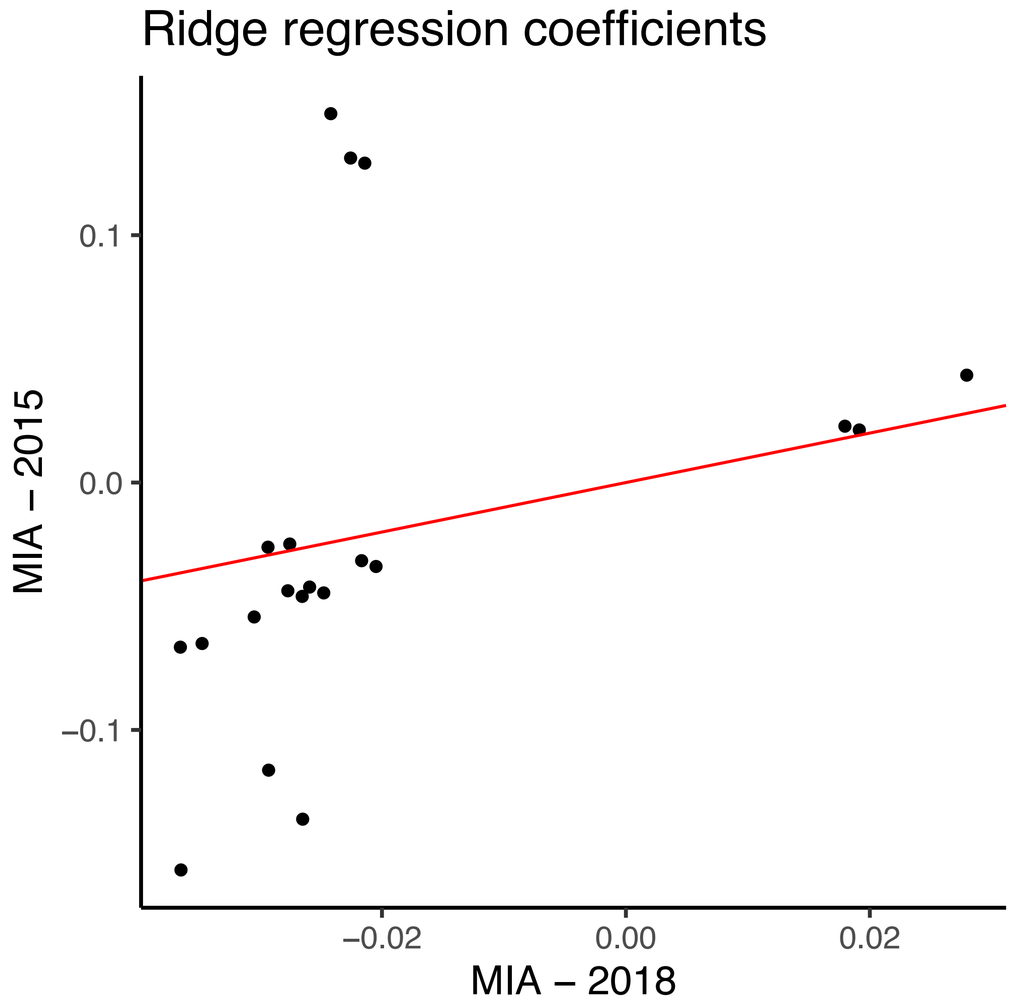
$$\mathcal{S}^{(1)} = \{j \mid \widehat{\boldsymbol{\beta}}^{(1)}_{1,j} \neq 0, \widehat{\boldsymbol{\beta}}^{(1)}_{2,j} \neq 0\}$$

$(\boldsymbol{Z}_{1,\mathcal{S}^{(1)}}, \boldsymbol{y}_1)$ $(\boldsymbol{Z}_{2,\mathcal{S}^{(1)}}, \boldsymbol{y}_2)$

fit Ridge to get $\widehat{\boldsymbol{\beta}}^{(2)}_1$ and $\widehat{\boldsymbol{\beta}}^{(2)}_2$

Step 2 feature set

$$\mathcal{S}^{(2)} = \{j \mid \text{sign}(\widehat{\boldsymbol{\beta}}^{(2)}_{1,j}) = \text{sign}(\widehat{\boldsymbol{\beta}}^{(2)}_{2,j})\}$$



*Not all features are shown

Step 2: feature estimation stability

Step 1 feature set

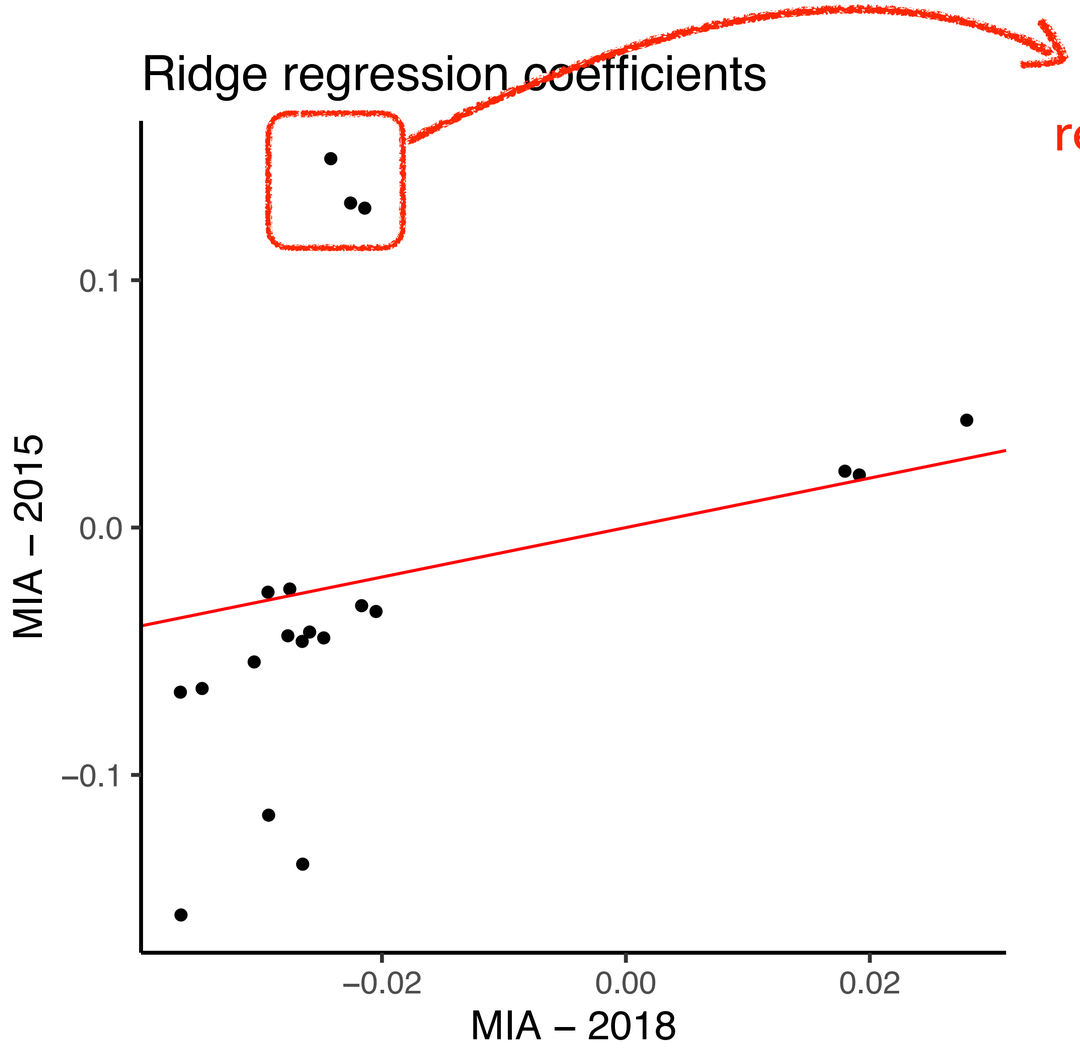
$$\mathcal{S}^{(1)} = \{j \mid \widehat{\boldsymbol{\beta}}^{(1)}_{1,j} \neq 0, \widehat{\boldsymbol{\beta}}^{(1)}_{2,j} \neq 0\}$$

$(\boldsymbol{Z}_{1,\mathcal{S}^{(1)}}, \boldsymbol{y}_1)$ $(\boldsymbol{Z}_{2,\mathcal{S}^{(1)}}, \boldsymbol{y}_2)$

fit Ridge to get $\widehat{\boldsymbol{\beta}}^{(2)}_1$ and $\widehat{\boldsymbol{\beta}}^{(2)}_2$

Step 2 feature set

$$\mathcal{S}^{(2)} = \{j \mid \text{sign}(\widehat{\boldsymbol{\beta}}^{(2)}_{1,j}) = \text{sign}(\widehat{\boldsymbol{\beta}}^{(2)}_{2,j})\}$$



remove features with unstable estimates across data

*Not all features are shown

Step 2: feature estimation stability

Step 1 feature set

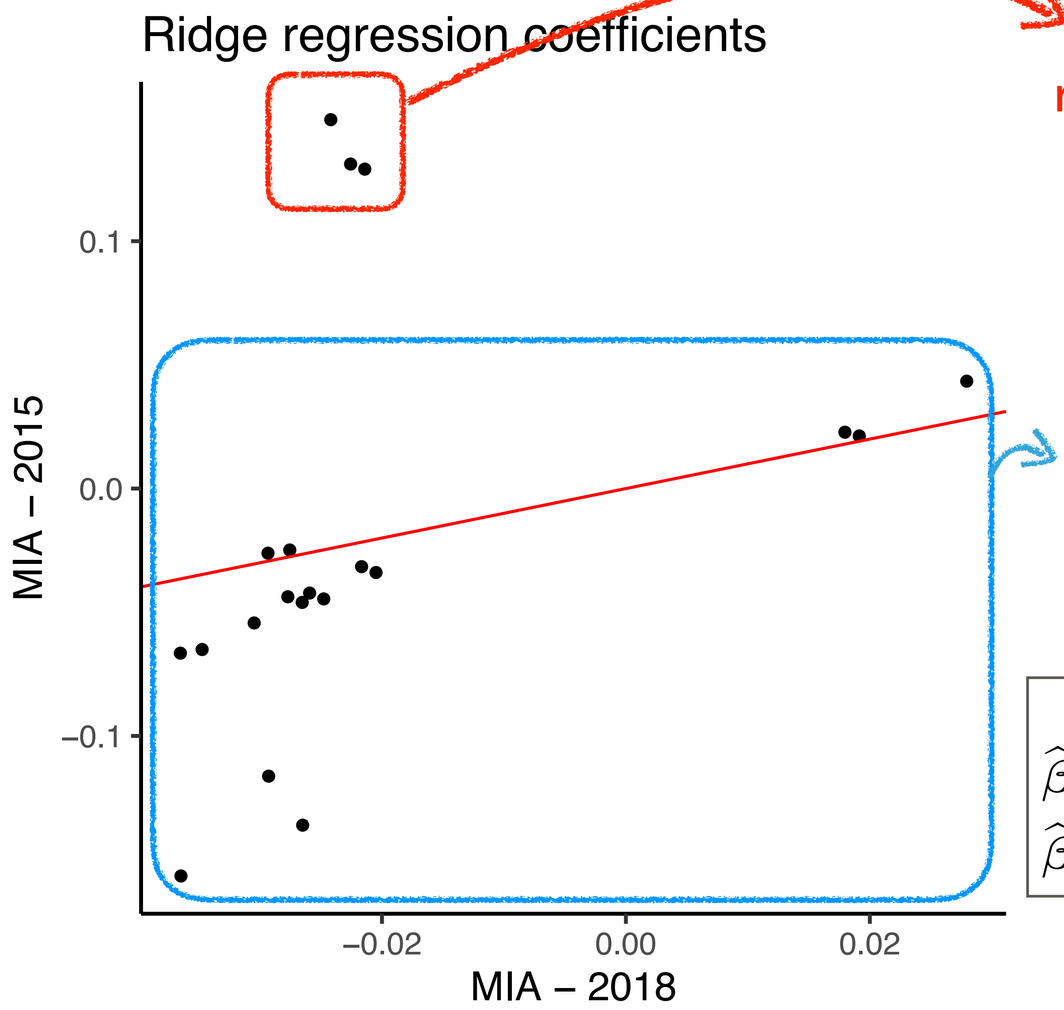
$$\mathcal{S}^{(1)} = \{j \mid \widehat{\boldsymbol{\beta}}^{(1)}_{1,j} \neq 0, \widehat{\boldsymbol{\beta}}^{(1)}_{2,j} \neq 0\}$$

$(\boldsymbol{Z}_{1,\mathcal{S}^{(1)}}, \boldsymbol{y}_1)$ $(\boldsymbol{Z}_{2,\mathcal{S}^{(1)}}, \boldsymbol{y}_2)$

fit Ridge to get $\widehat{\boldsymbol{\beta}}^{(2)}_1$ and $\widehat{\boldsymbol{\beta}}^{(2)}_2$

Step 2 feature set

$$\mathcal{S}^{(2)} = \{j \mid \text{sign}(\widehat{\boldsymbol{\beta}}^{(2)}_{1,j}) = \text{sign}(\widehat{\boldsymbol{\beta}}^{(2)}_{2,j})\}$$



remove features with unstable estimates across data

features suitable for stable prediction:

Final model

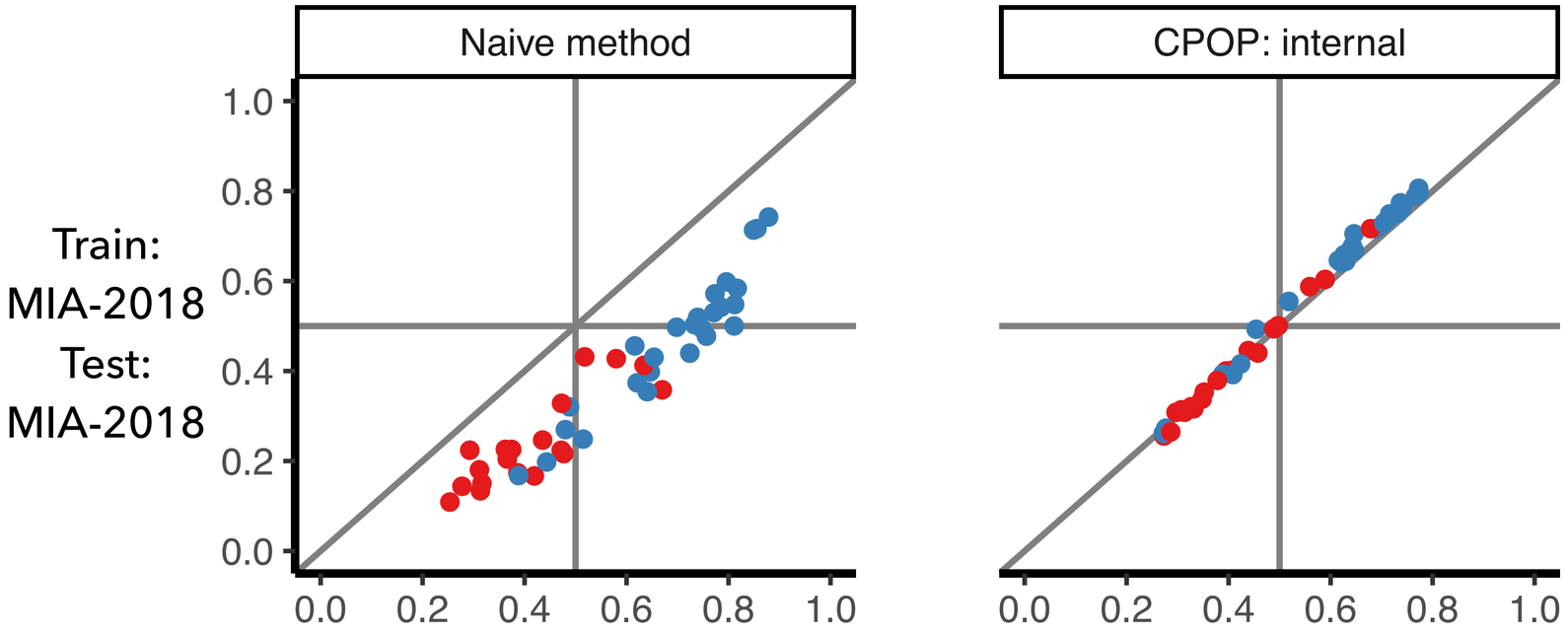
$\widehat{\boldsymbol{\beta}}_1$: fit Ridge on $(\boldsymbol{Z}_{1,\mathcal{S}^{(2)}}, \boldsymbol{y}_1)$

$\widehat{\boldsymbol{\beta}}_2$: fit Ridge on $(\boldsymbol{Z}_{2,\mathcal{S}^{(2)}}, \boldsymbol{y}_2)$

*Not all features are shown

Results

Result 1: MIA 2015 vs 2018 data

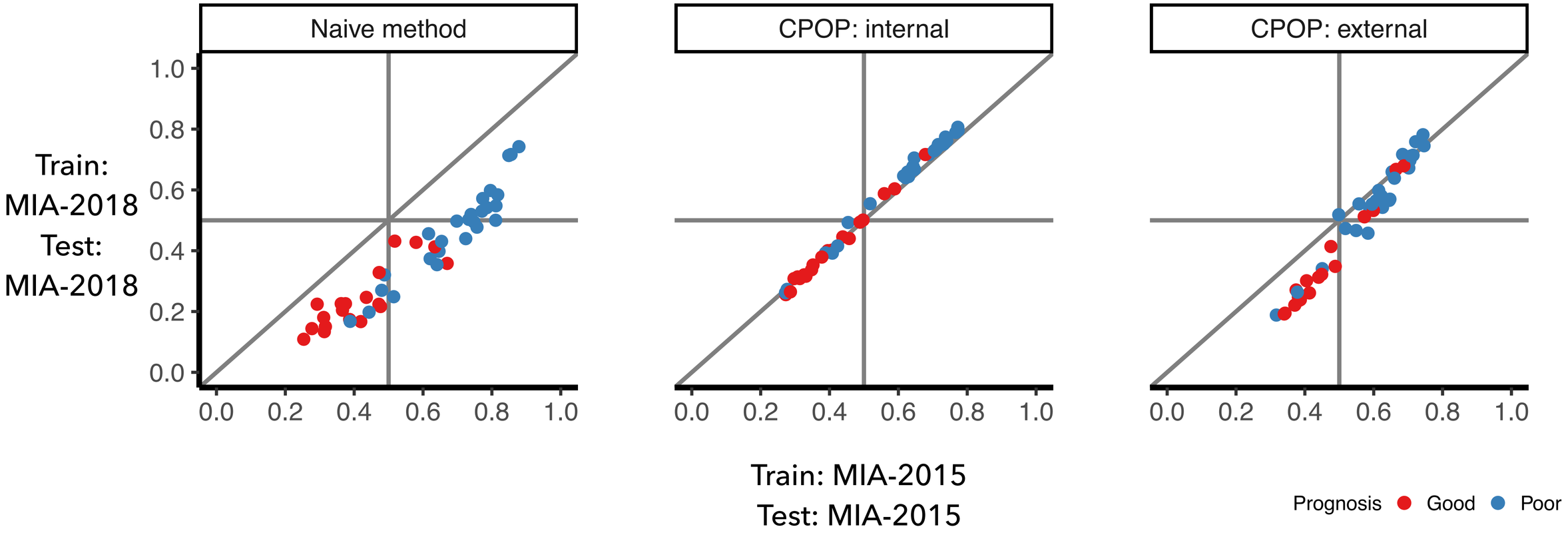


Train: MIA-2015
Test: MIA-2015

Prognosis ● Good ● Poor

Small deviation in predicted values across datasets

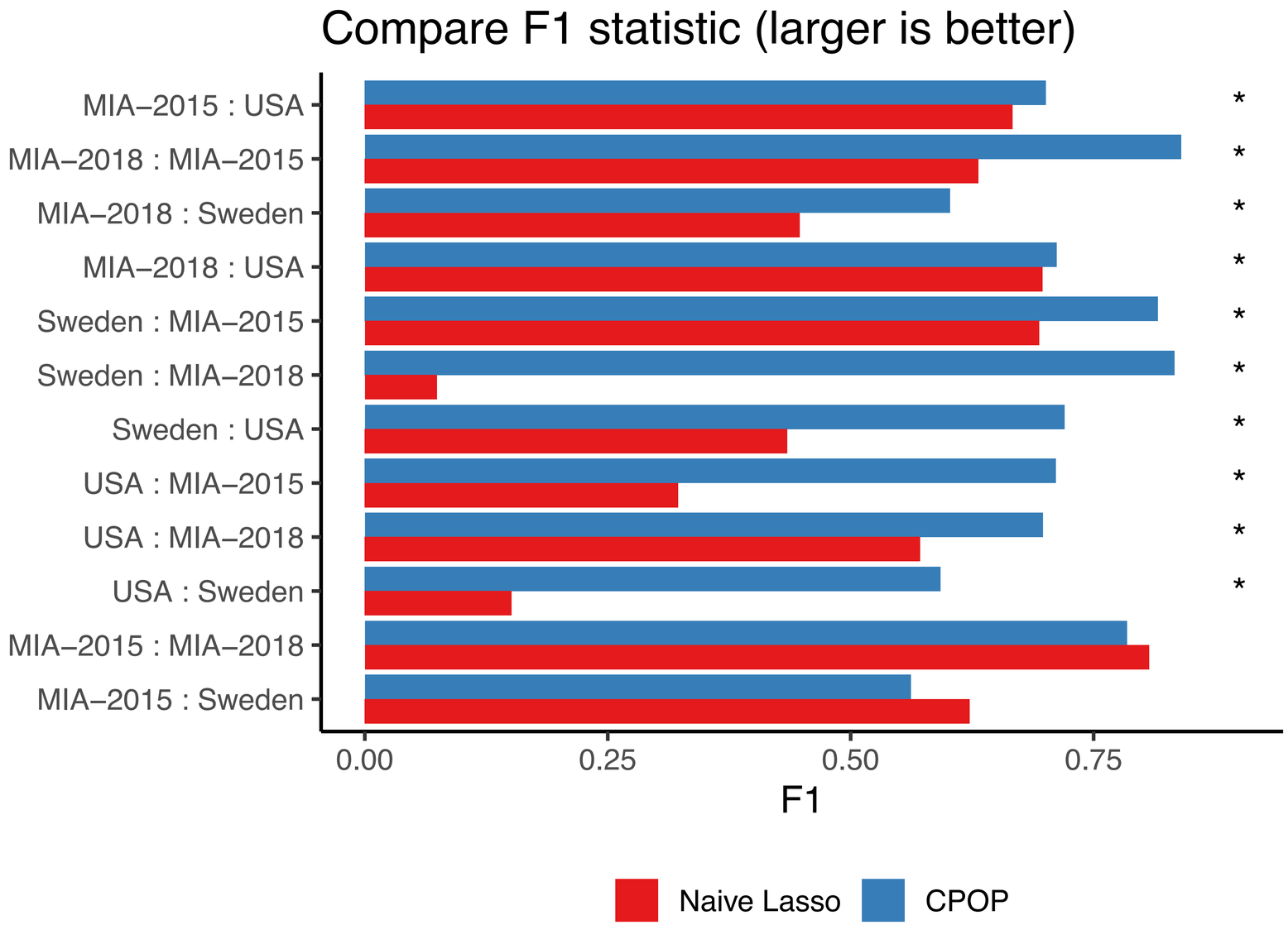
Result 1: MIA 2015 vs 2018 data



Small deviation in predicted values across datasets

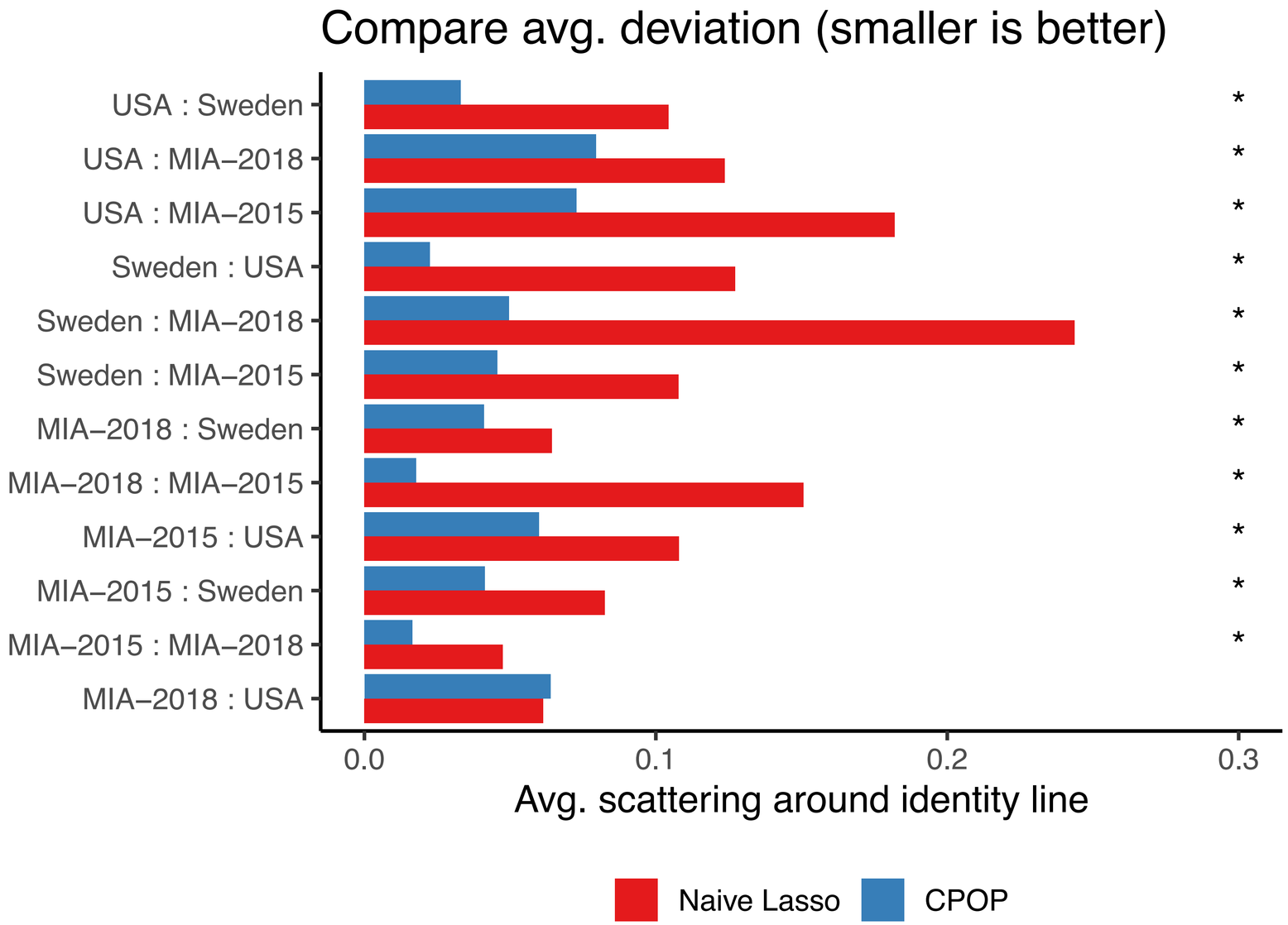
Result 2: four melanoma data

CPOP is highly predictive



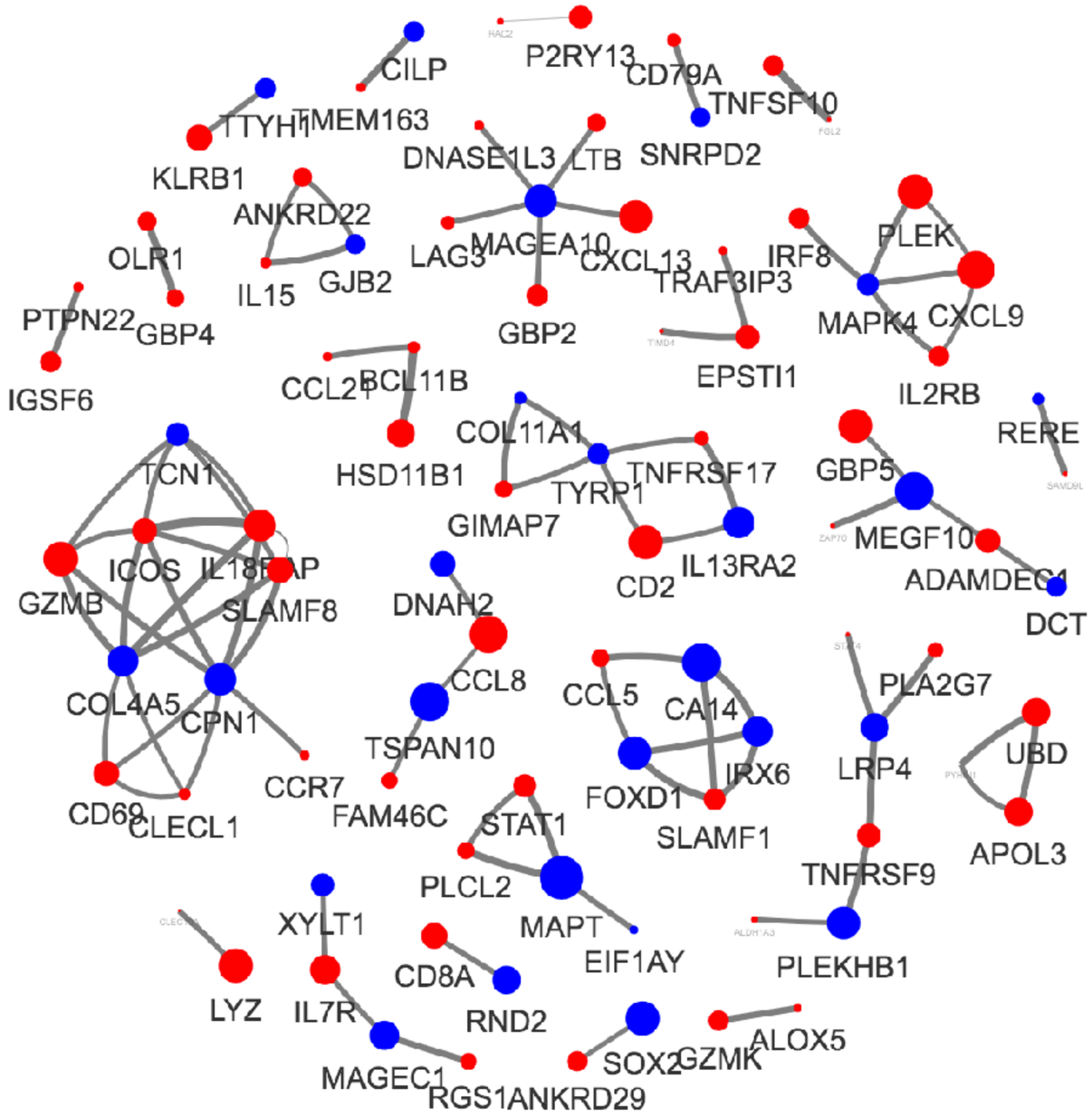
Result 2: four melanoma data

CPOP is highly stable



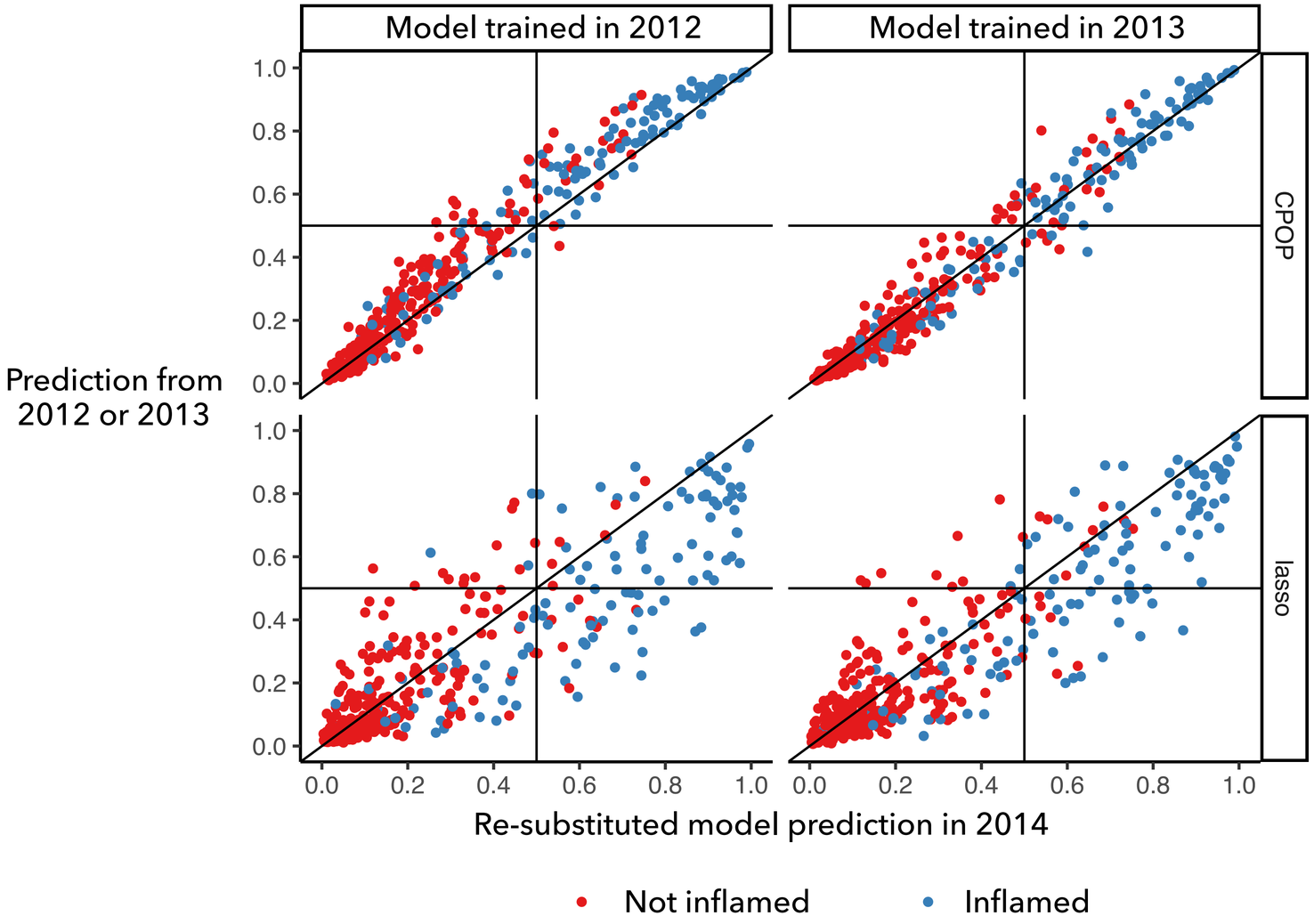
Result 3: melanoma feature set

CPOP features offer new biological interpretations



Result 4: prospective prediction on inflammatory bowel disease

CPOP works on prospective experiments



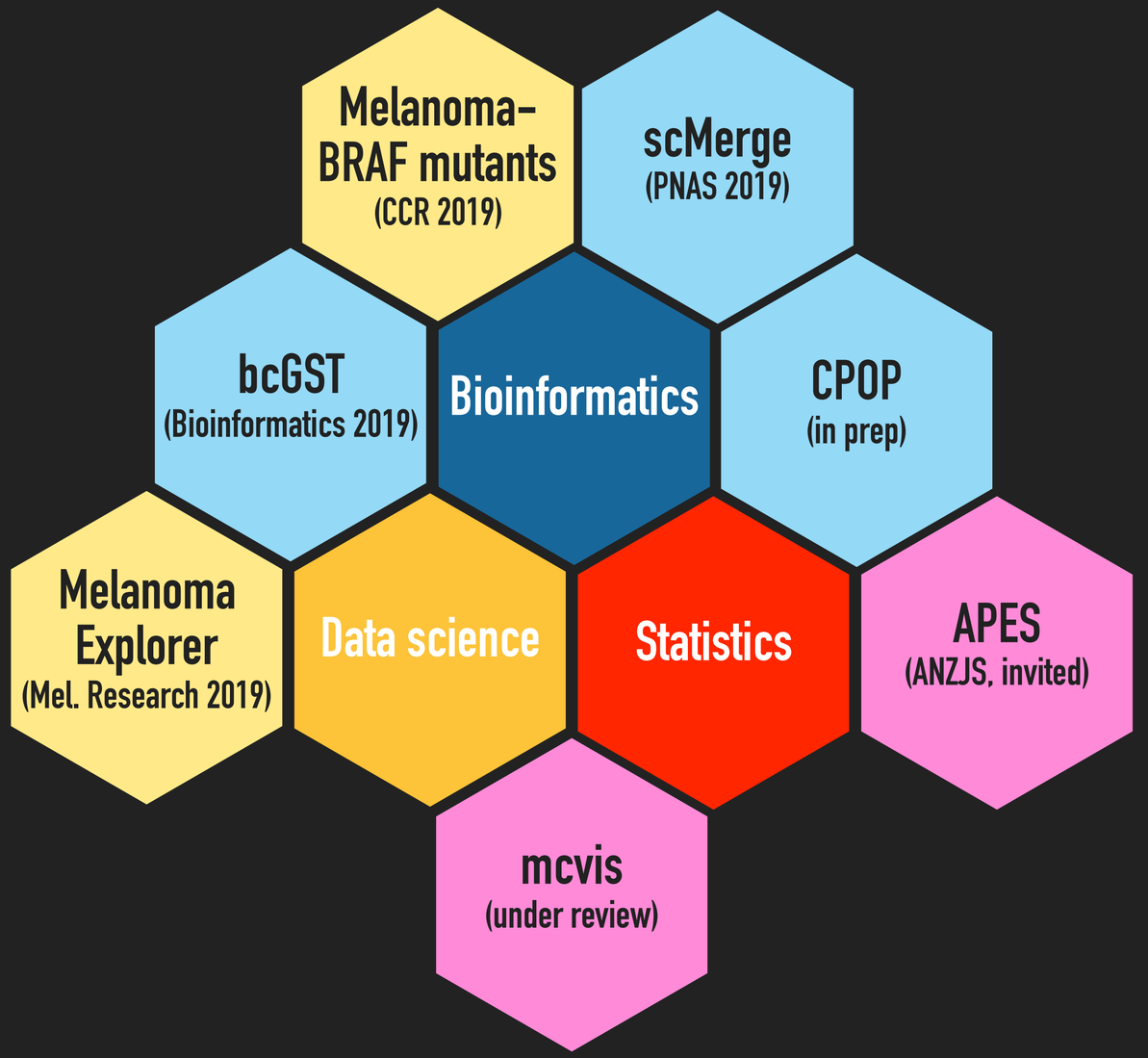
Concluding remarks

1. Integrates clinical implementation constraints into the model
2. Log-ratios enable prospective and multi-centres prediction
3. Stable variable selection and estimation components

- A flexible framework with many adaptable components
- Potential to handle data with higher relevance to precision medicine (e.g. drug sensitivity)

Acknowledgement

- Supervision:
  - Jean Yang
  - Samuel Mueller
  - Garth Tarr
- Melanoma Institute Australia
- Sydney Precision Bioinformatics Group



Clinical constraints

Data

$$(X_1, y_1)$$

$$(X_2, y_2)$$

Model

$$\hat{\beta}_1$$

Prediction

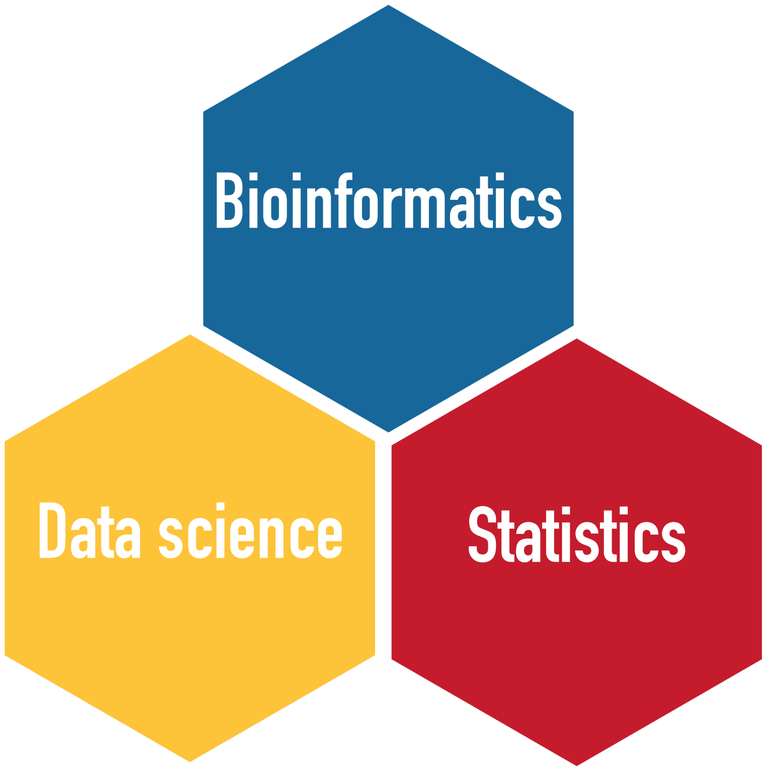
$$\hat{y}_1 = X_1 \hat{\beta}_1$$

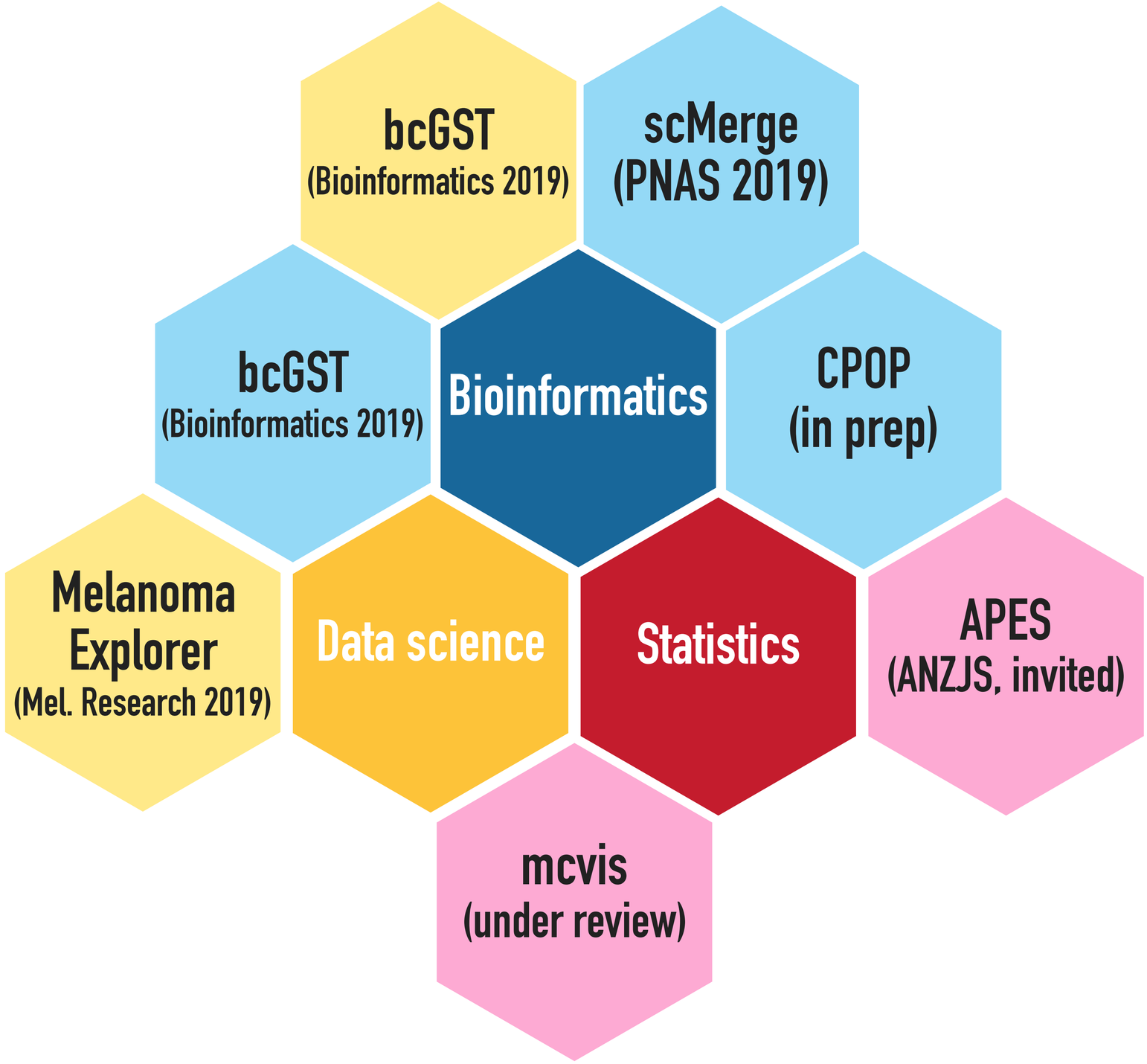
$$\hat{y}_2 = X_2 \hat{\beta}_1$$

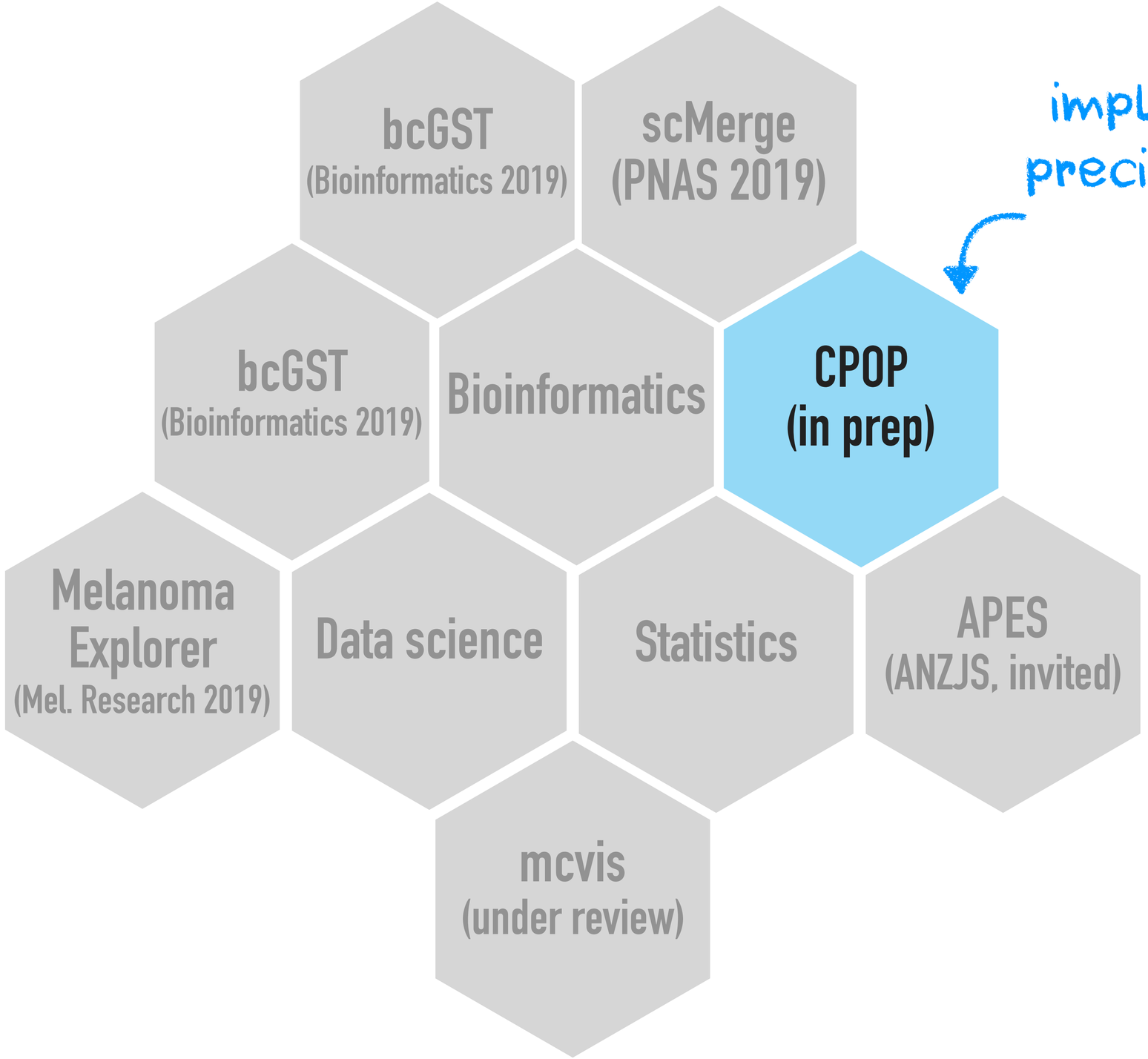
No re-normalisation

No model re-training

Scale-equivalent prediction

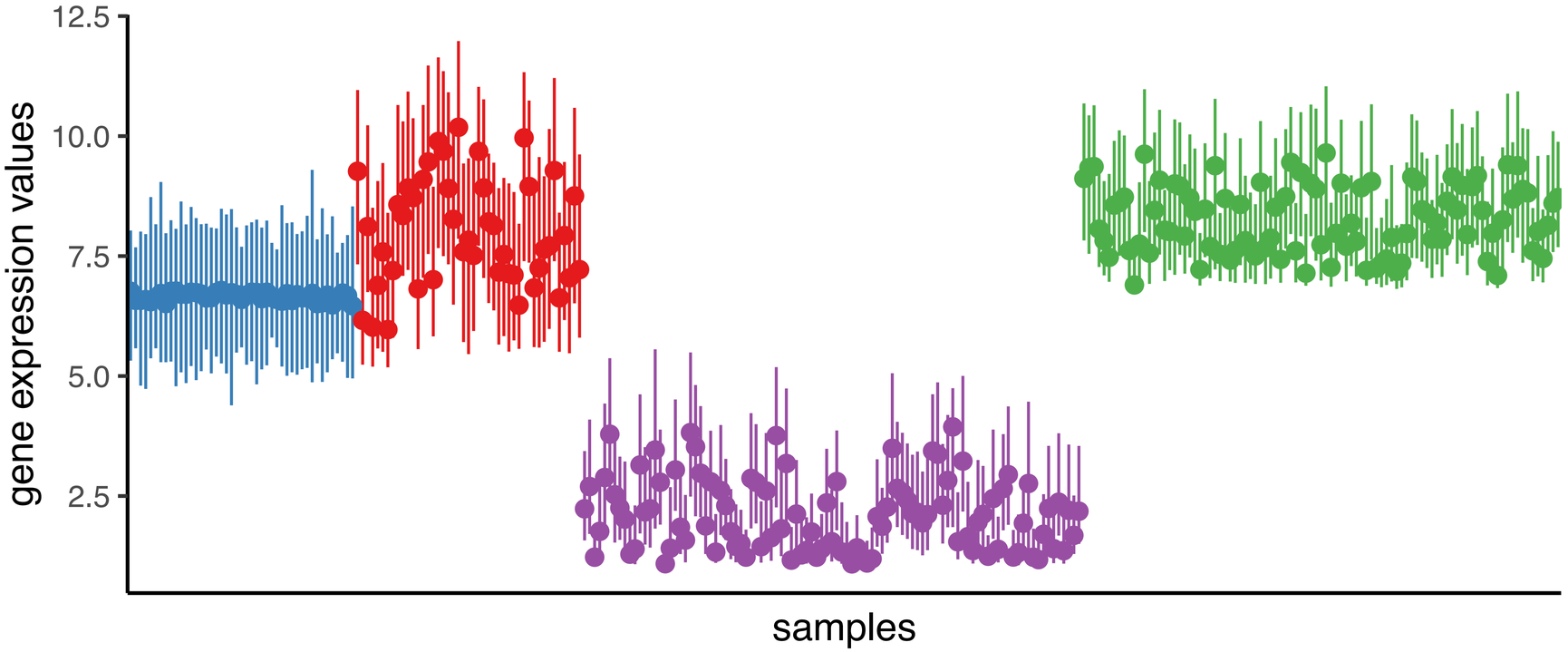

Bioinformatics
Data science
Statistics


bcGST
(Bioinformatics 2019)
scMerge
(PNAS 2019)
bcGST
(Bioinformatics 2019)
Bioinformatics
CPOP
(in prep)
Melanoma
Explorer
(Mel. Research 2019)
Data science
Statistics
APES
(ANZJS, invited)
mcvis
(under review)



Towards implementation precision medicine

Vision of a risk score



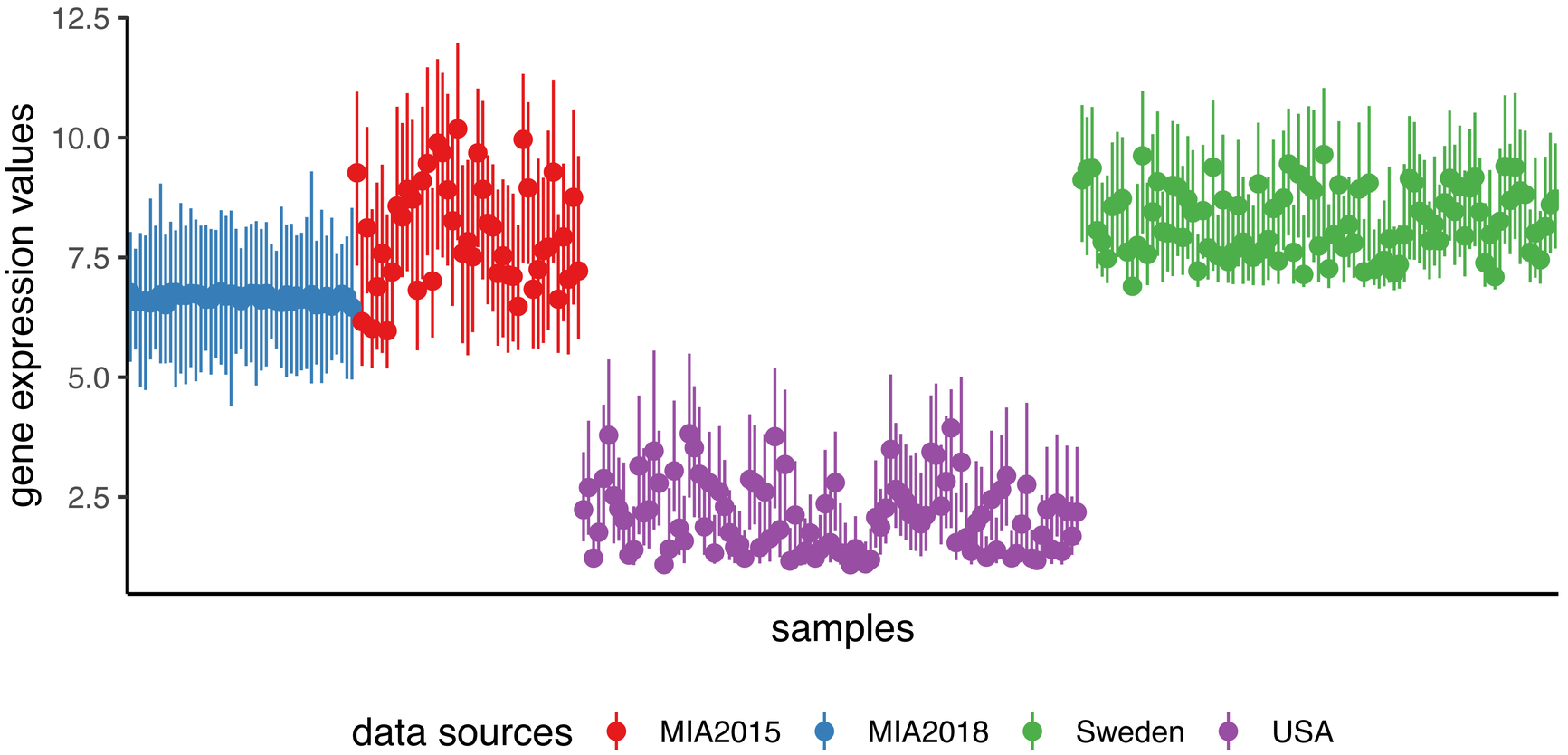
$$\hat{y} = X\hat{\beta}$$

regression-based risk score

Melanoma Institute Australia

https://researchdata.ands.org.au/centre-research-excellence-outcomes-melanoma/1318231

Melanoma Institute Australia



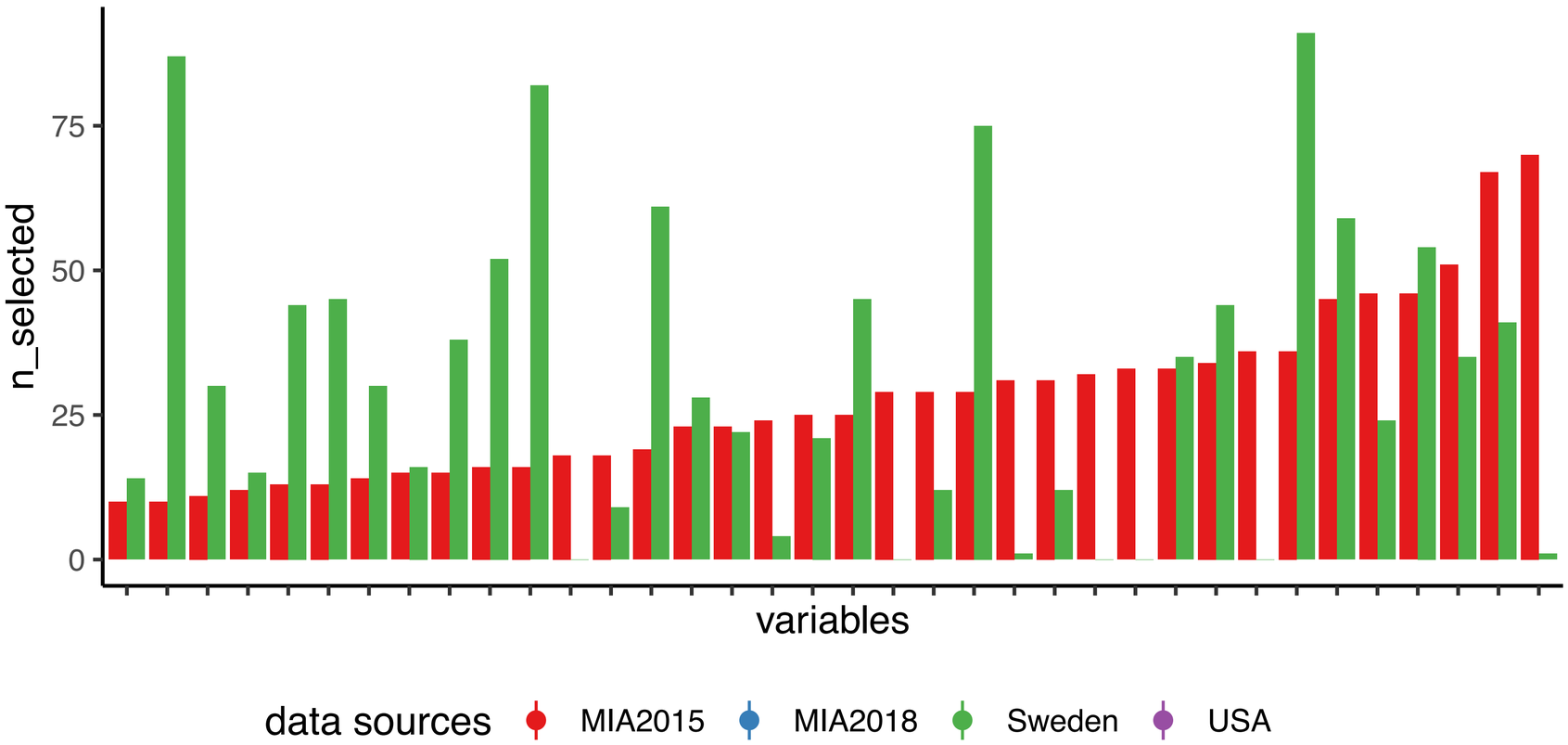
- Predict patient outcomes using gene expression:
  - prospectively
  - multi-centres
- CRE grant for implementation

https://researchdata.ands.org.au/centre-research-excellence-outcomes-melanoma/1318231

Lasso variable selection is not stable!

- Bach 2008
- Meinshausen & Bühlmann 2010
- Lim & Yu 2016

all pointed out that Lasso is unstable under cross-validation



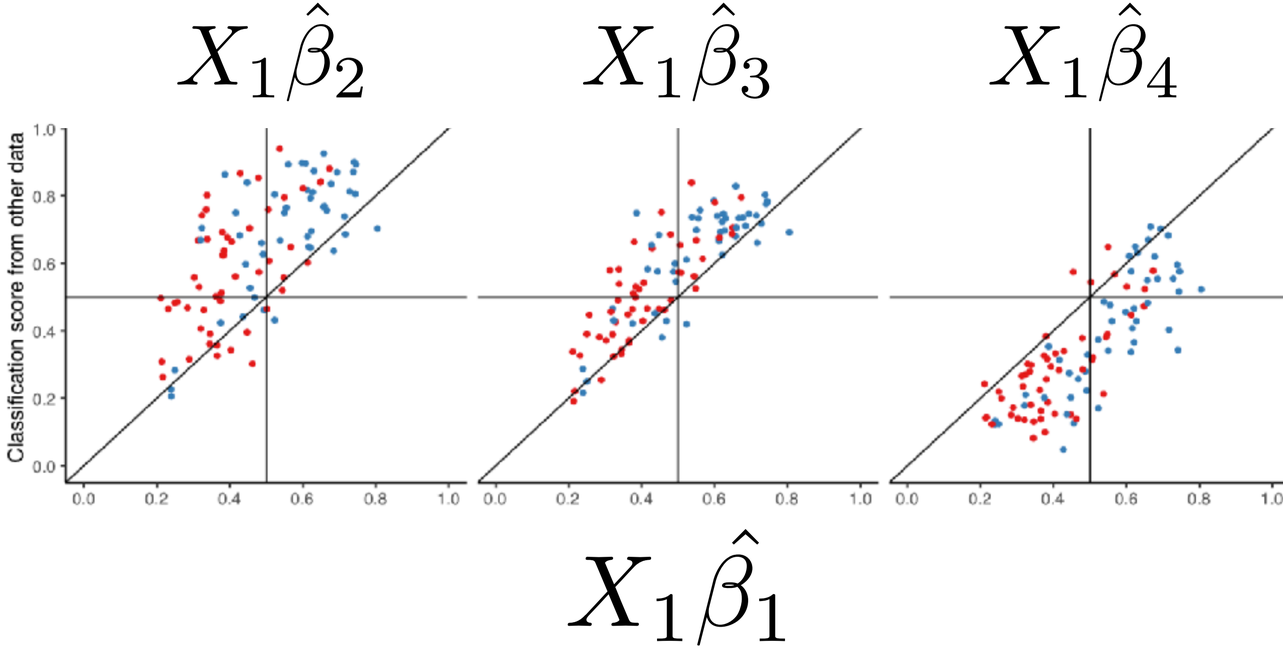
Statistical challenges

1. Concordance in features scaling across datasets

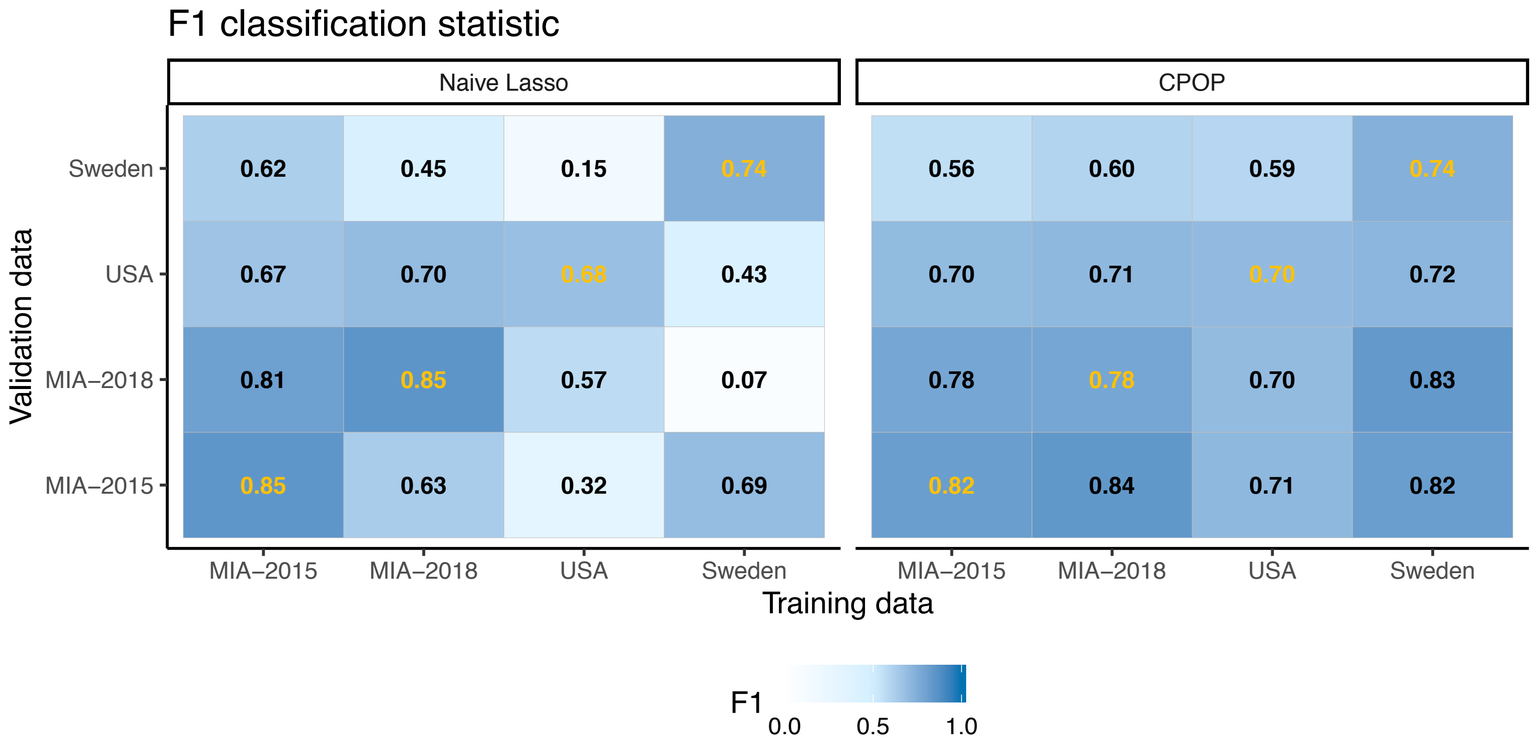
2. Concordance in feature selection and estimates

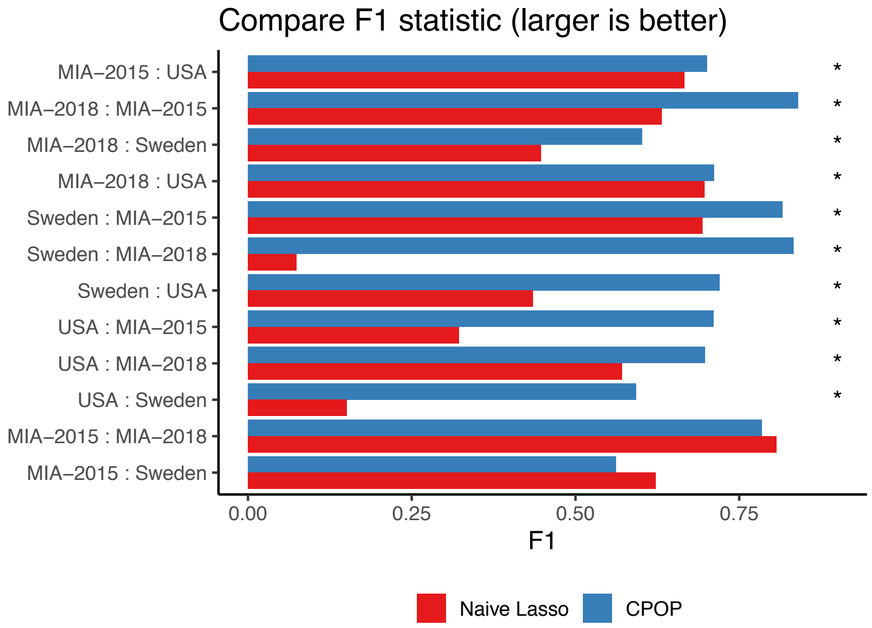
3. Single-patient prediction

Transferability implies that patients should be evenly scattered around the identity line



CPOP results 2: four melanoma data





CPOP is highly predictive

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

$$X_1\hat{\beta}_1 \approx X_2\hat{\beta}_2$$

loosely translate to

$$X_1 \approx X_2$$

column-wise

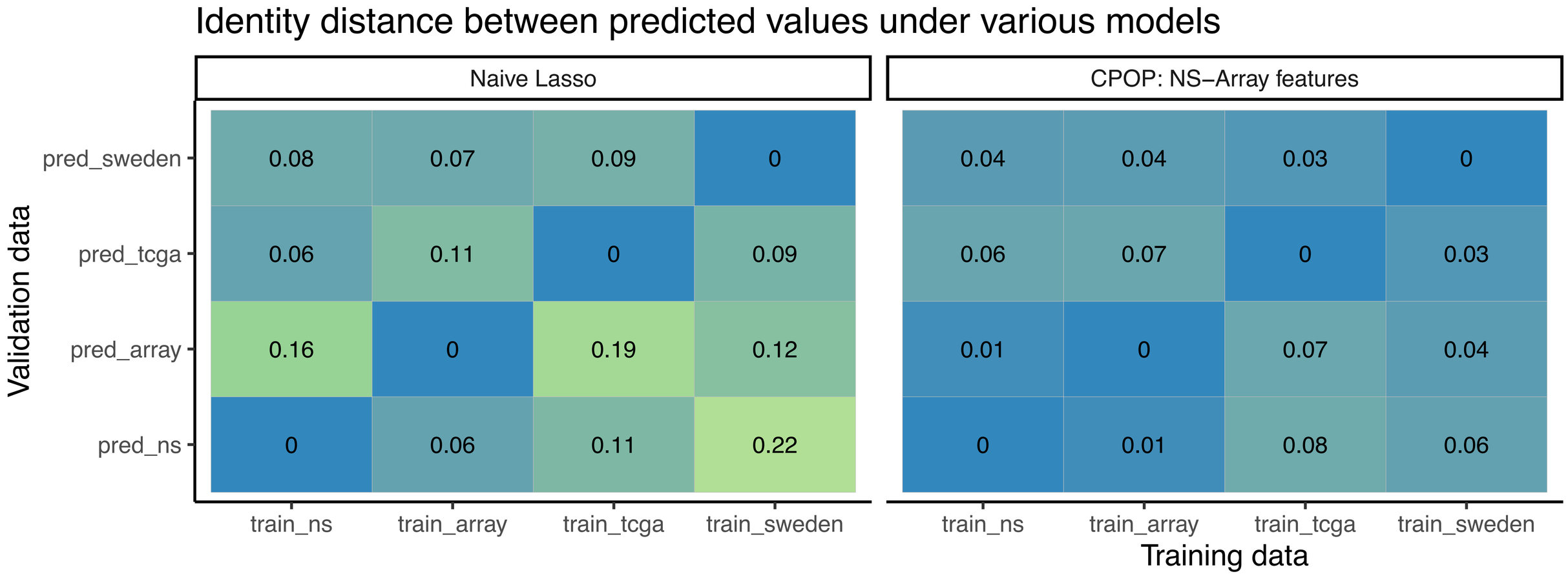
(feature distribution stability)

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

element-wise

(mode estimation stability)

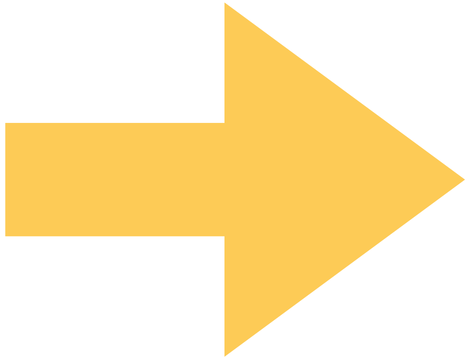
CPOP results 1: four melanoma data



Small deviation in predicted values across datasets

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

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. insert random pictures into your slides)

But what about breast cancer?

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

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

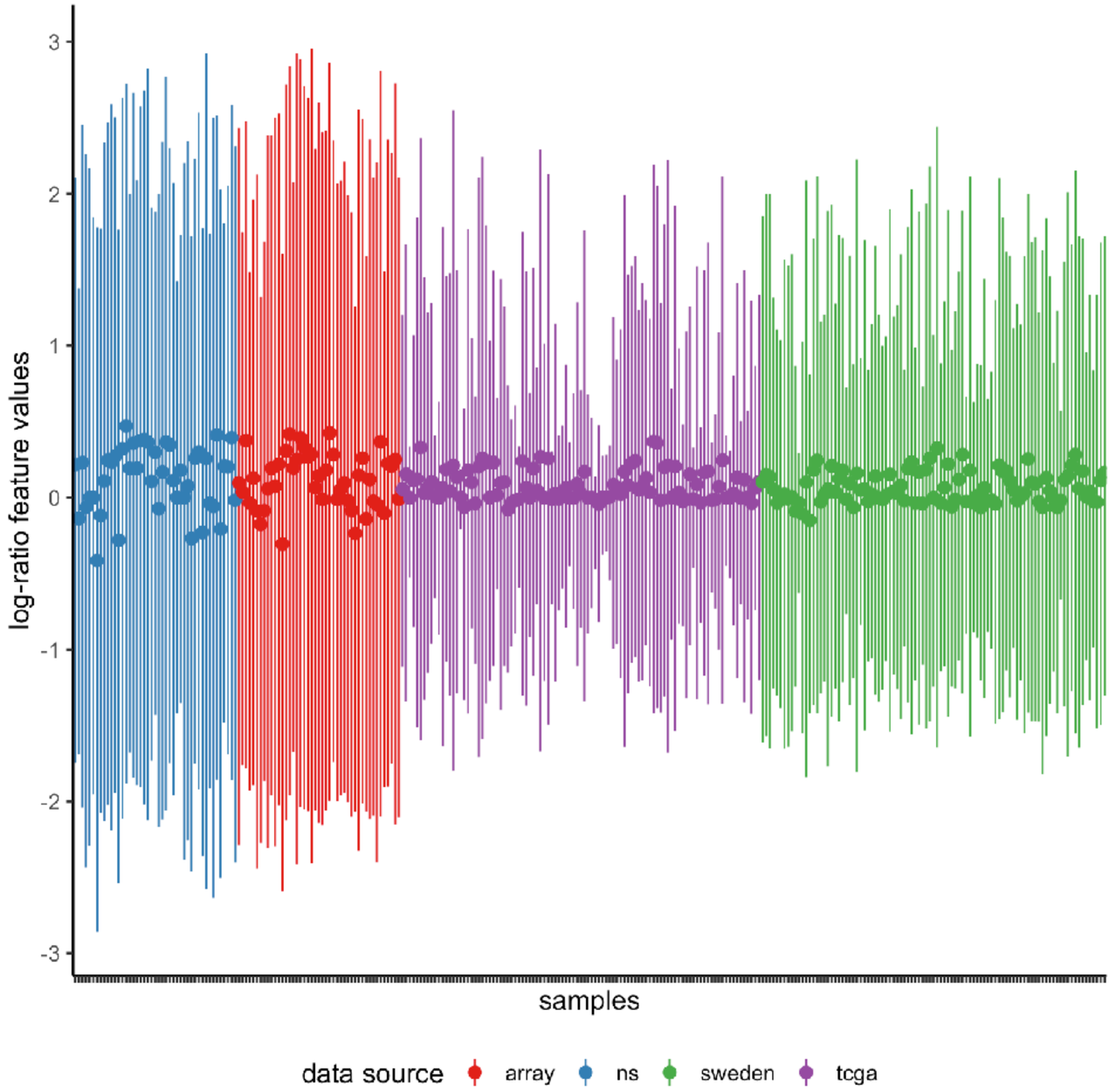
Is log-ratio really a new innovation?

Here is a list of papers that uses genes as predictors

Here is a list of papers that uses a single ratio for prediction

Our contribution is the advocation using a whole collection of ratios for prediction

This has extra implications in terms of the statistics, but we are happy to tackle these.



Within-sample feature standardisation