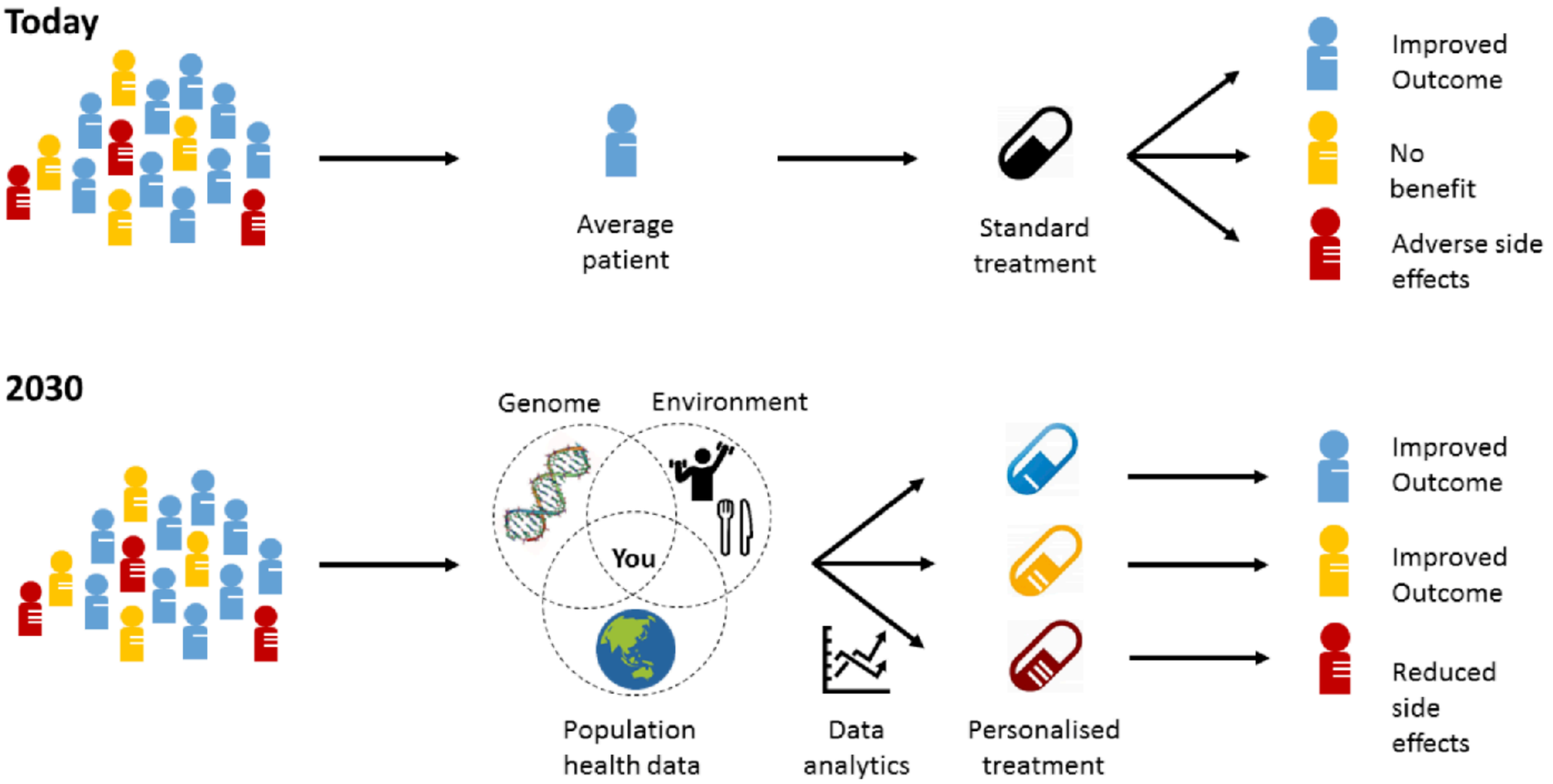
Kevin Wang

Cross-Platform Omics Prediction: a step towards precision medicine

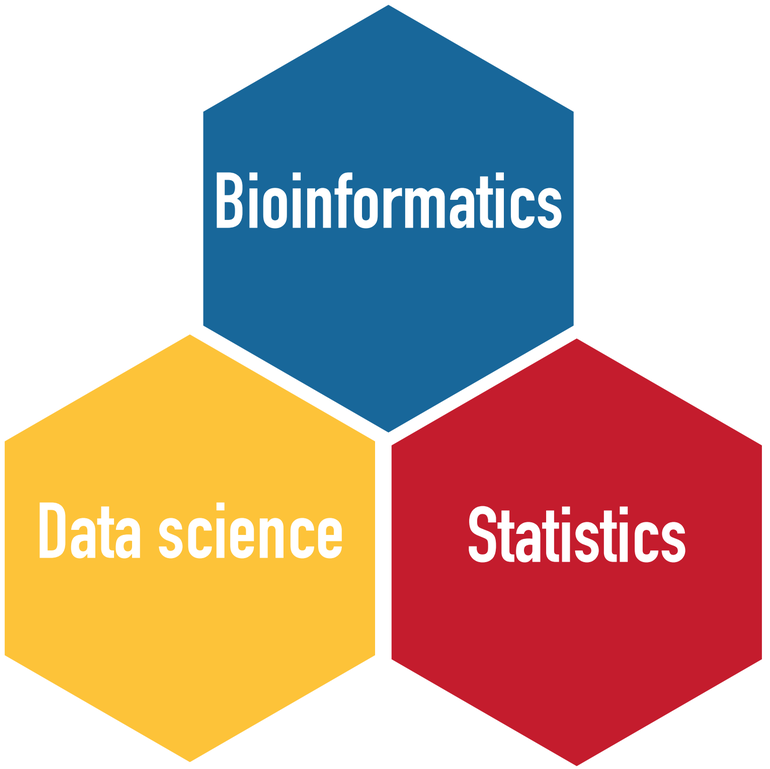
Precision medicine: predicting best cause of action using omics data

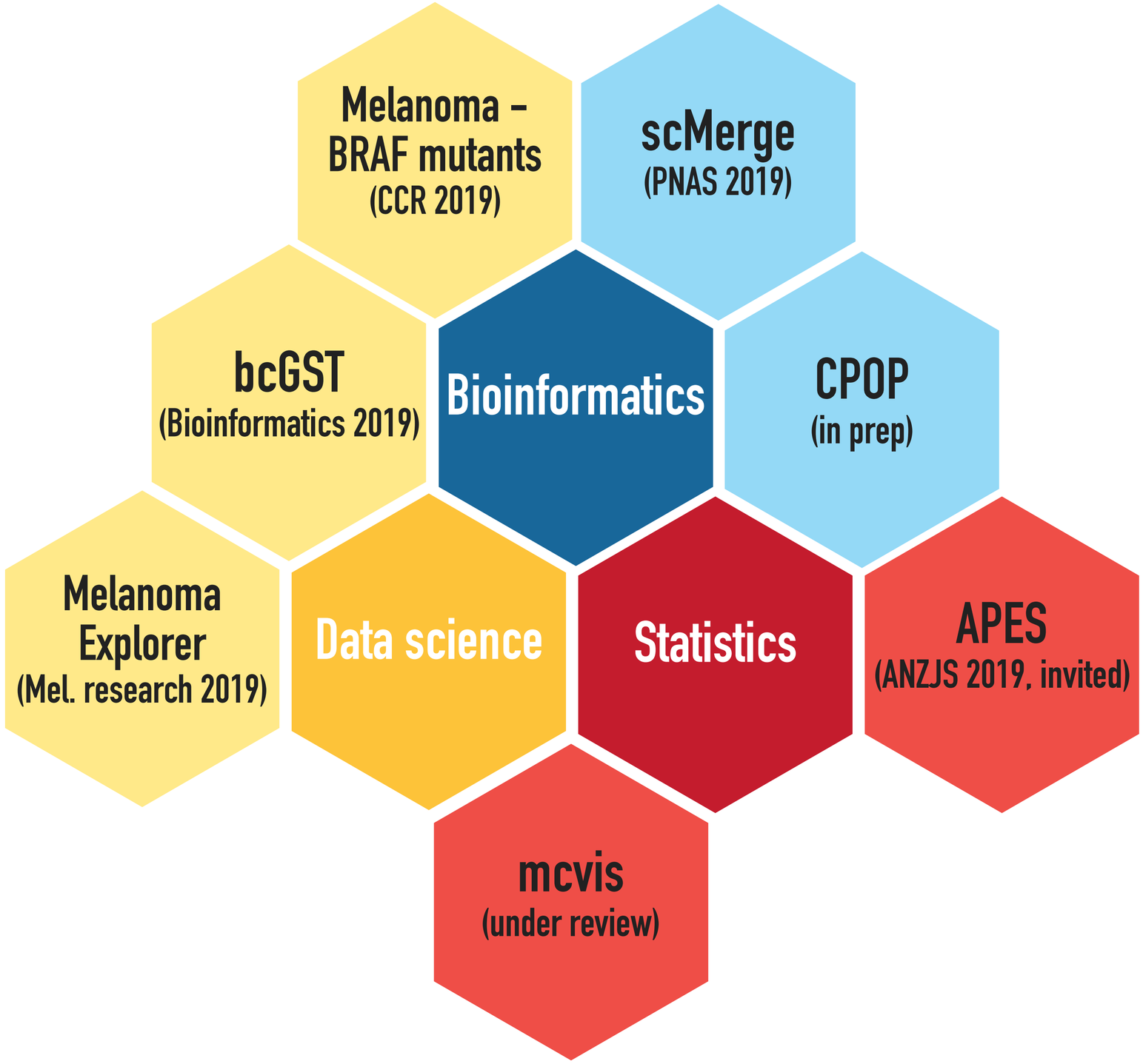
https://www.chiefscientist.gov.au/wp-content/uploads/Precision-medicine-final.pdf

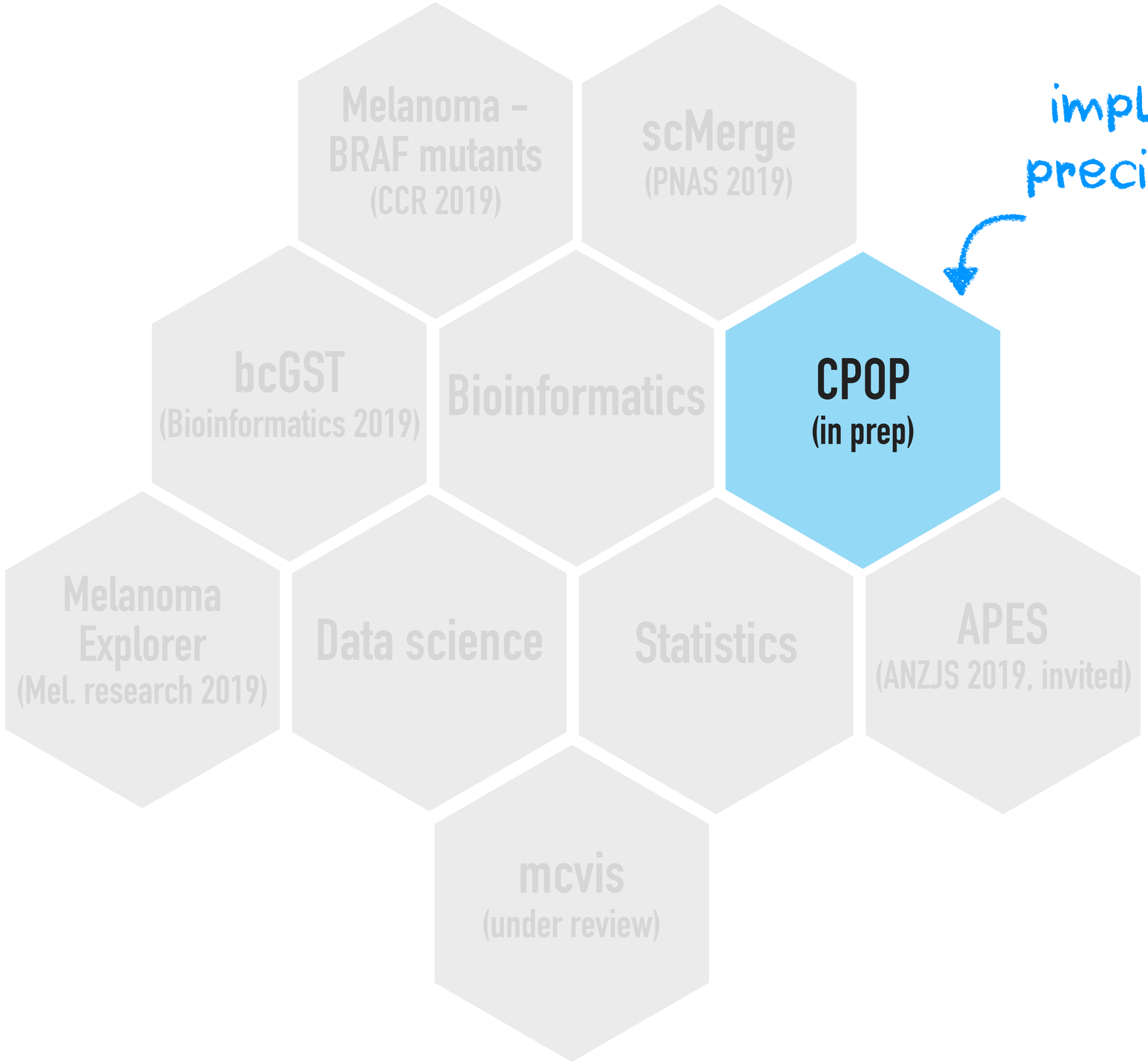
Precision medicine: predicting best cause of action using omics data



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

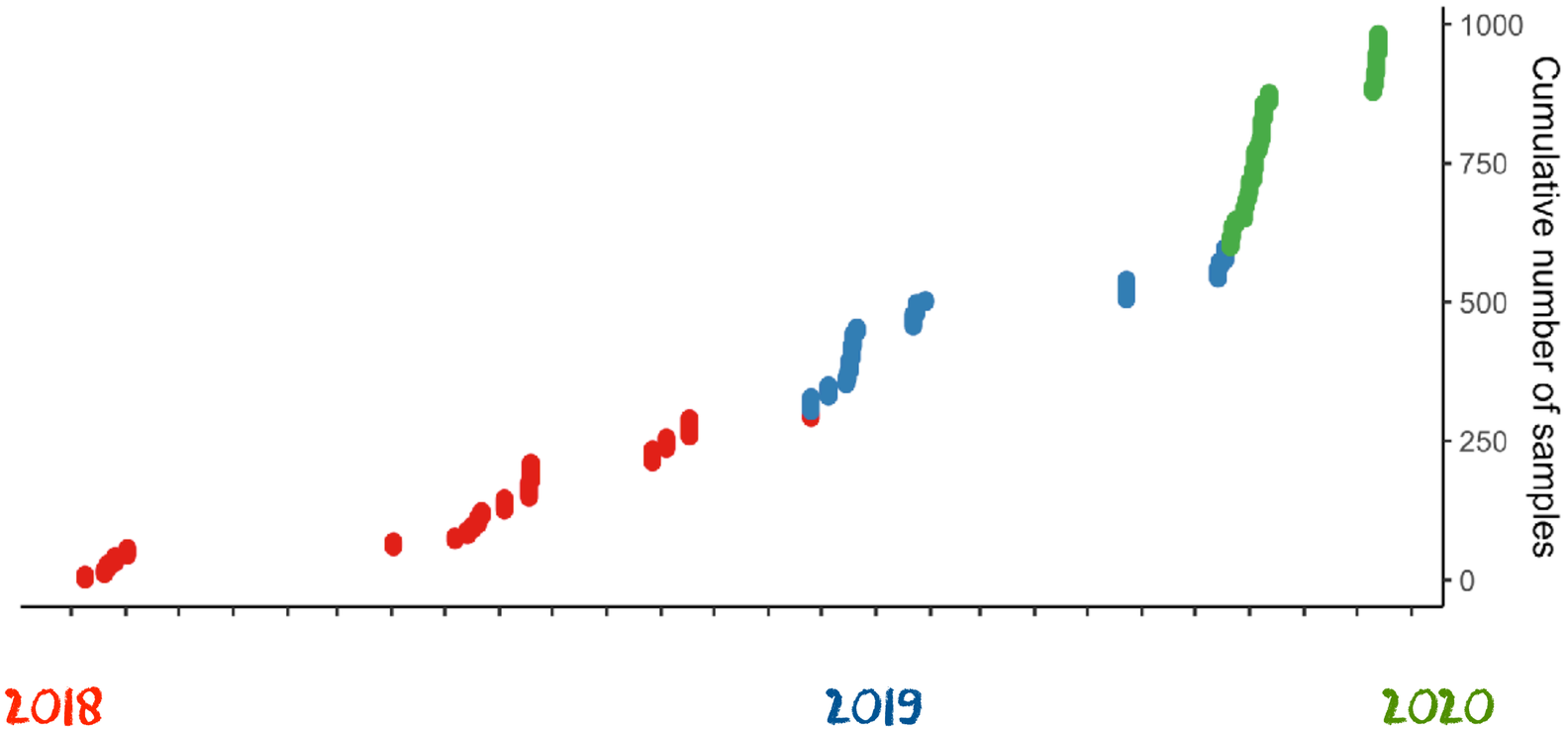

Bioinformatics
Data science
Statistics


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


Melanoma – BRAF mutants (CCR 2019)
scMerge (PNAS 2019)
bcGST (Bioinformatics 2019)
Bioinformatics
CPOP (in prep)
Melanoma Explorer (Mel. research 2019)
Data science
Statistics
APES (ANZJS 2019, invited)
mcvis (under review)
Towards implementation precision medicine

Melanoma Institute Australia

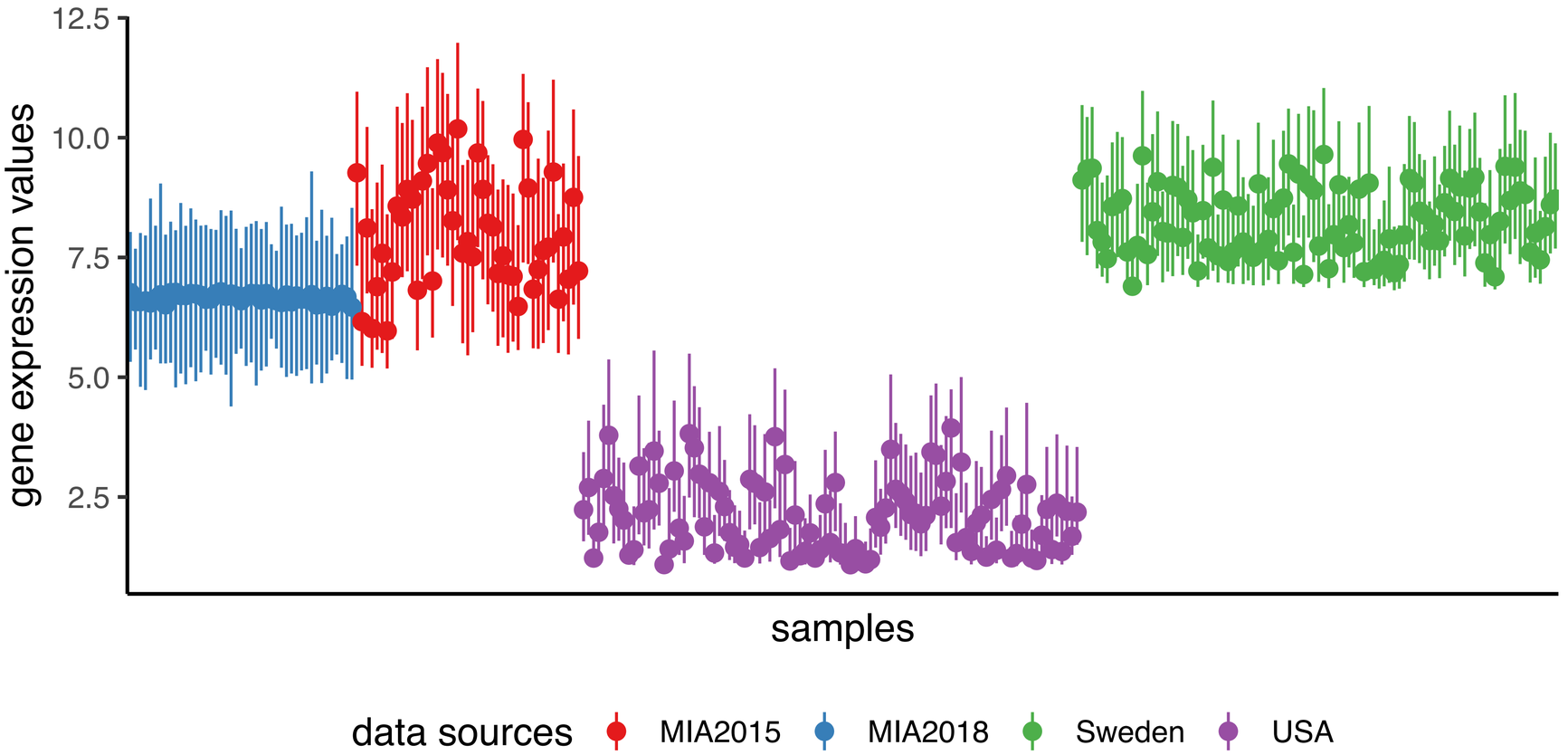
- Predict patient outcomes using gene expression:
  - prospectively



Melanoma Institute Australia

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

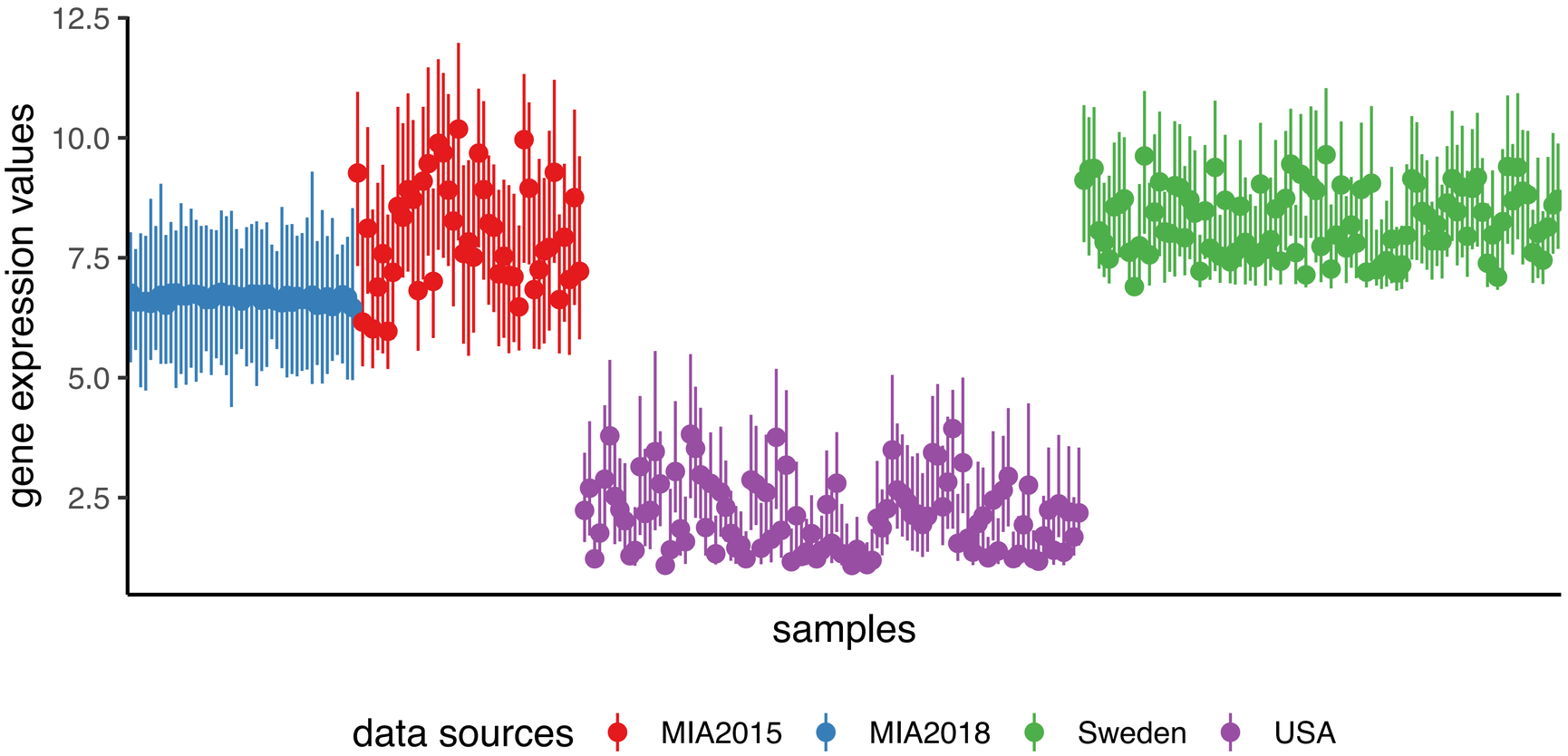
Melanoma Institute Australia



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

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

Vision of a risk score



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

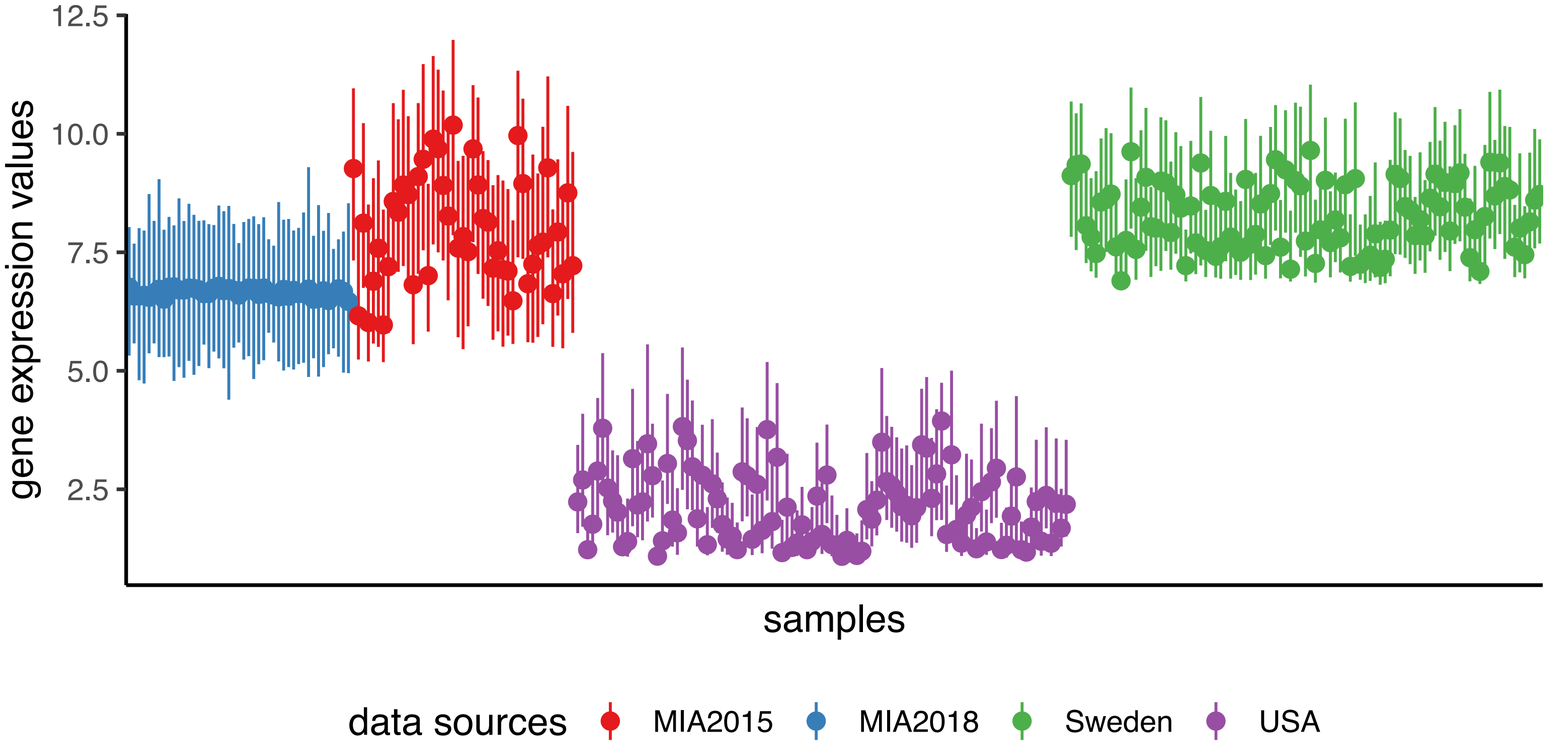
regression-based risk score

Melanoma Institute Australia

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

Omics-based risk score

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



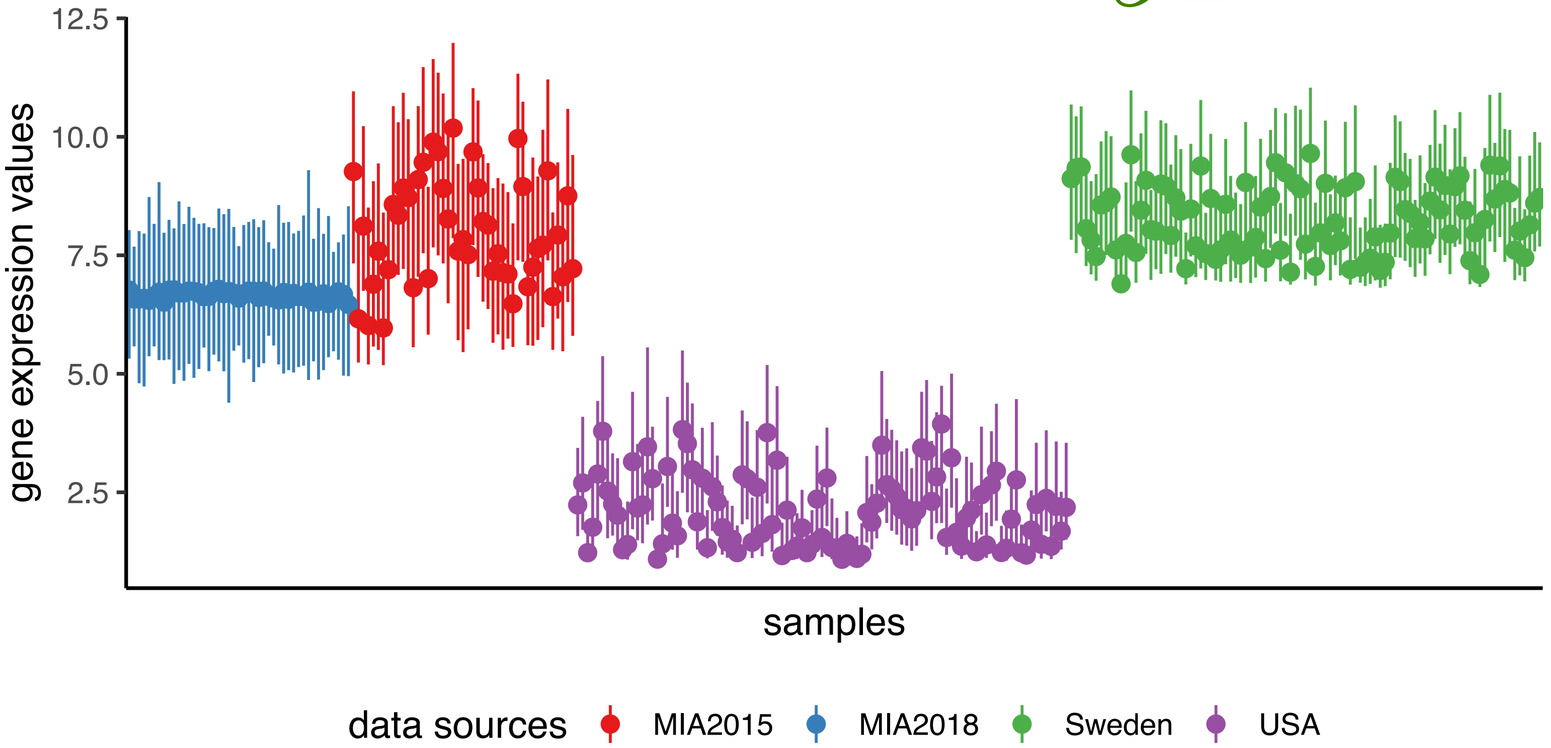
Transferability

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

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

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

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



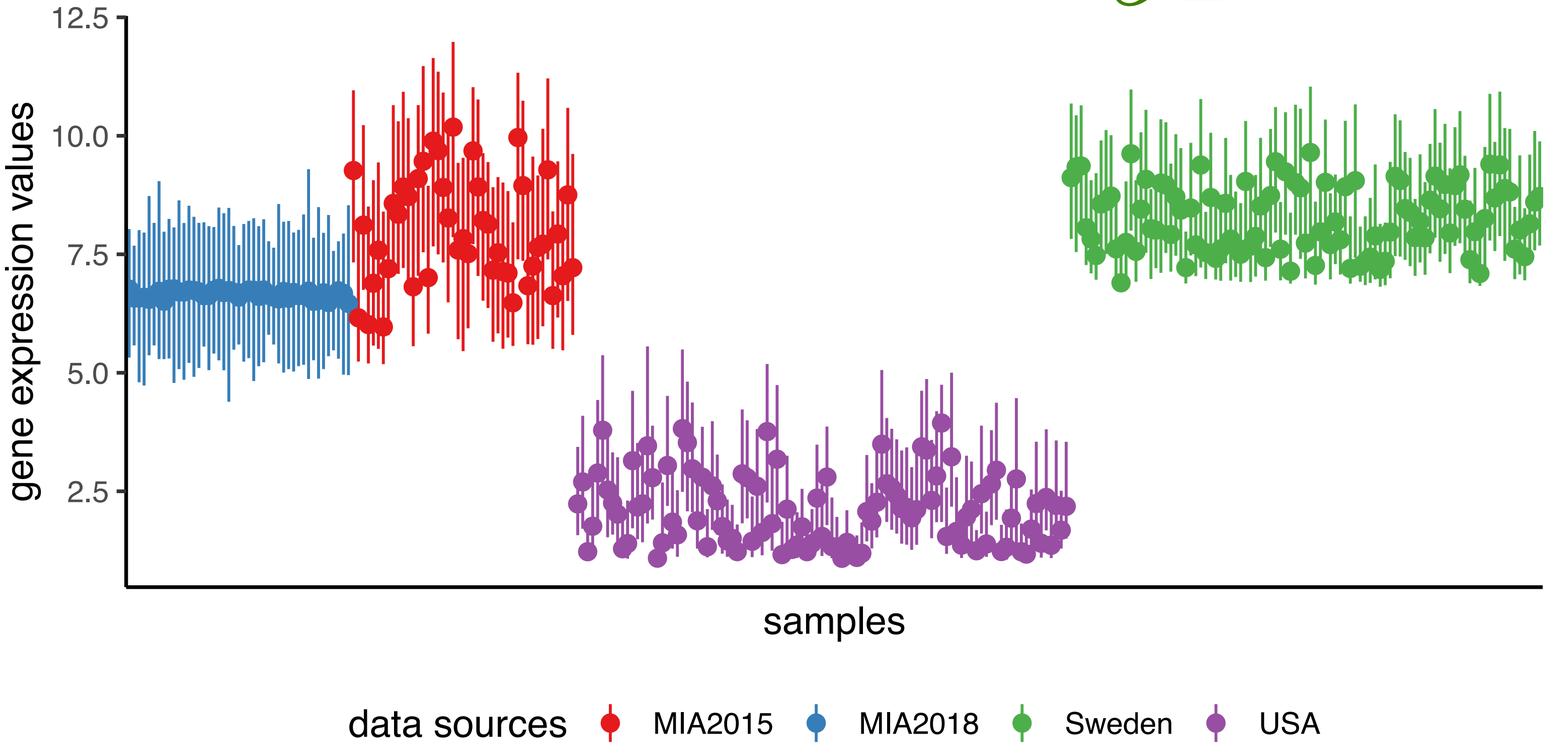
Omics features are on a relative and unit-less scale, a typical value in one data can be an impossible value on another

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

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

Assuming a noiseless shift

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



Omics features are on a relative and unit-less scale, a typical value in one data can be an impossible value on another

Components of a risk score

Data

$$(X_1, y_1)$$

$$(X_2, y_2)$$

Model

$$\hat{\beta}_1$$

Prediction

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

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

Components of a risk score

Data

$$(X_1, y_1)$$

$$(X_2, y_2)$$

Model

$$\hat{\beta}_1$$

Prediction

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

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

No renormalisation

No model retraining

Scale-equivalent prediction

Existing approach vs CPOP

Existing approach: data-harmonisation

Data

(X_1, y_1)

(X_2, y_2)

Harmonisation

Model

$\hat{\beta}_1$

Classical estimation methods

Prediction

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

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

Classical model prediction

1. Data-harmonisation: normalisation or standardisation

Data

(X_1, y_1)

(X_2, y_2)

Harmonisation

Introduces strong dependence between training samples and models

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

2. Model-based approach: CPOP

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

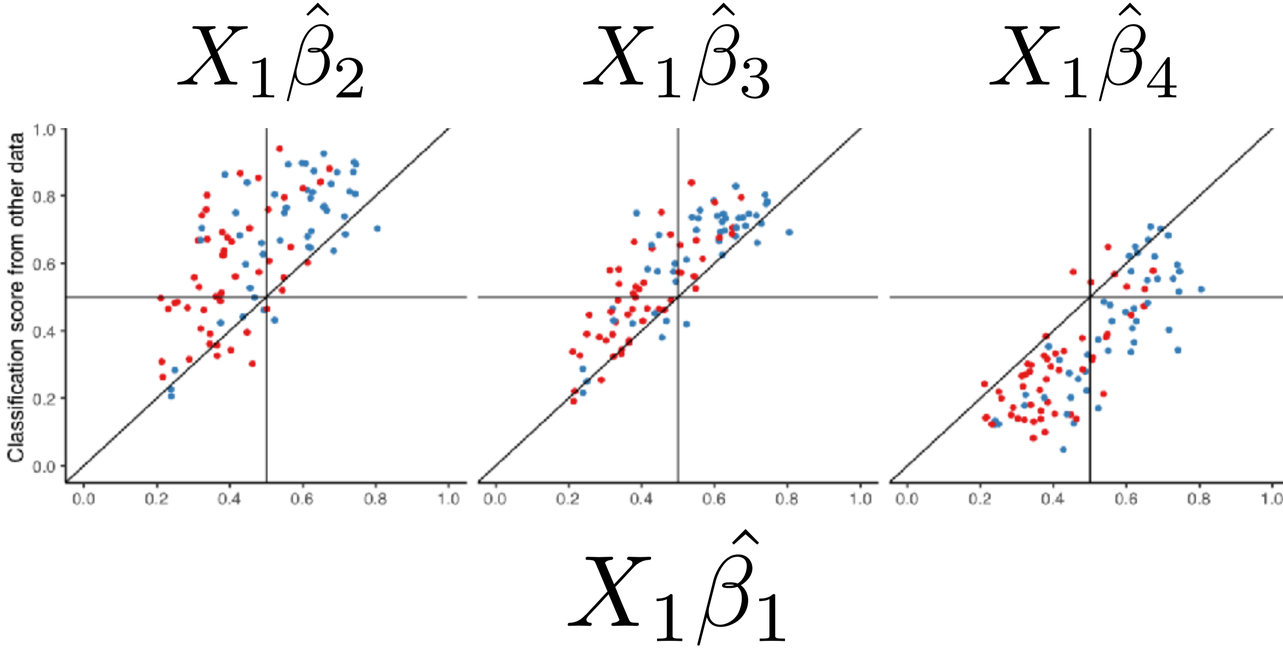
Statistical challenges

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



First component of CPOP: feature transform

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

Log-ratio transformation

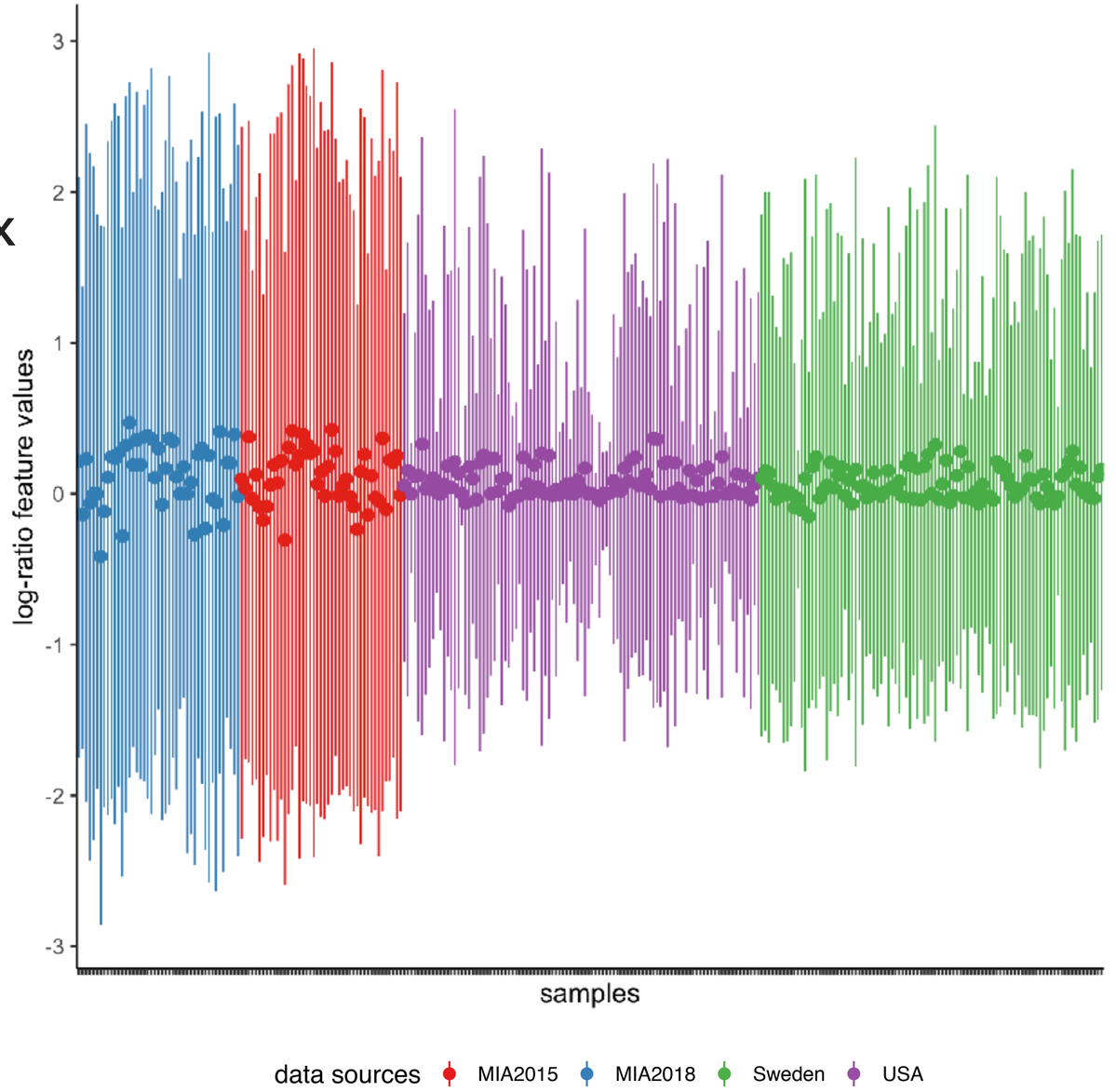
For each column in the gene expression matrix

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

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

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

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

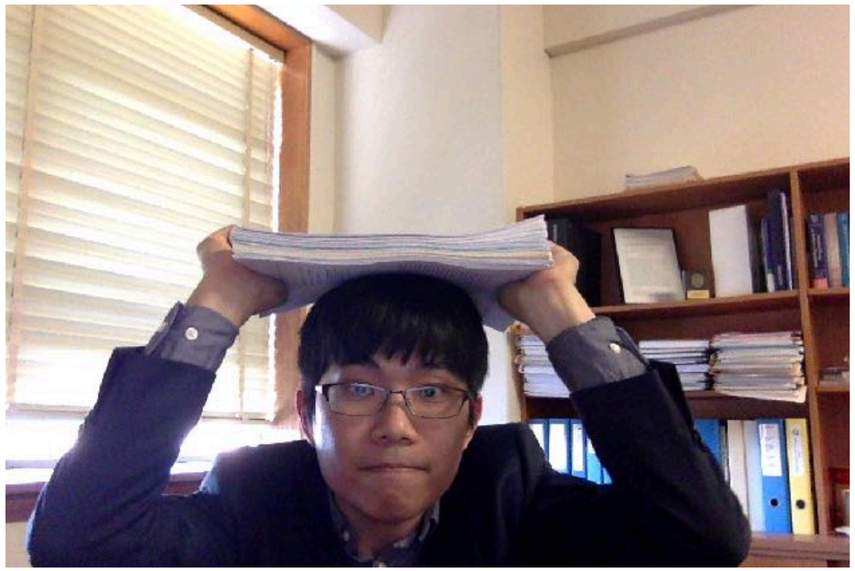


Why log-ratios?

- Within-sample standardisation avoids re-normalisation and model re-training
- Under-used in patient outcome prediction
- Potentials for further method developments

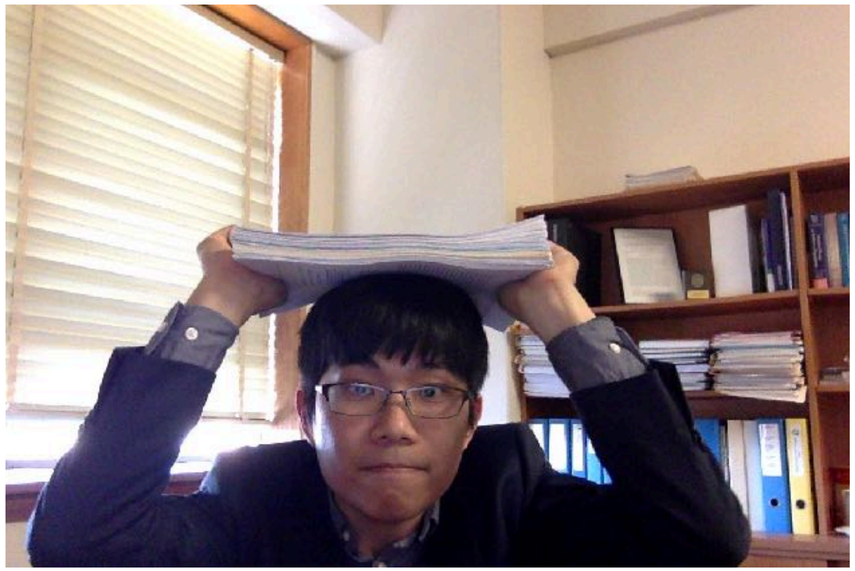
Why log-ratios?

- Within-sample standardisation avoids re-normalisation and model re-training
- Under-used in patient outcome prediction
- Potentials for further method developments

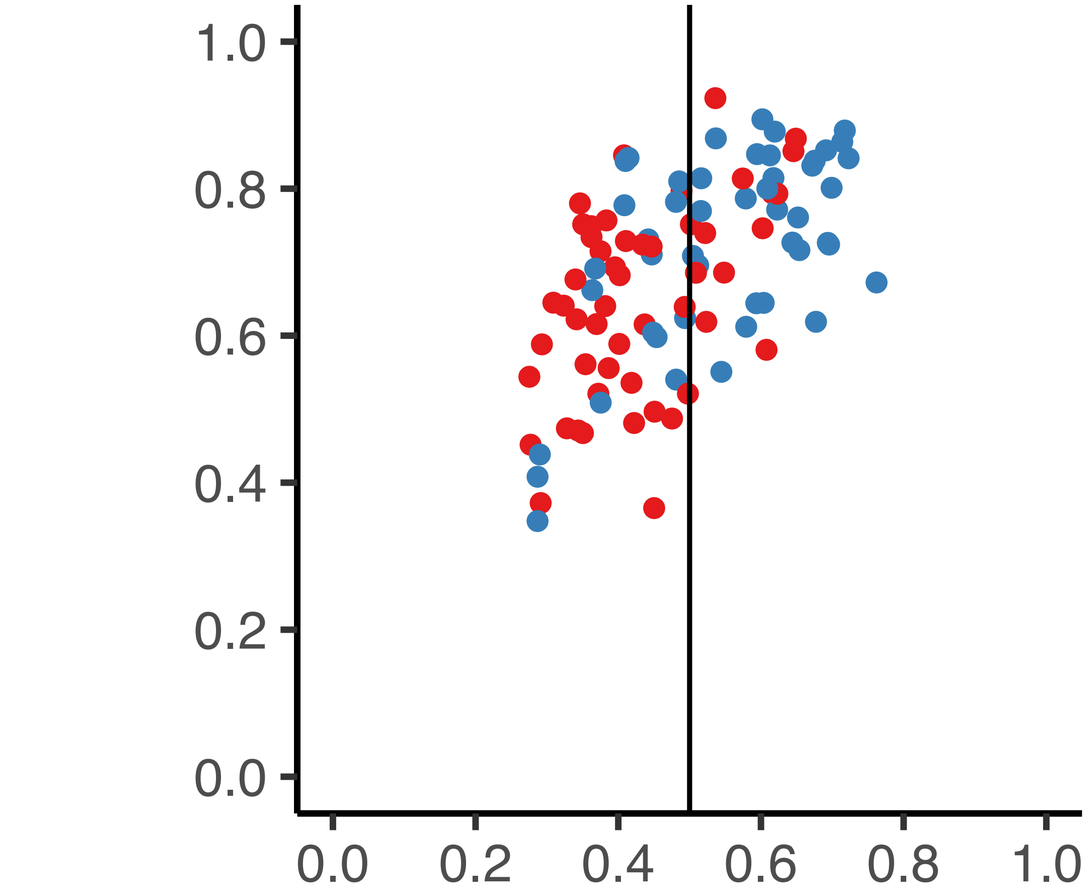


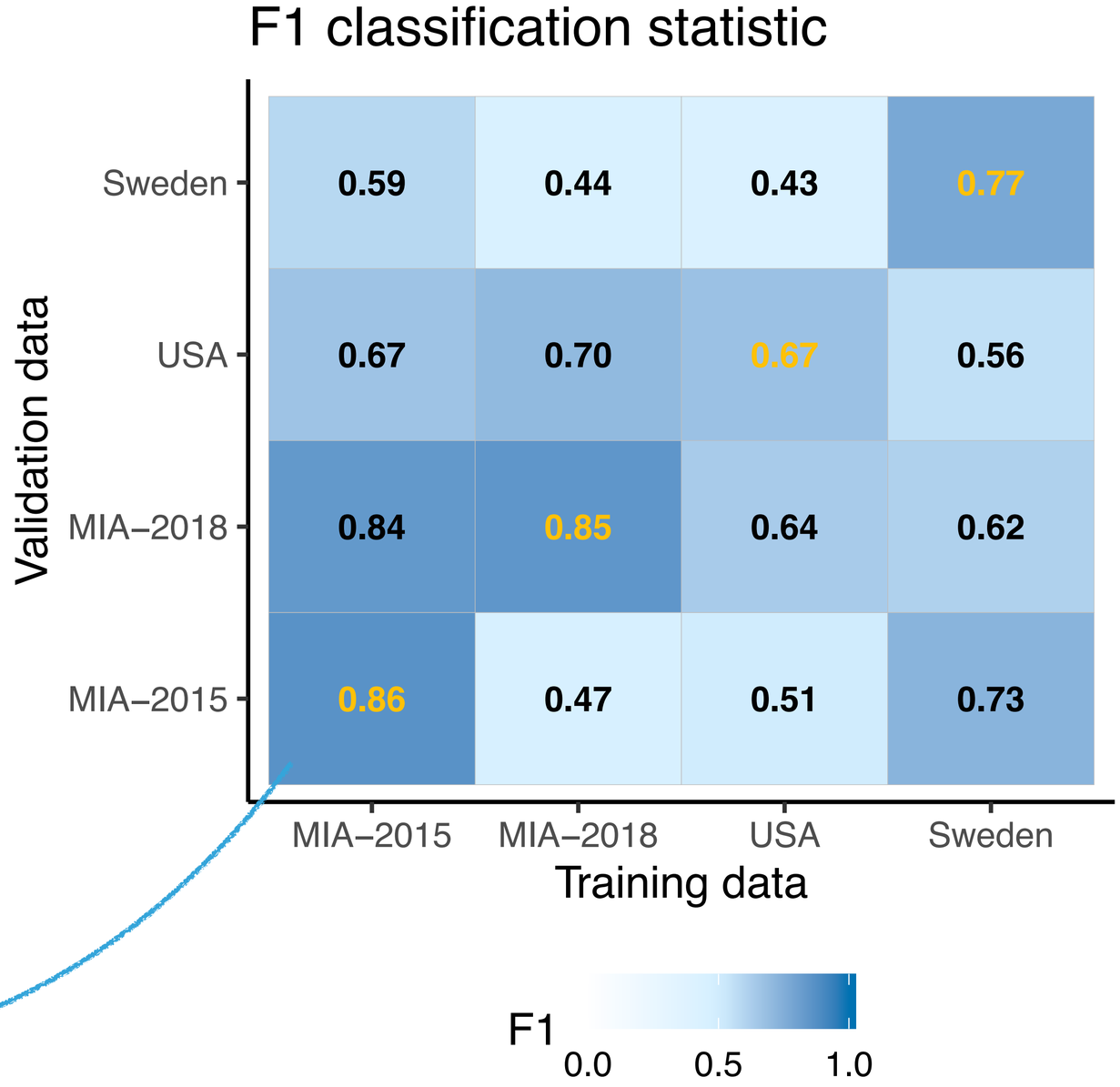
Why log-ratios?

- Within-sample standardisation avoids re-normalisation and model re-training
- Under-used in patient outcome prediction
- Potentials for further method developments

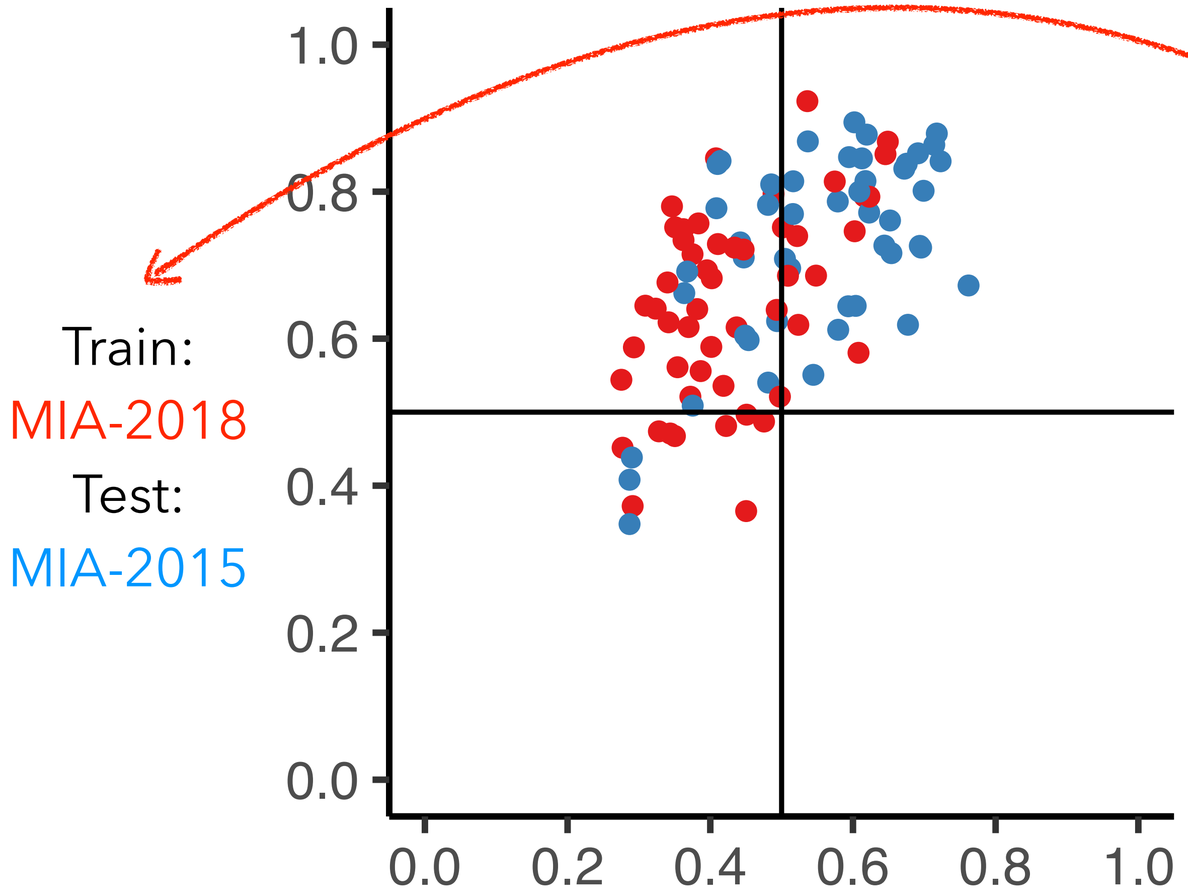


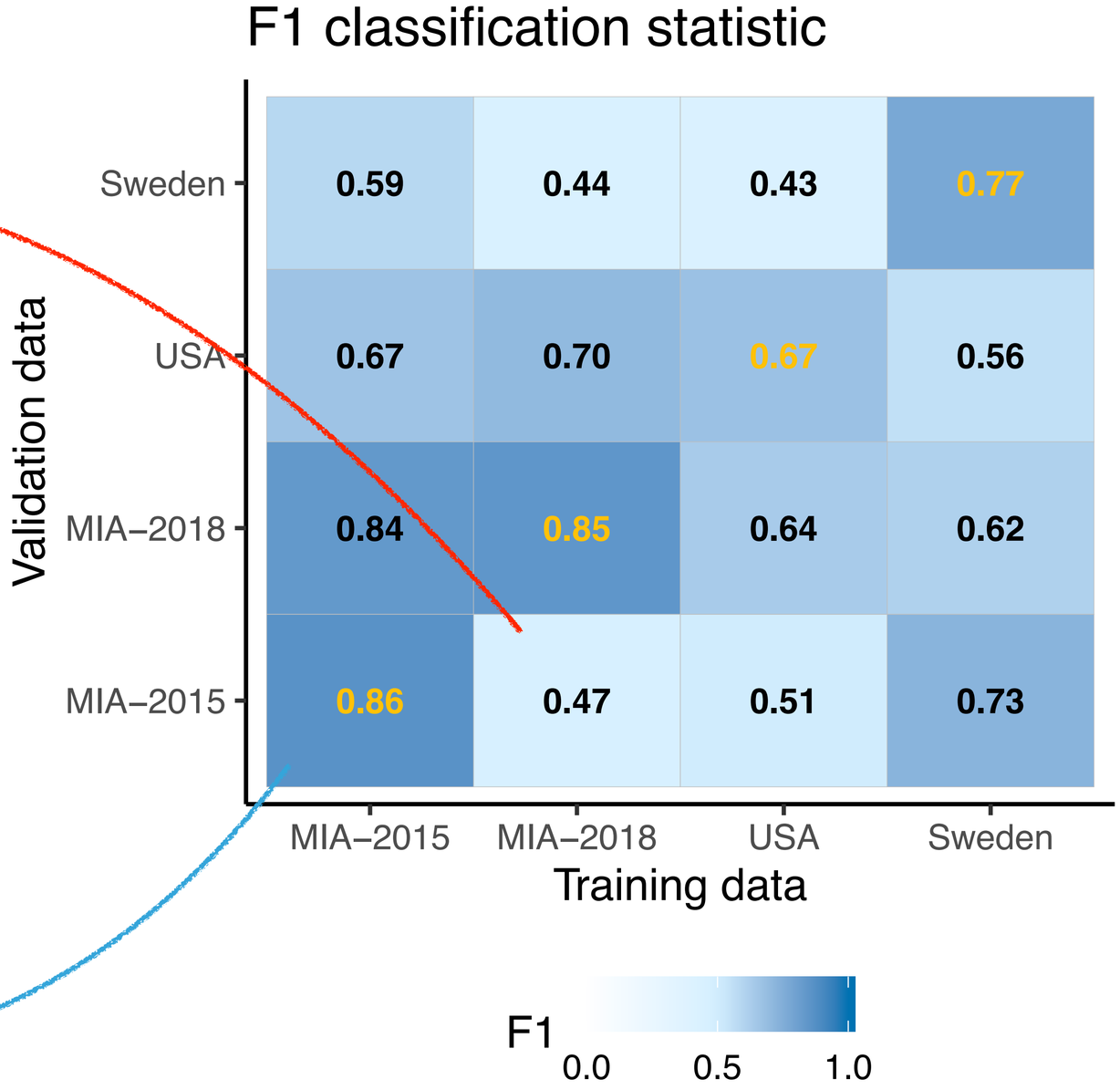
Is log-ratio enough?



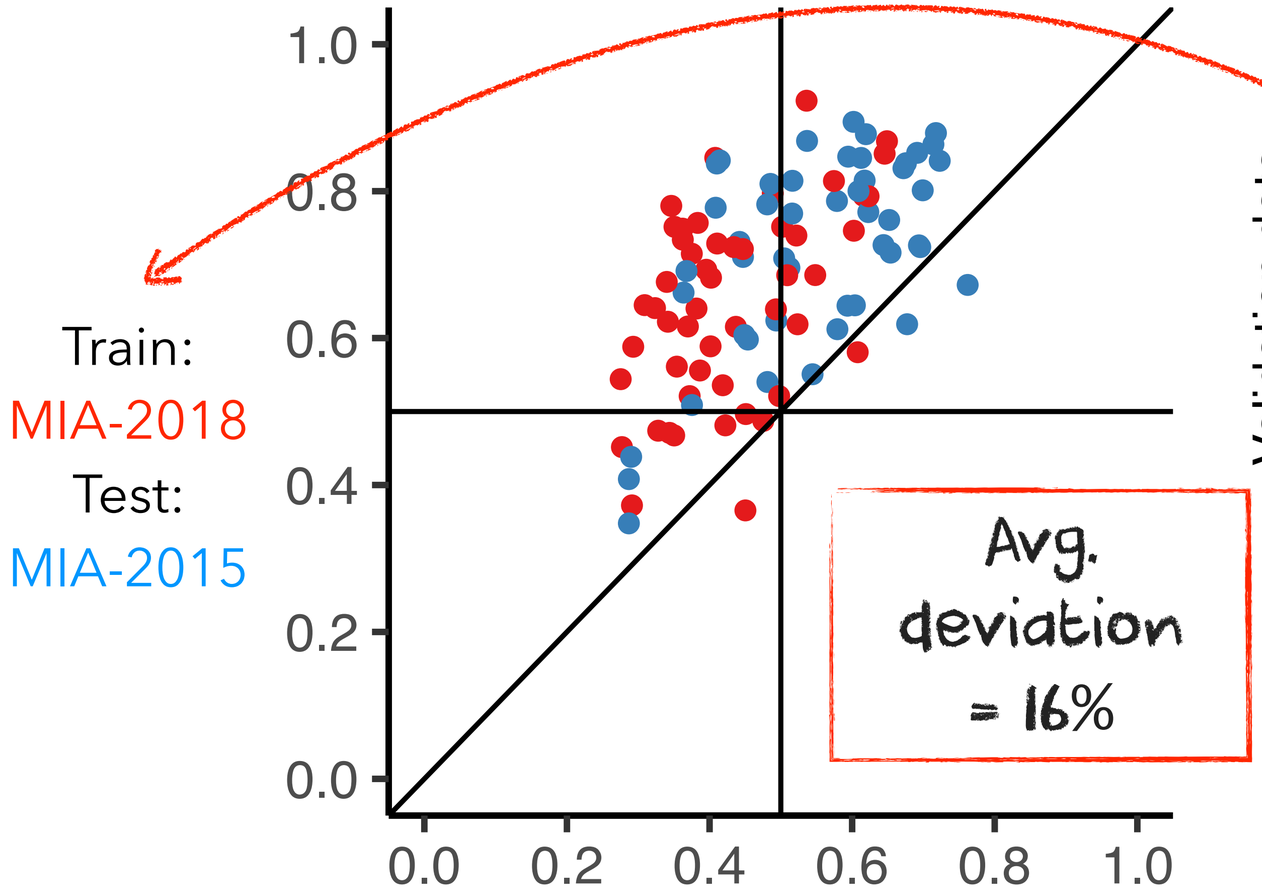


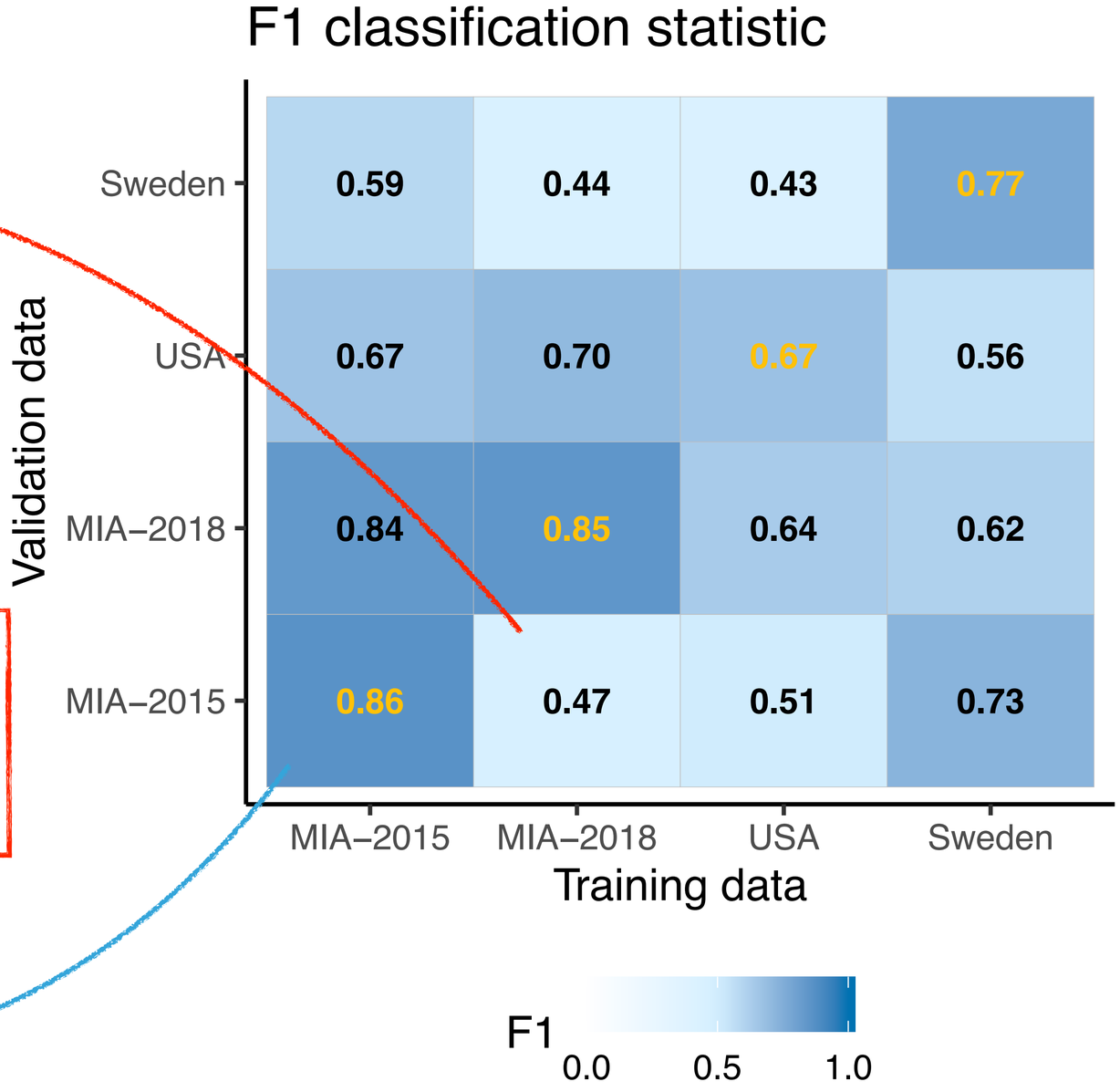
Is log-ratio enough?





Is log-ratio enough?

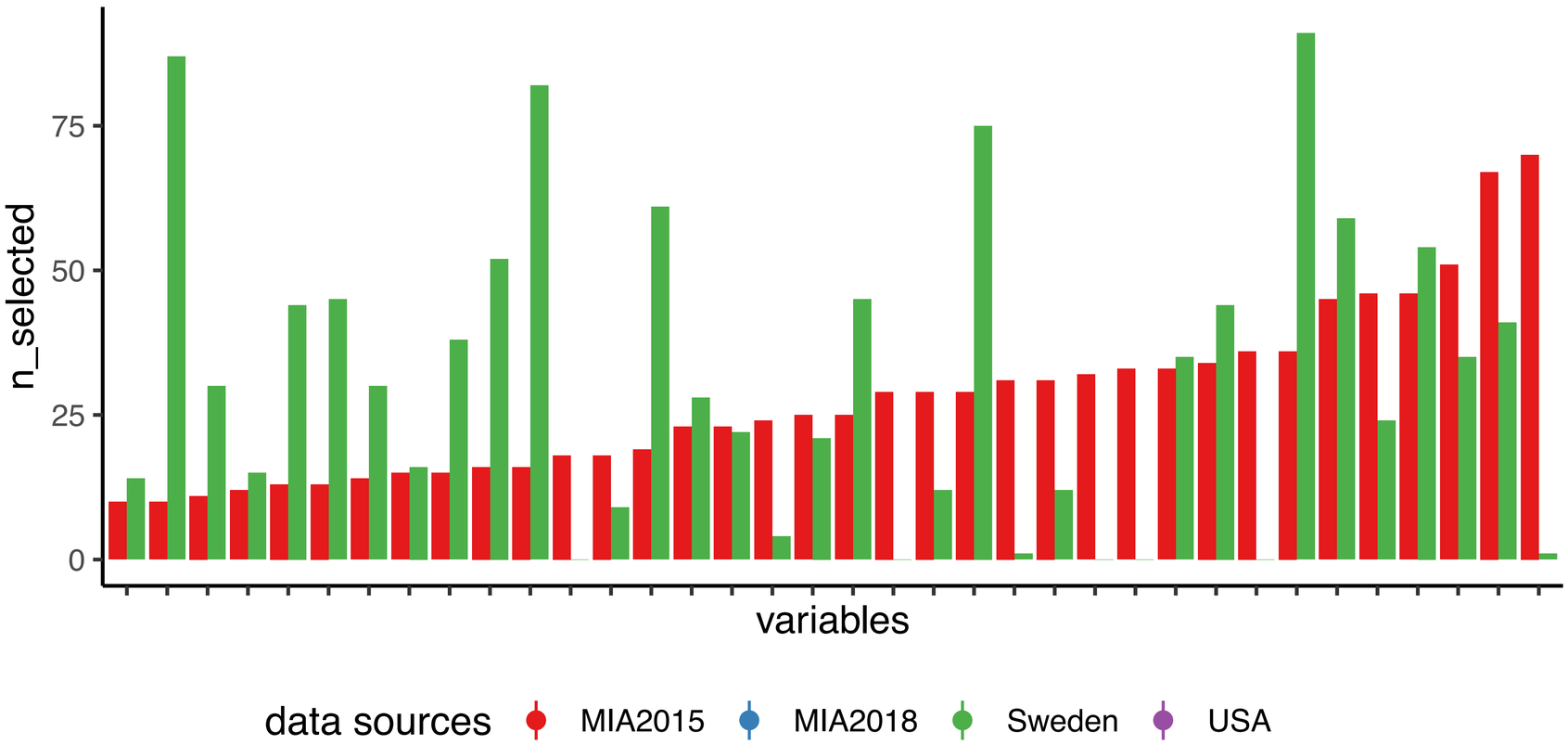




Lasso variable selection is not stable!

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

all pointed out that Lasso is unstable under cross-validation



Second component of CPOP: stable feature selection and estimation

我曾经毁了我的一切
只想永远地离开
我曾经陷入无边黑暗
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

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

Loss function, could be logistic or Cox

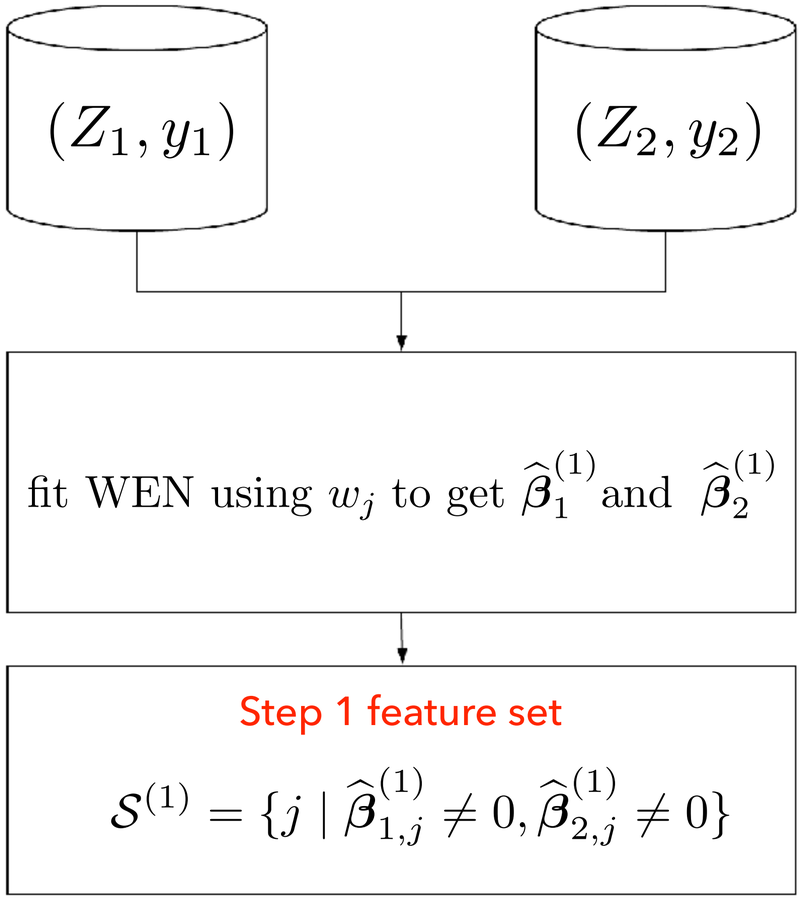
A mix of L1 and L2 penalties

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

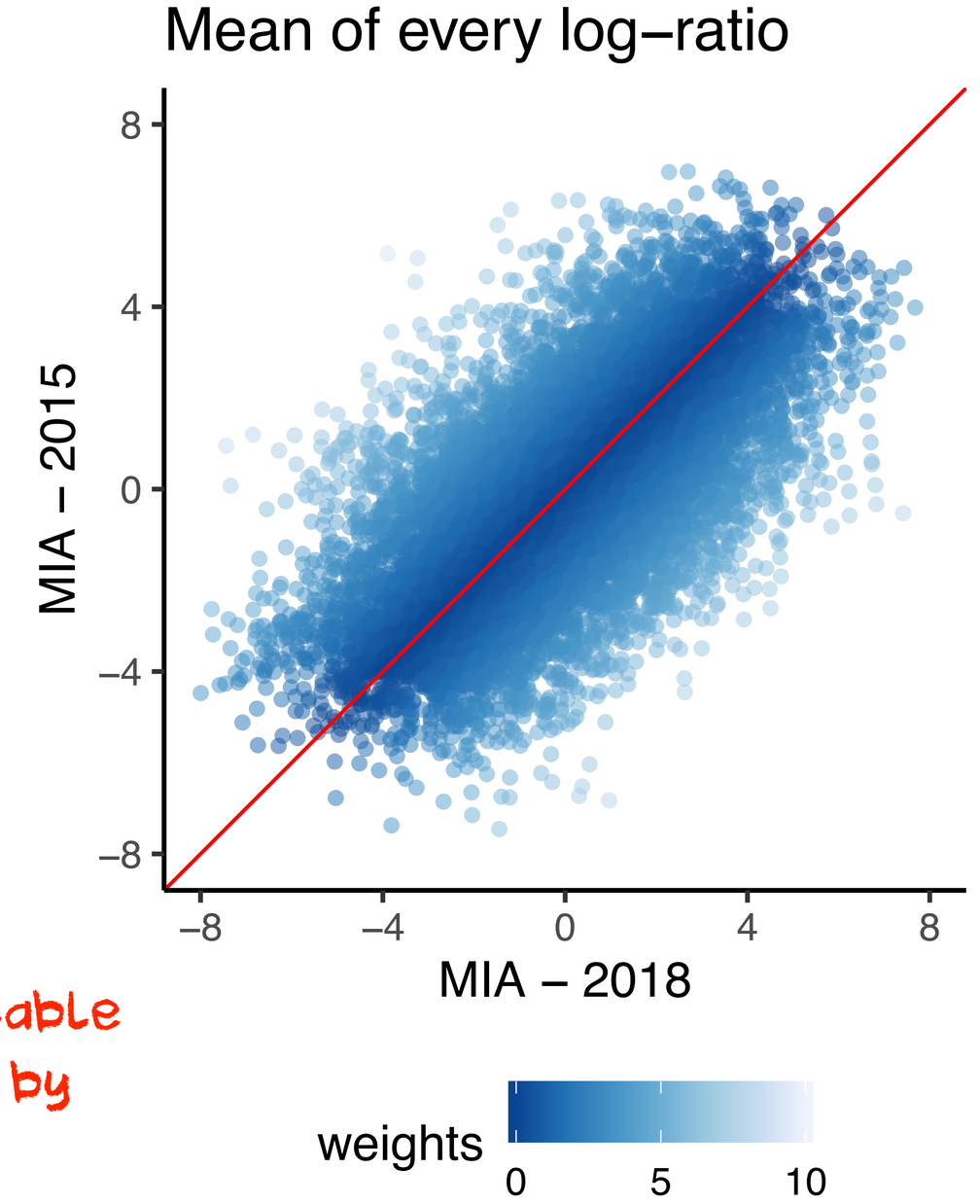
Modelling on log-ratio

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

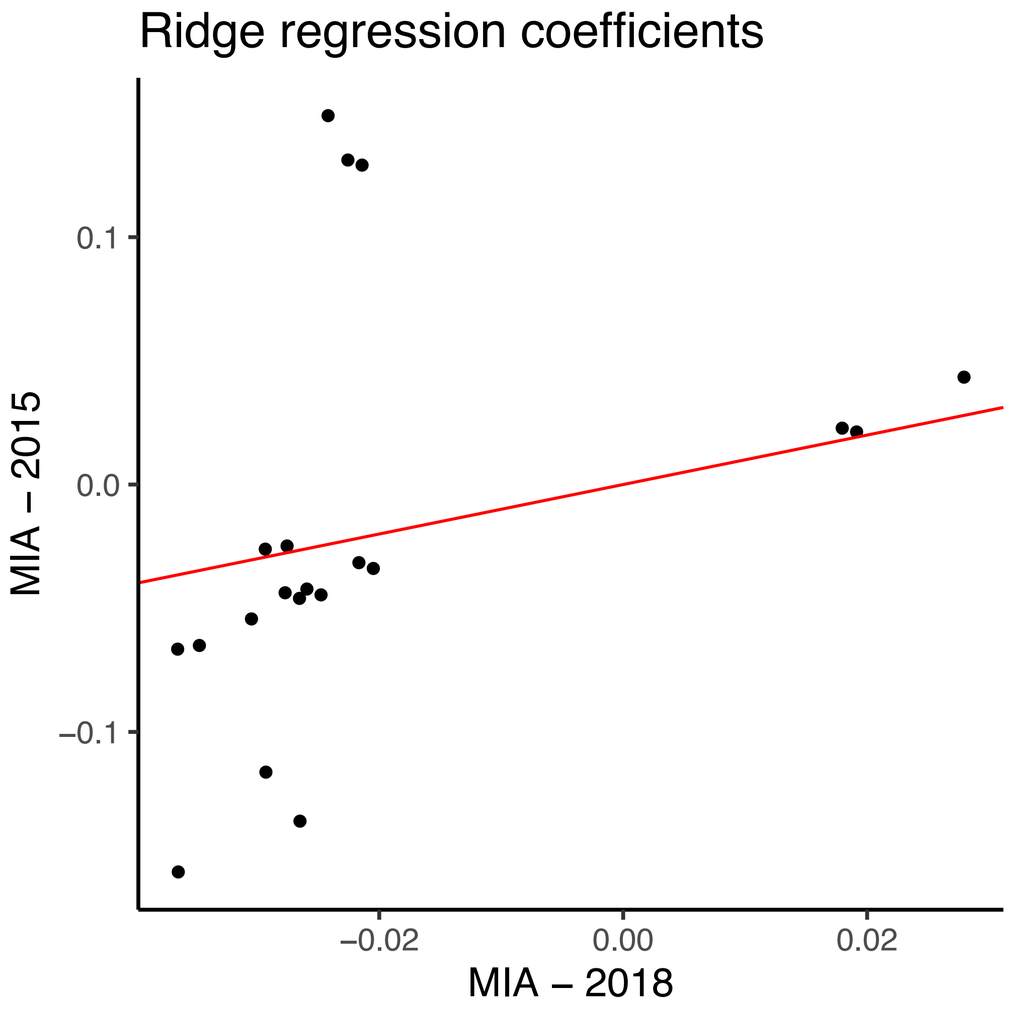
$$\mathcal{S}^{(1)} = \{j \mid \widehat{\boldsymbol{\beta}}^{(1)}_{1,j} \neq 0, \widehat{\boldsymbol{\beta}}^{(1)}_{2,j} \neq 0\}$$

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

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

Step 2 feature set

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



Step 2: feature estimation stability

Step 1 feature set

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

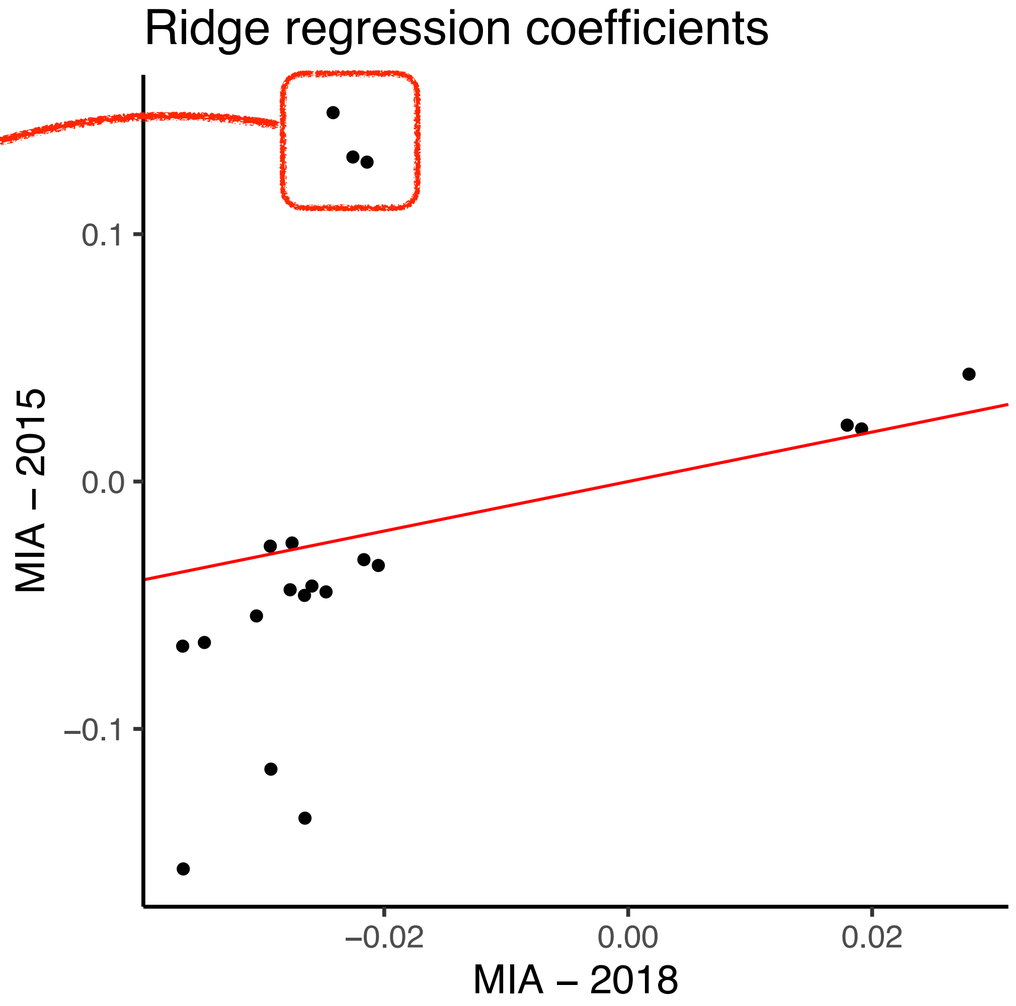
$(\boldsymbol{Z}_{1,\mathcal{S}^{(1)}}, \boldsymbol{y}_1)$ $(\boldsymbol{Z}_{2,\mathcal{S}^{(1)}}, \boldsymbol{y}_2)$

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

Step 2 feature set

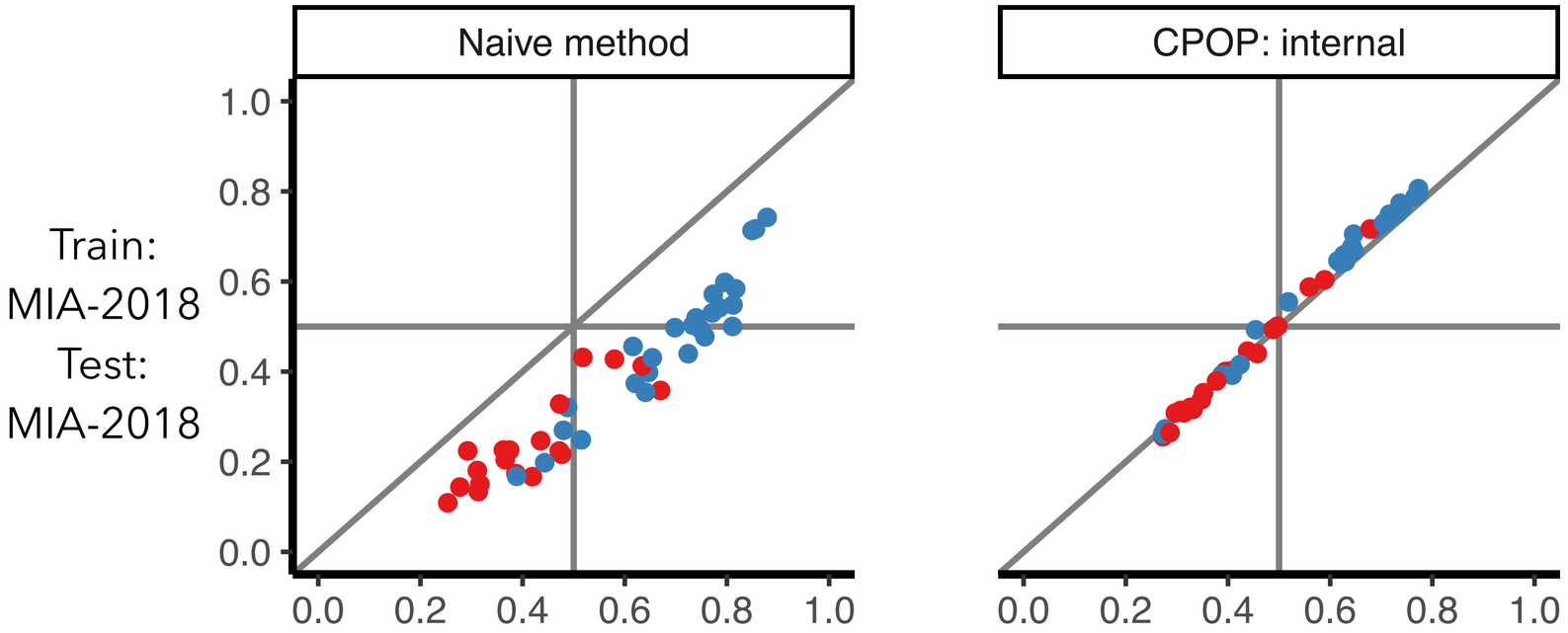
$$\mathcal{S}^{(2)} = \{j \mid \text{sign}(\widehat{\boldsymbol{\beta}}^{(2)}_{1,j}) = \text{sign}(\widehat{\boldsymbol{\beta}}^{(2)}_{2,j})\}$$

features with unstable estimates across data



Results

CPOP results 1: MIA 2015 vs 2018 data

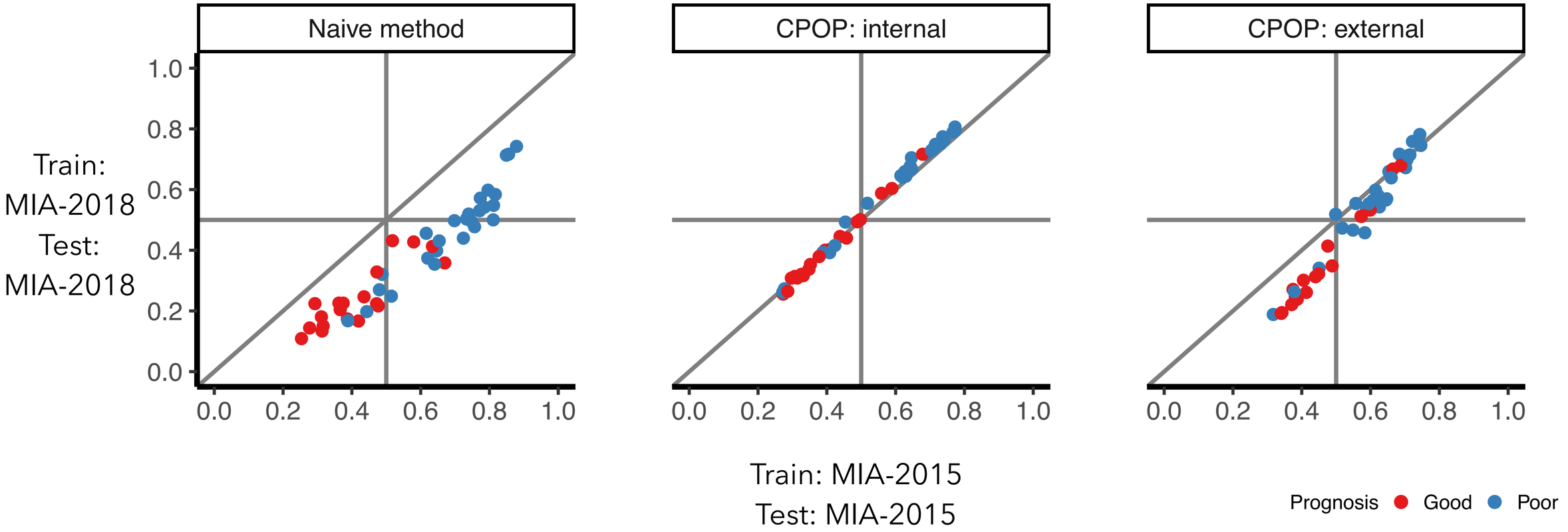


Train: MIA-2015
Test: MIA-2015

Prognosis Good Poor

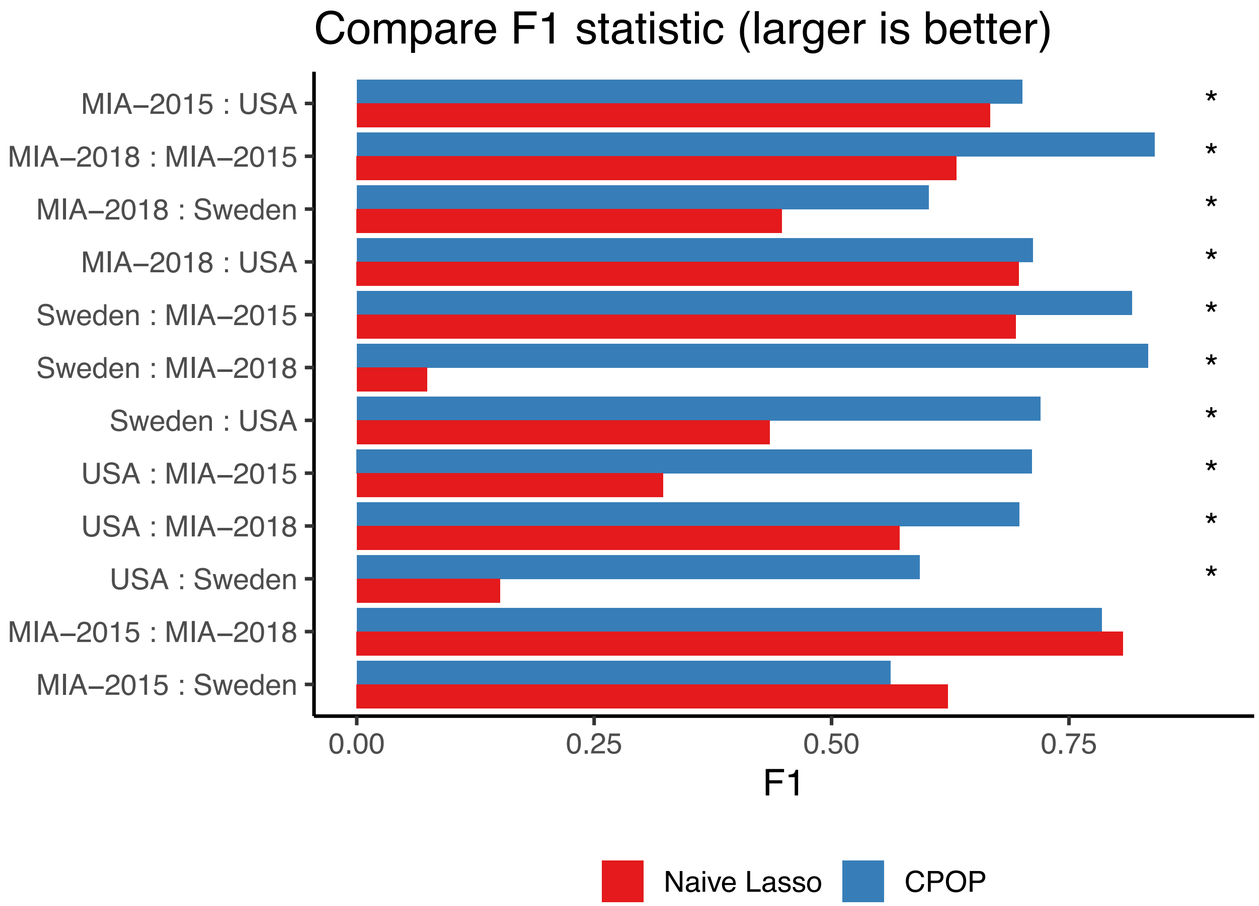
Small deviation in predicted values across datasets

CPOP results 1: MIA 2015 vs 2018 data



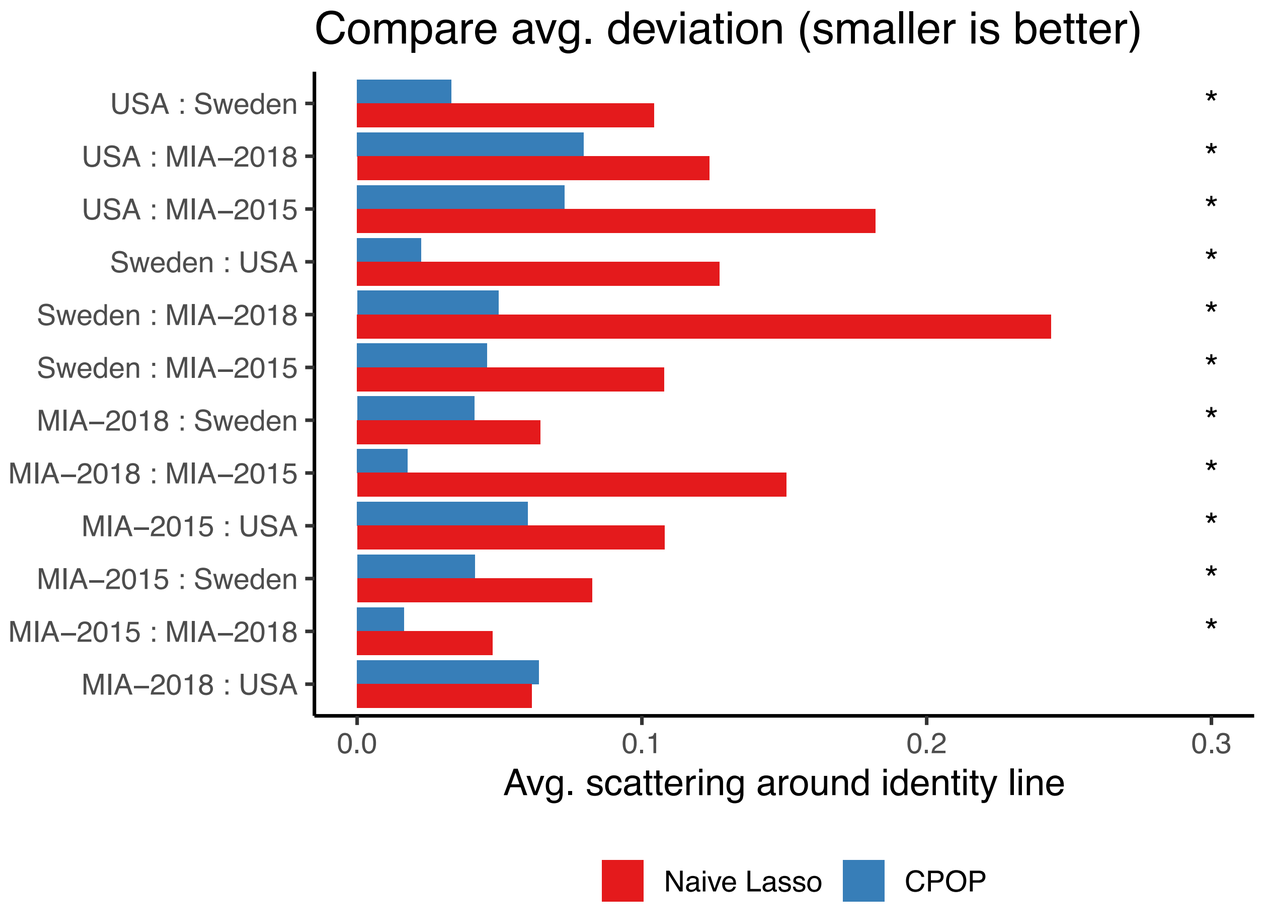
Small deviation in predicted values across datasets

CPOP results 2: four melanoma data



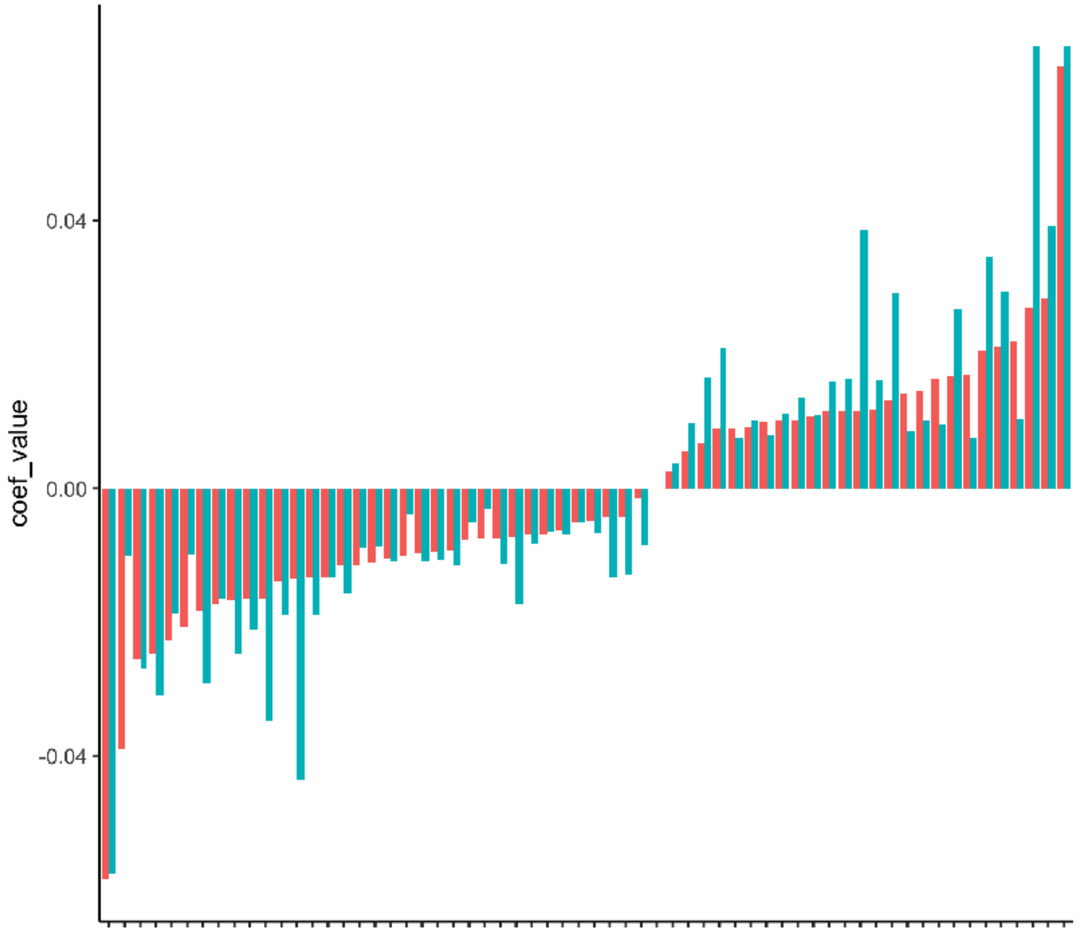
CPOP is highly predictive

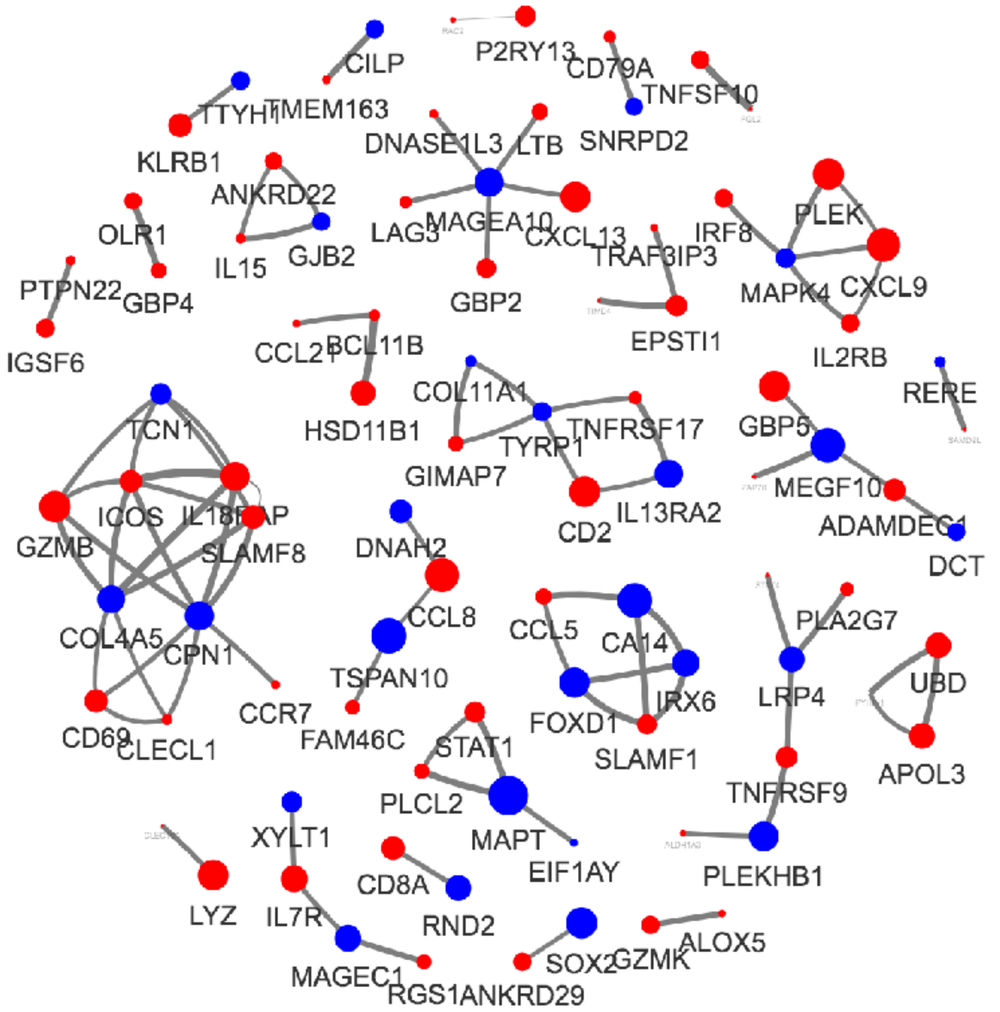
CPOP results 2: four melanoma data



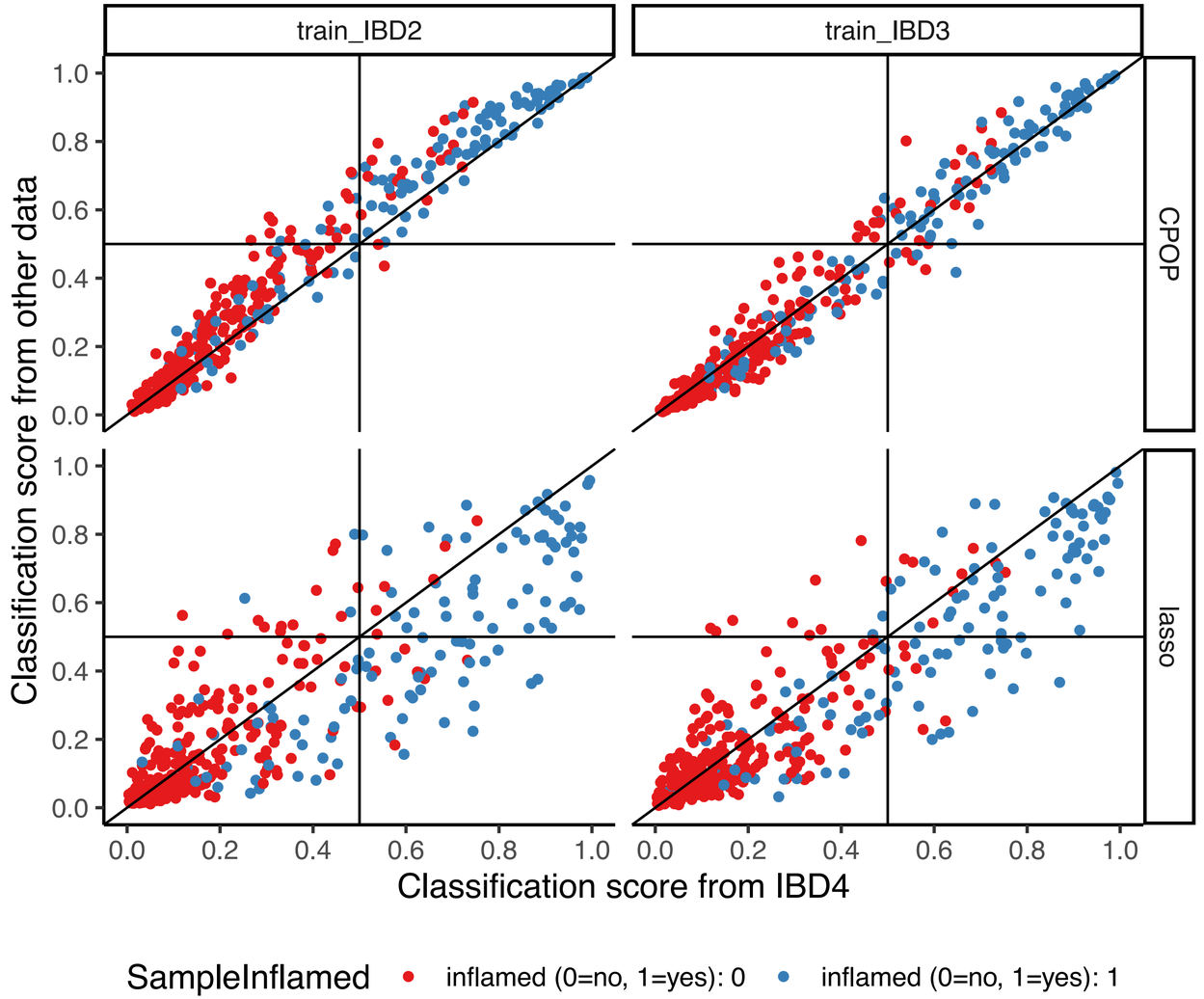
CPOP is highly stable

CPOP features





CPOP results 3: prospective prediction on inflammatory bowel disease



Concluding remarks

- Integrates clinical implementation constraints into the model

    1. Log-ratios enable prospective and multi-centres prediction

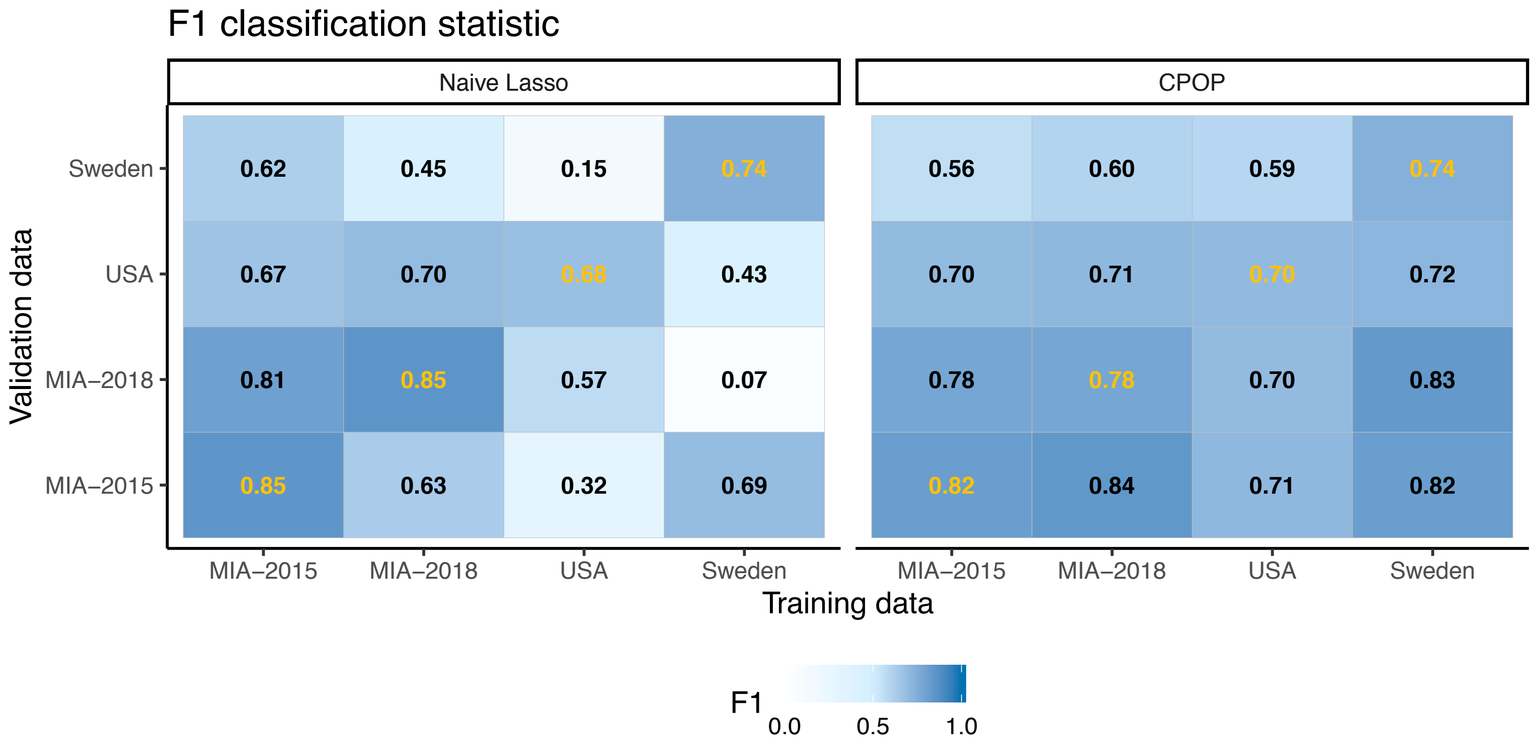
    2. Stable variable selection and estimation components

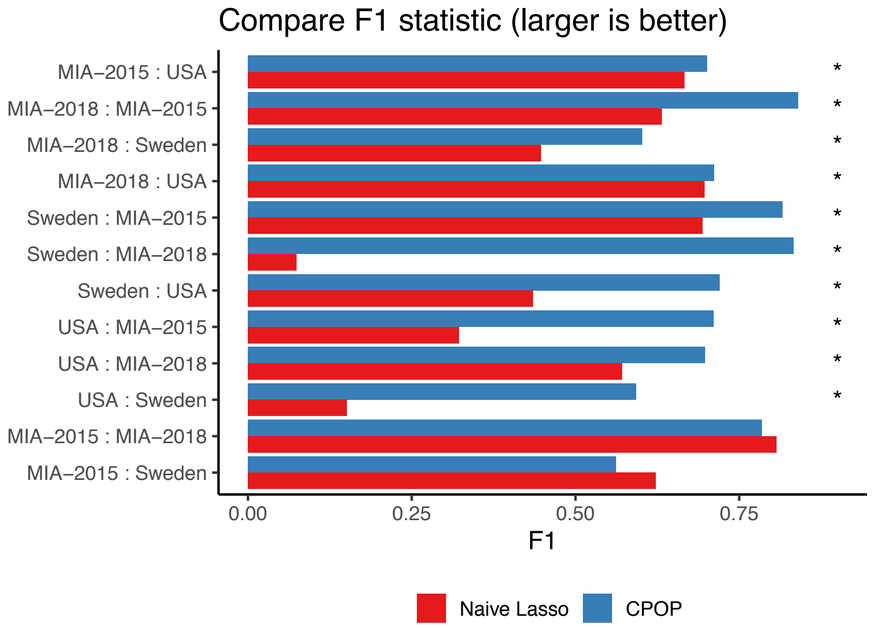
- A flexible framework with many adaptable components

- Potentials to handle data with higher relevance to precision medicine (e.g. drug sensitivity)

I think that is all.

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

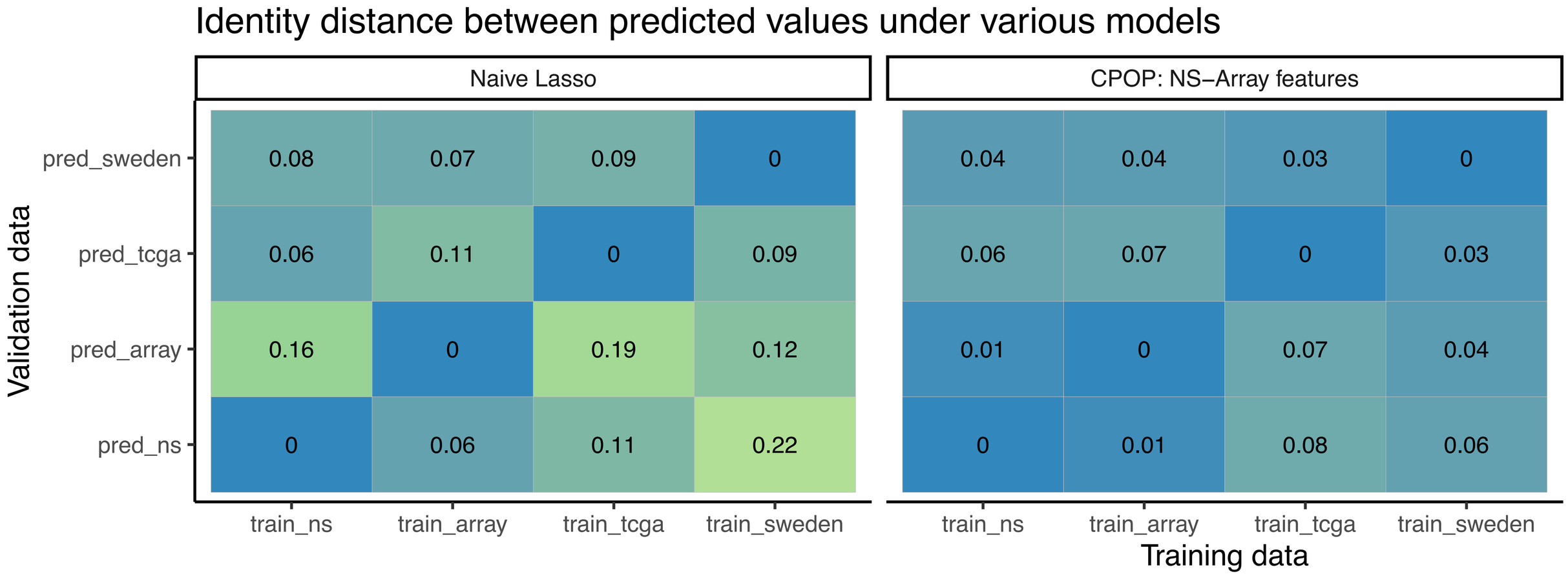
(feature distribution stability)

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

element-wise

(mode estimation stability)

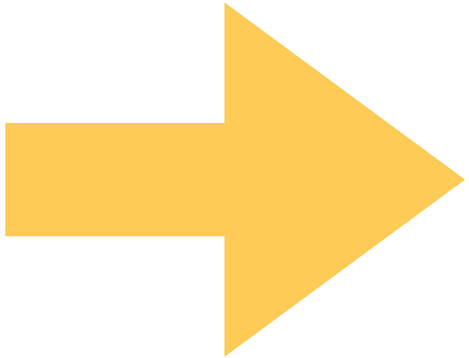
CPOP results 1: four melanoma data



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

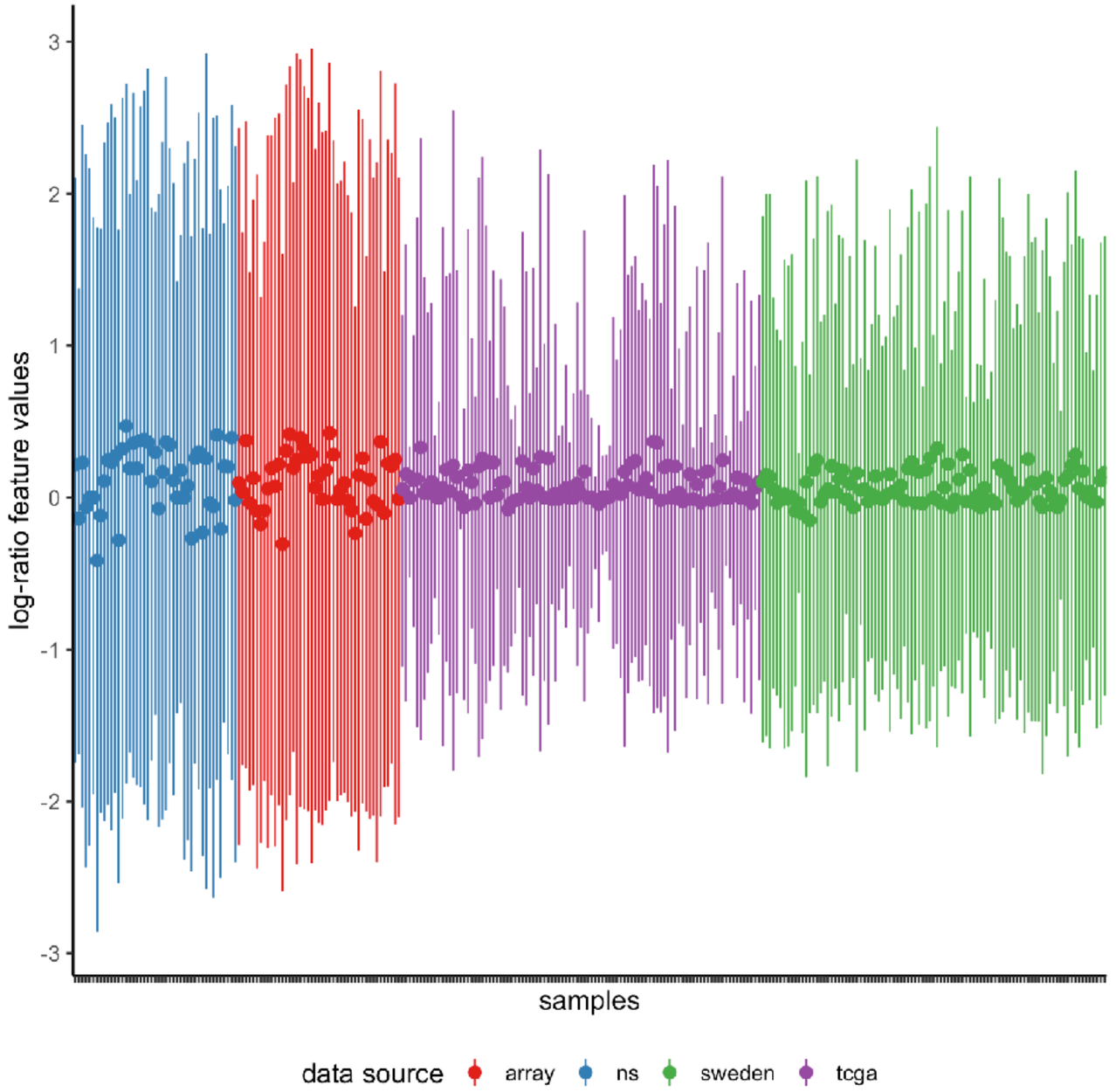
Is log-ratio really a new innovation?

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