Kevin Wang

School of Mathematics and Statistics, University of Sydney

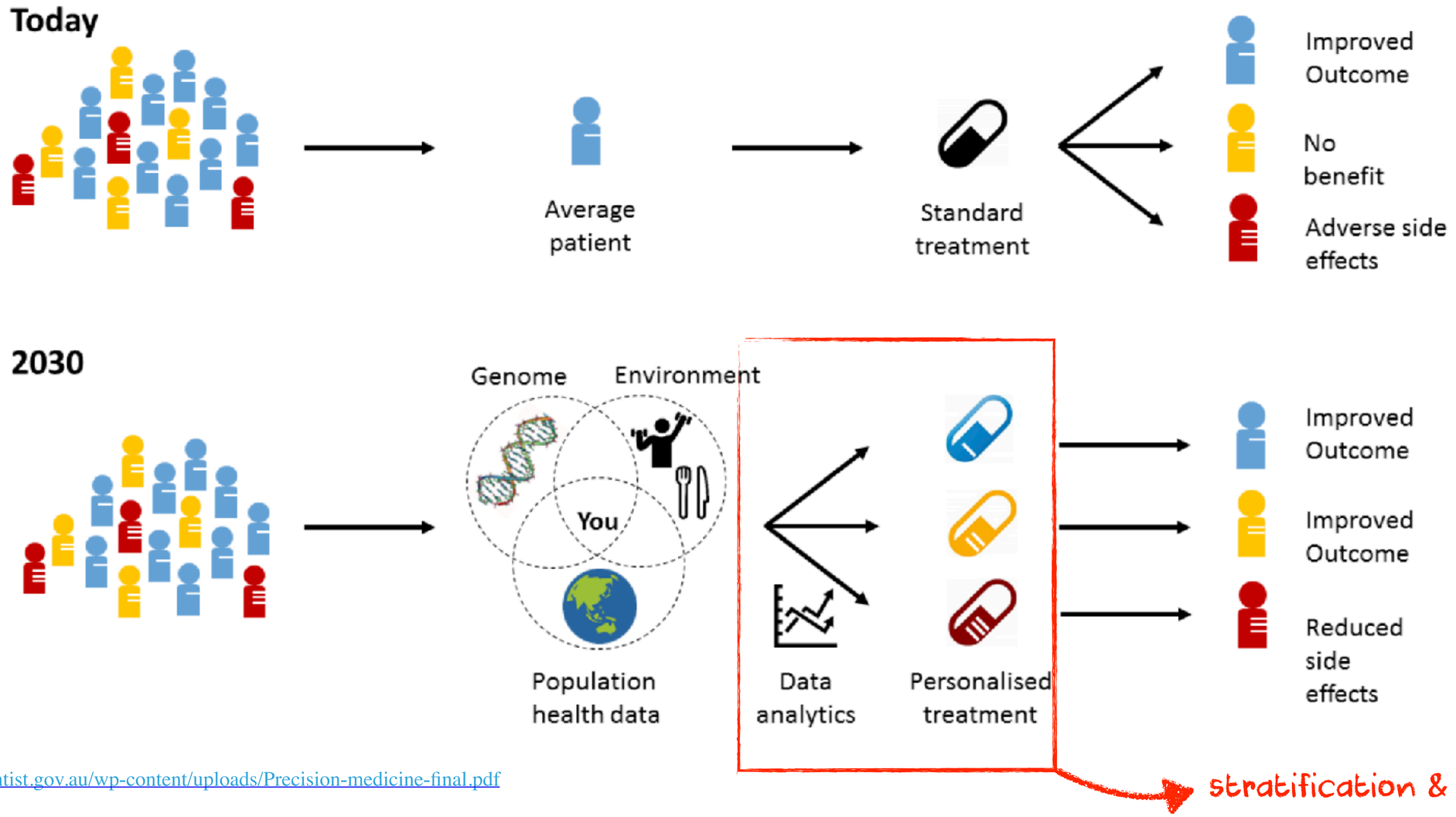
Cross-Platform Omics Prediction: towards implementation of precision medicine

13 Nov 2019

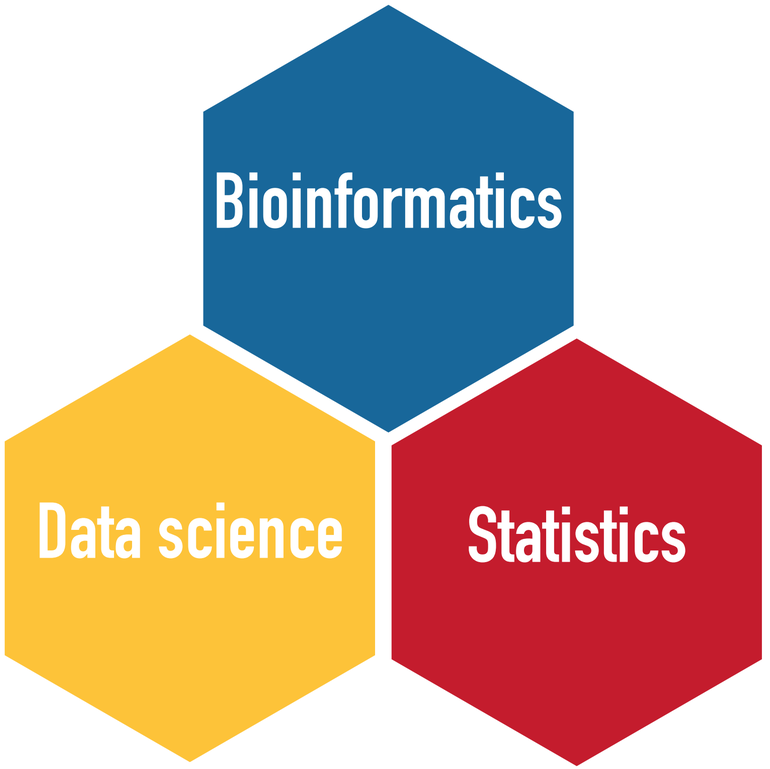
Precision medicine: predicting best cause of action using omics data

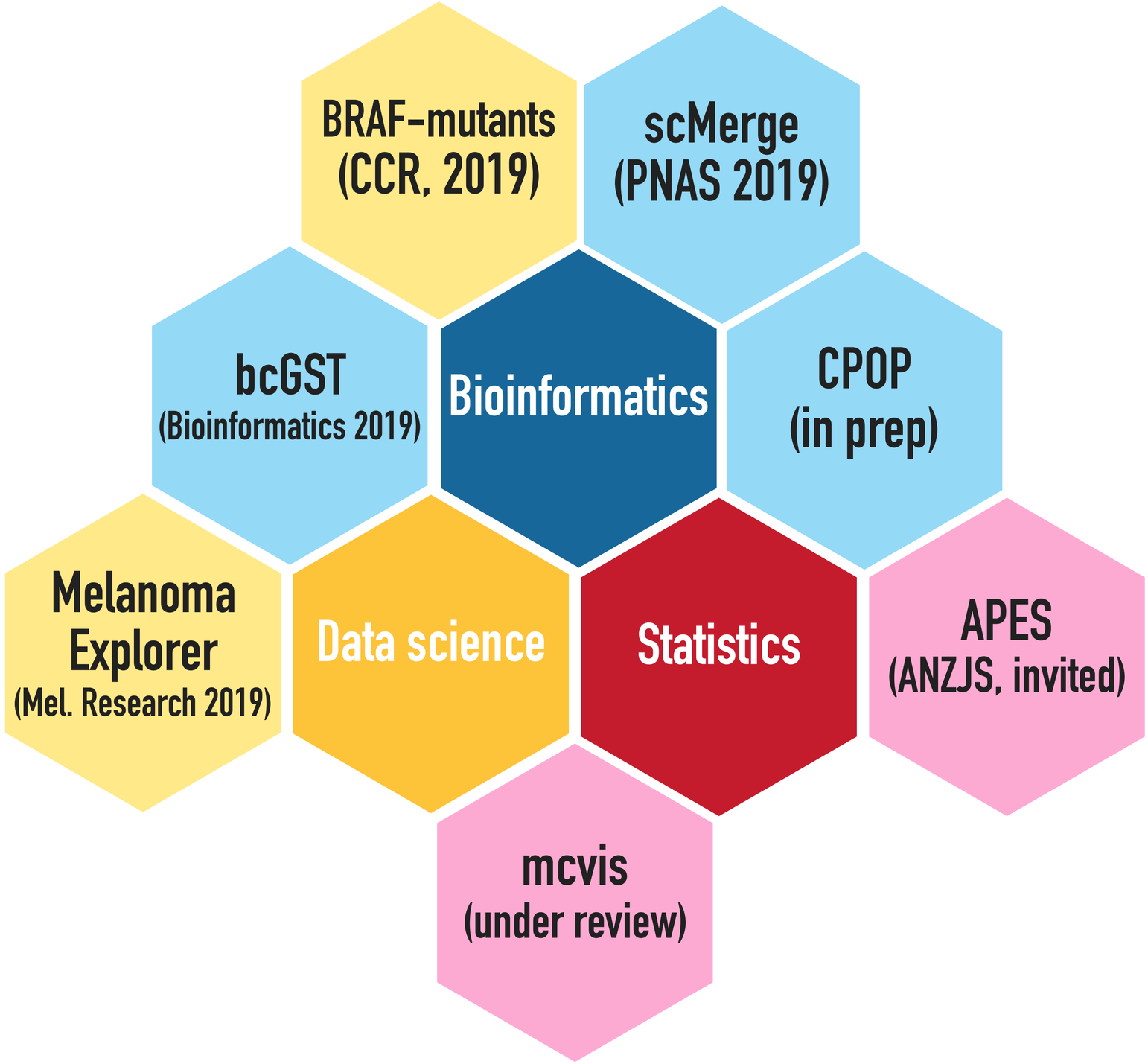
https://www.chiefscientist.gov.au/wp-content/uploads/Precision-medicine-final.pdf

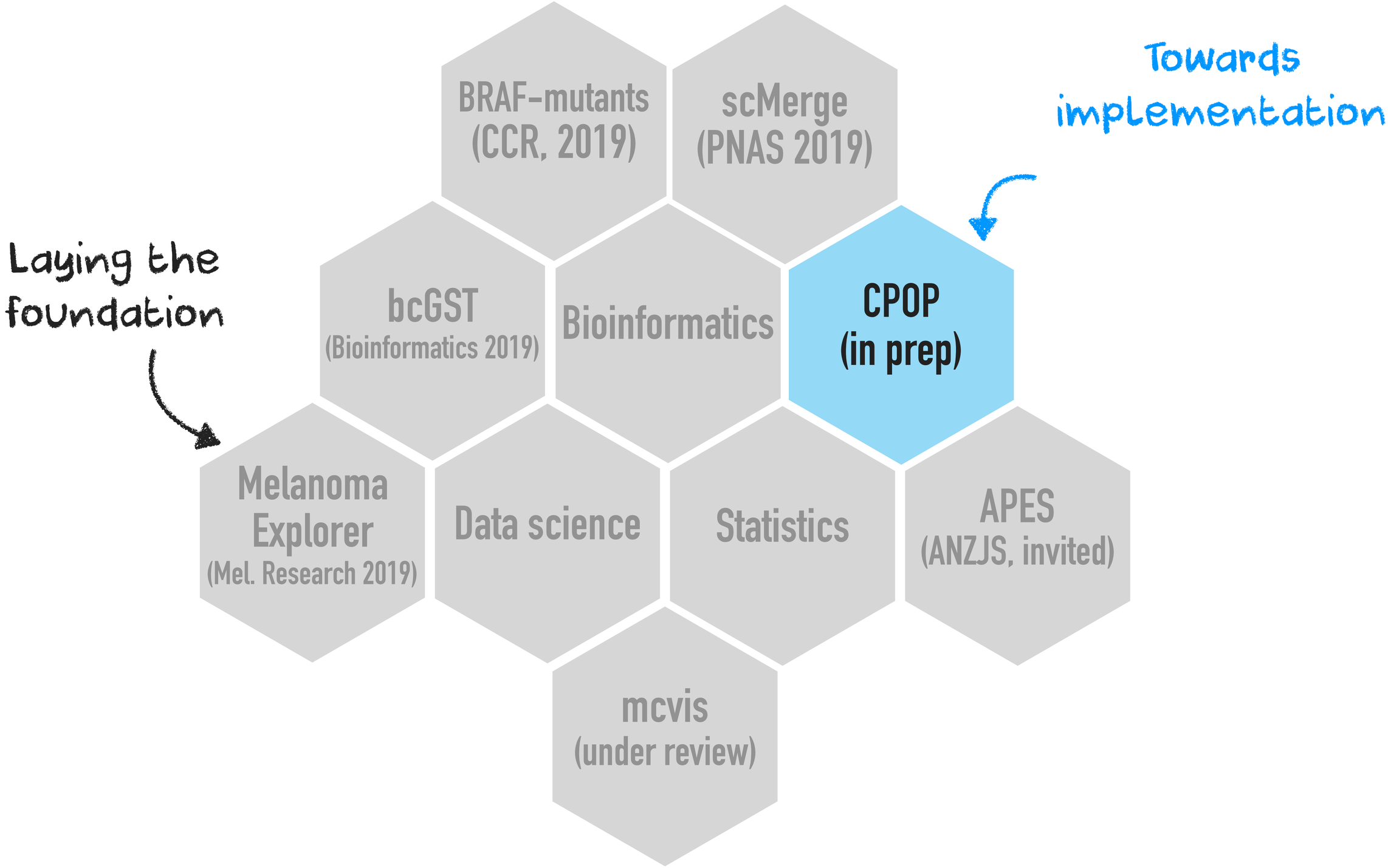
Precision medicine: predicting best cause of action using omics data



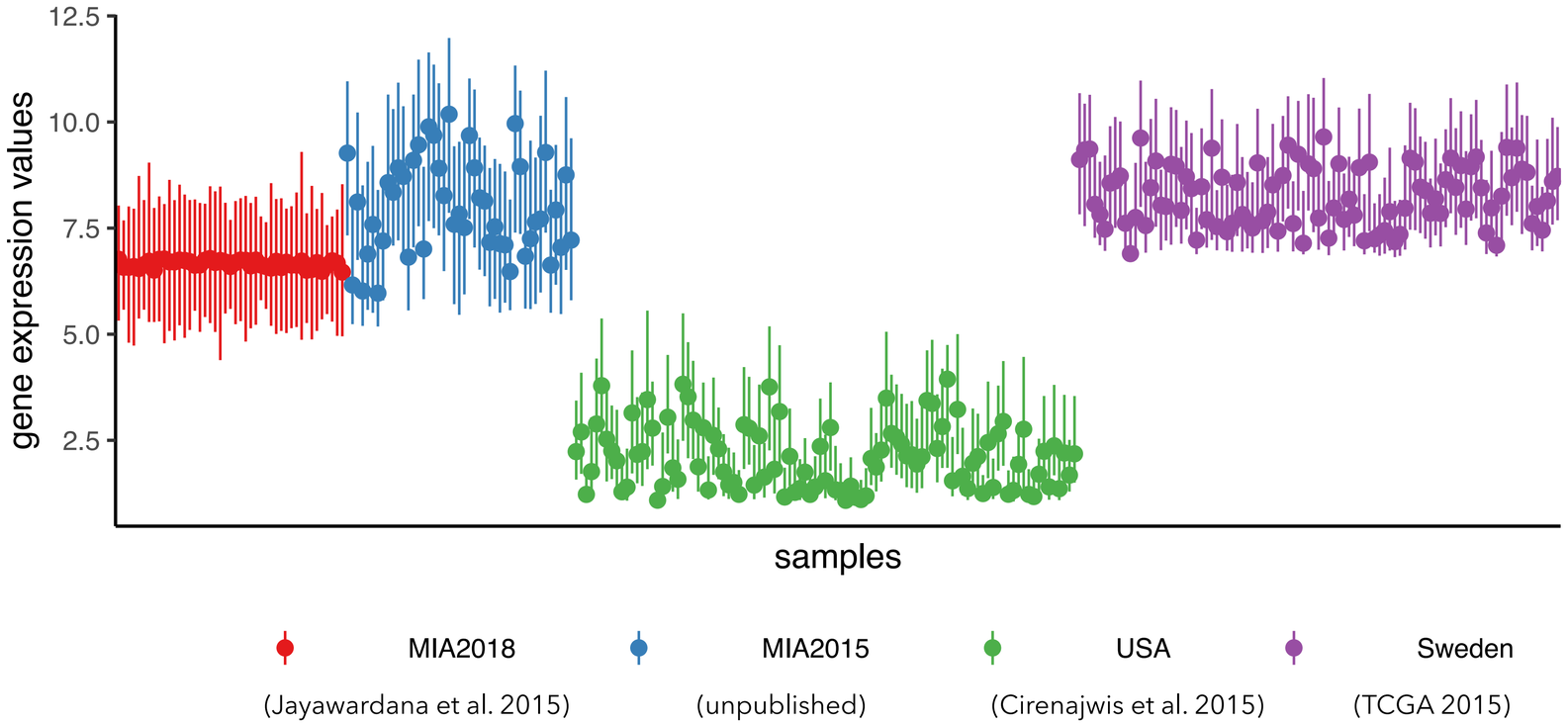
https://www.chiefscientist.gov.au/wp-content/uploads/Precision-medicine-final.pdf


Bioinformatics
Data science
Statistics


BRAF-mutants
(CCR, 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
BRAF-mutants (CCR, 2019)
scMerge (PNAS 2019)
Laying the foundation
bcGST (Bioinformatics 2019)
Bioinformatics
CPOP (in prep)
Melanoma Explorer (Mel. Research 2019)
Data science
Statistics
APES (ANZJS, invited)
mcvis (under review)

Melanoma Institute Australia



- Risk score using gene expression

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

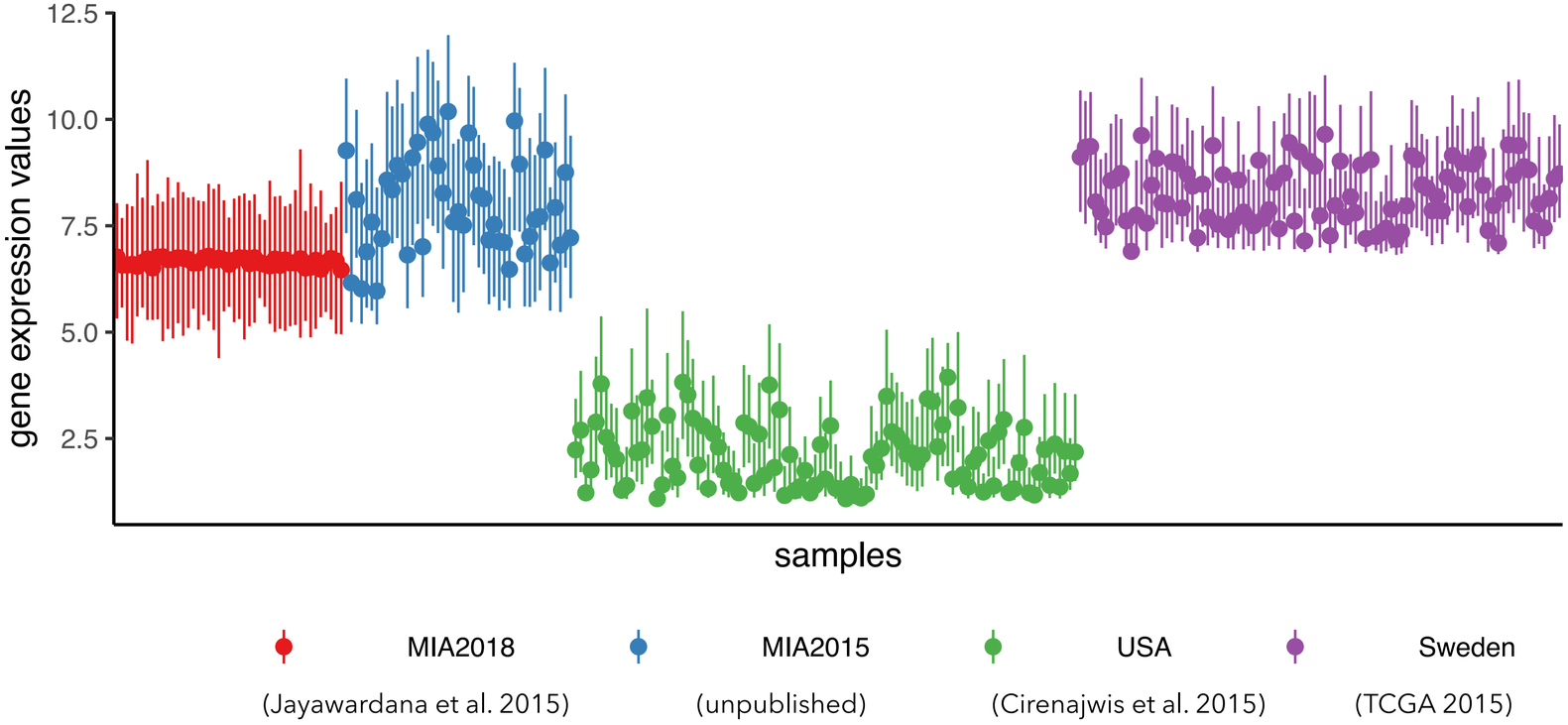
- CRE grant will support implementation
  - prospective
  - **multi-centres**

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?



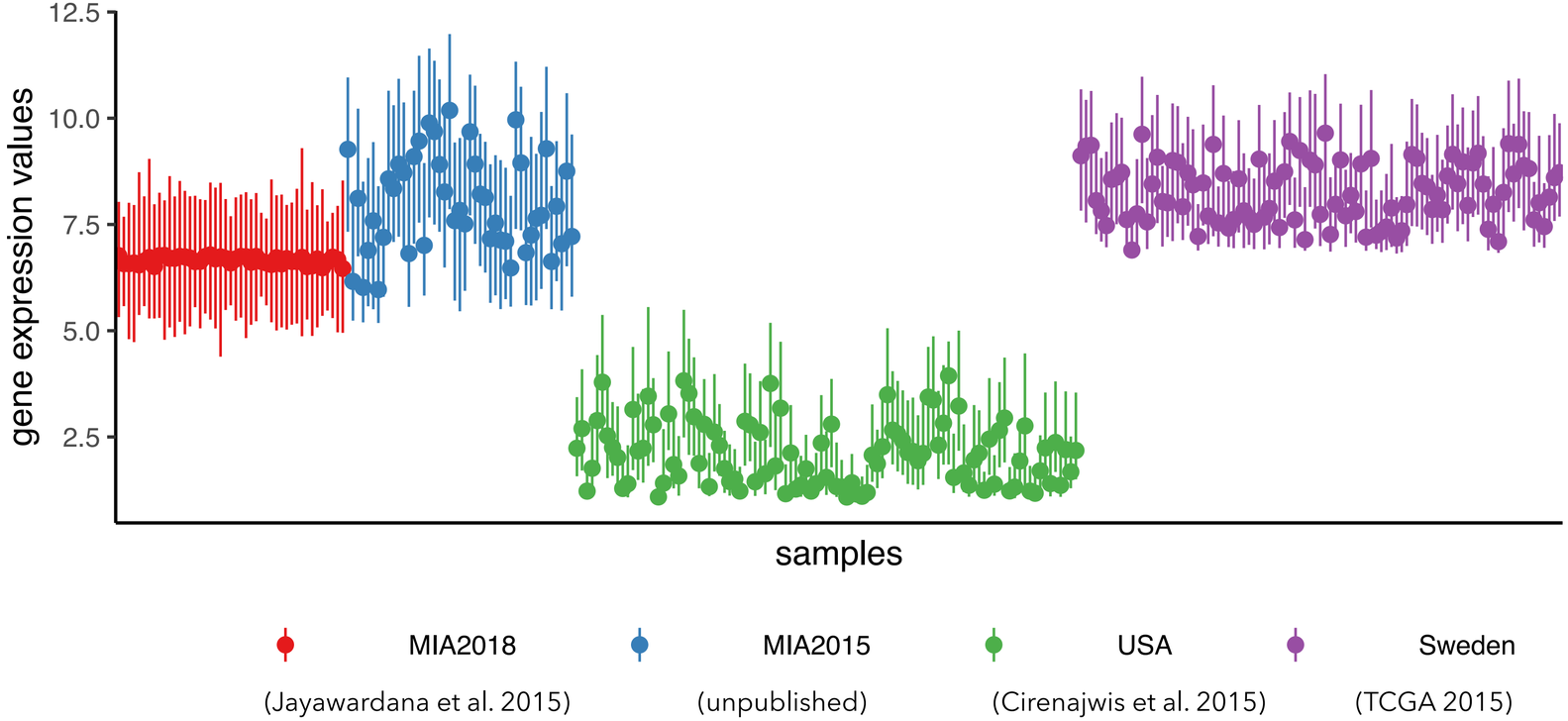
Gene expression features on a unit-less scale.

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.

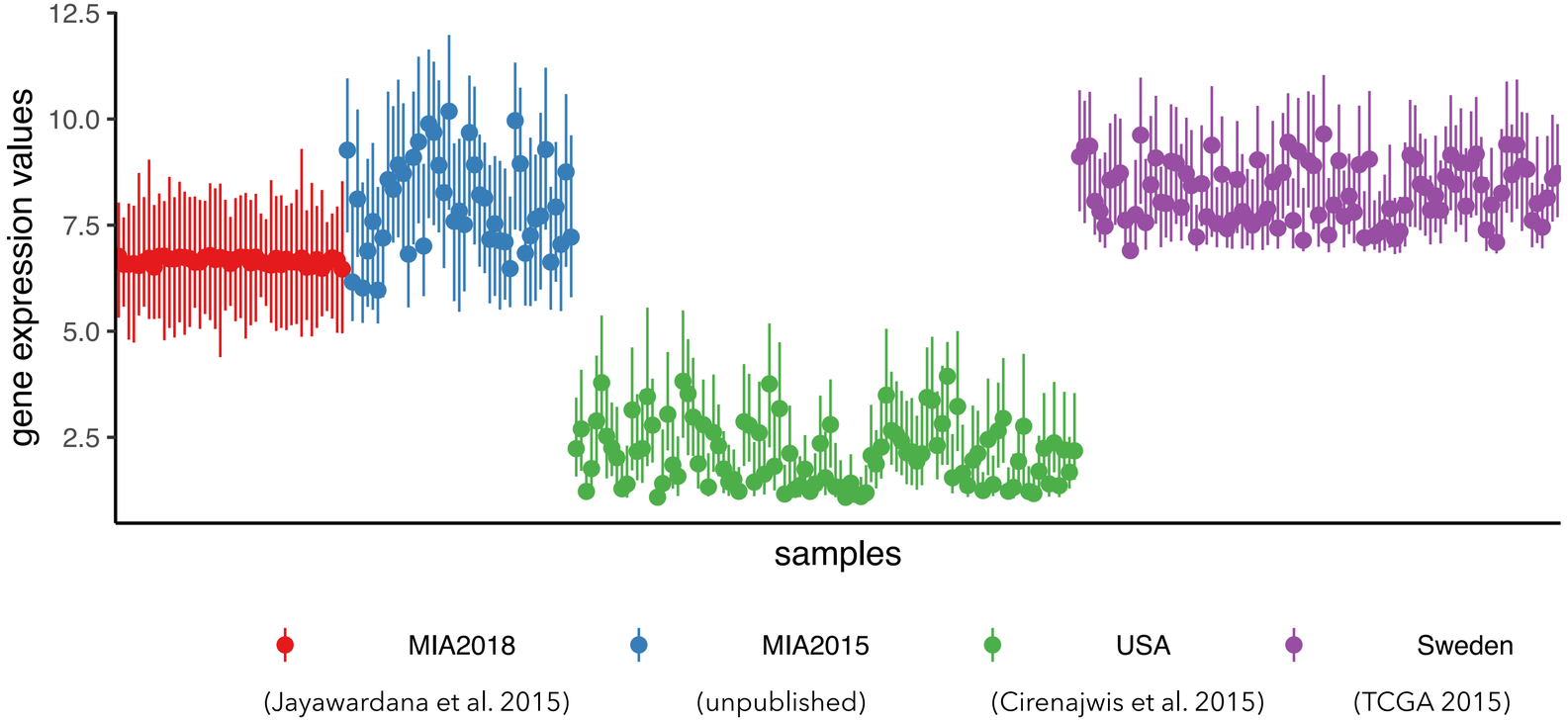
A typical value in one data can be an impossible value on another.

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

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

Assuming a noiseless shift

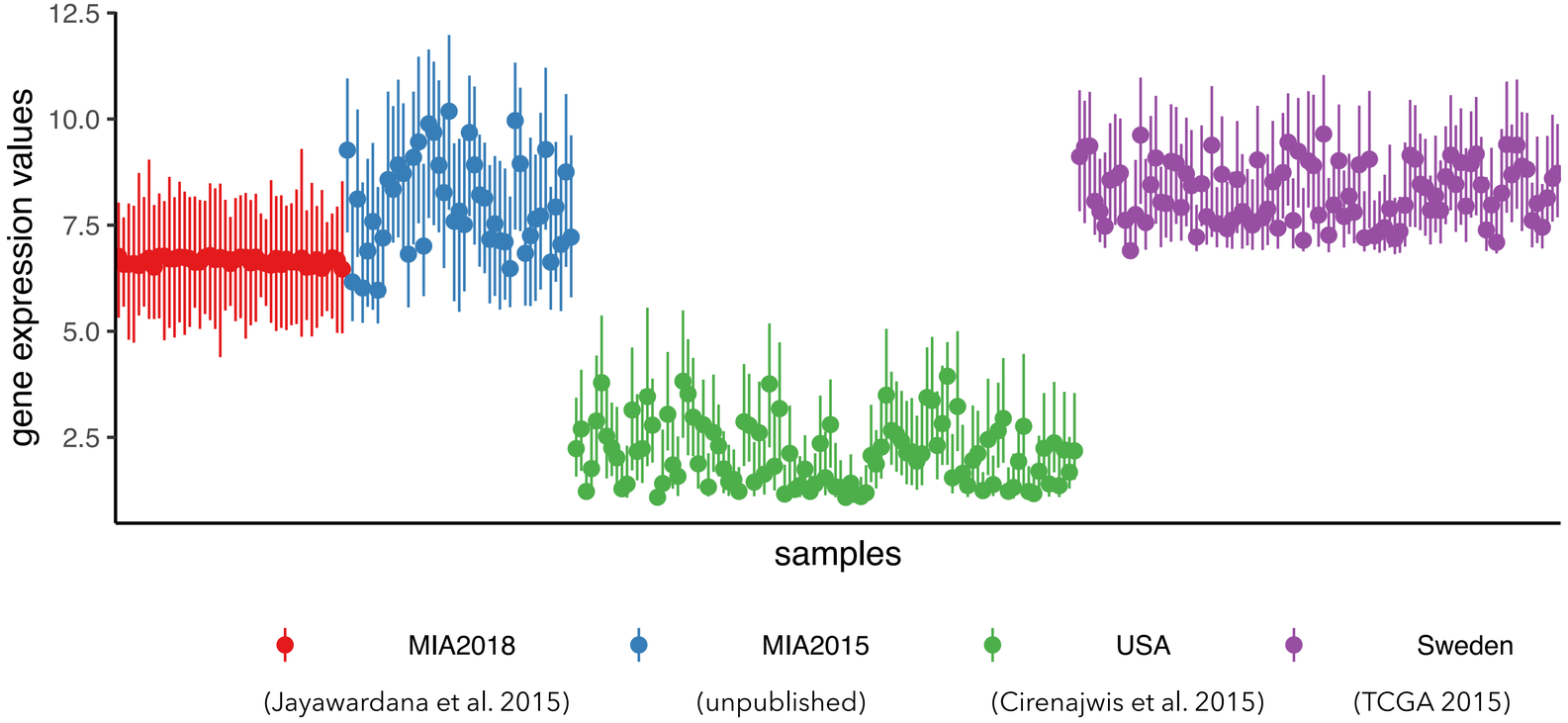
$$\hat{y}_2 = X_2 \hat{\beta}_1 = (X_1 + \mathbf{1}) \hat{\beta}_1$$



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 predictions**

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 harmonisation

No re-training

Scale-equivalent prediction

CPOP: prediction model that respects implementation 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

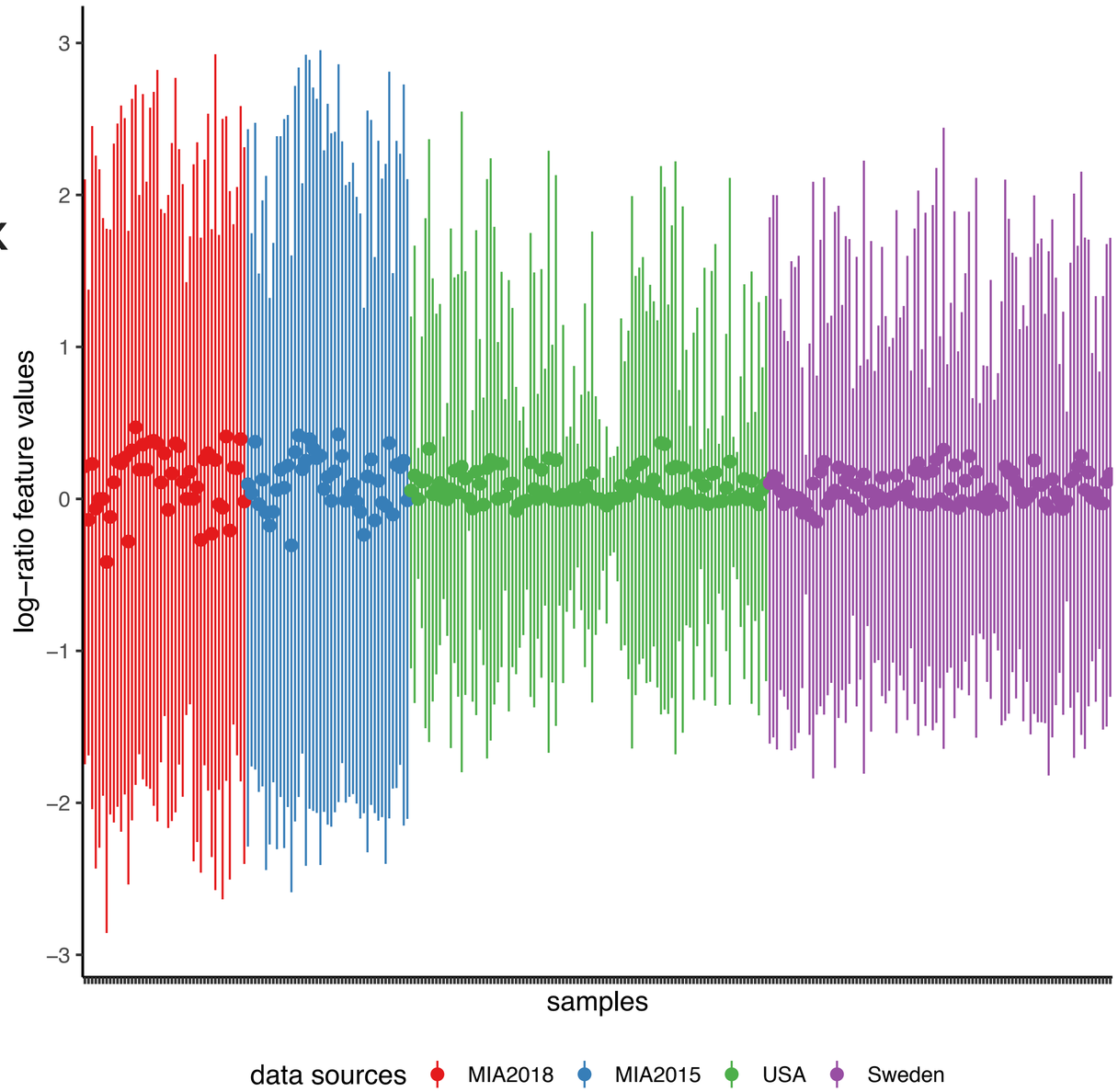
For each column in the gene expression matrix

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

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

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

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

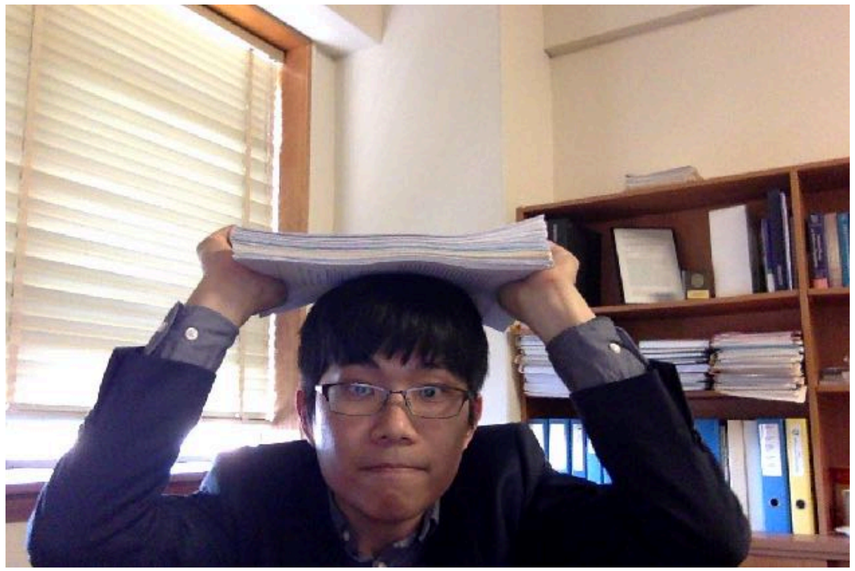


Why log-ratios?

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

Why log-ratios?

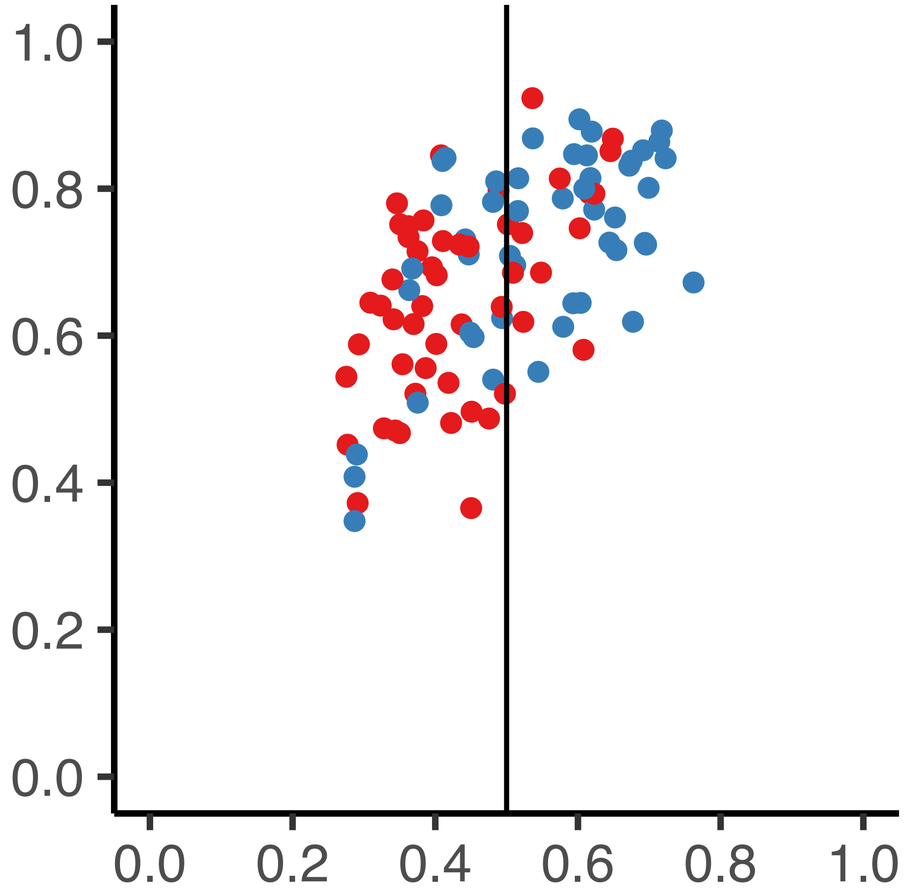
- Within-sample standardisation, conforms to implementation constraints
- Modelling on relative expression of genes
- Opportunities for method developments



Papers that used gene expression

The paper that used all log-ratios

Is log-ratio enough?

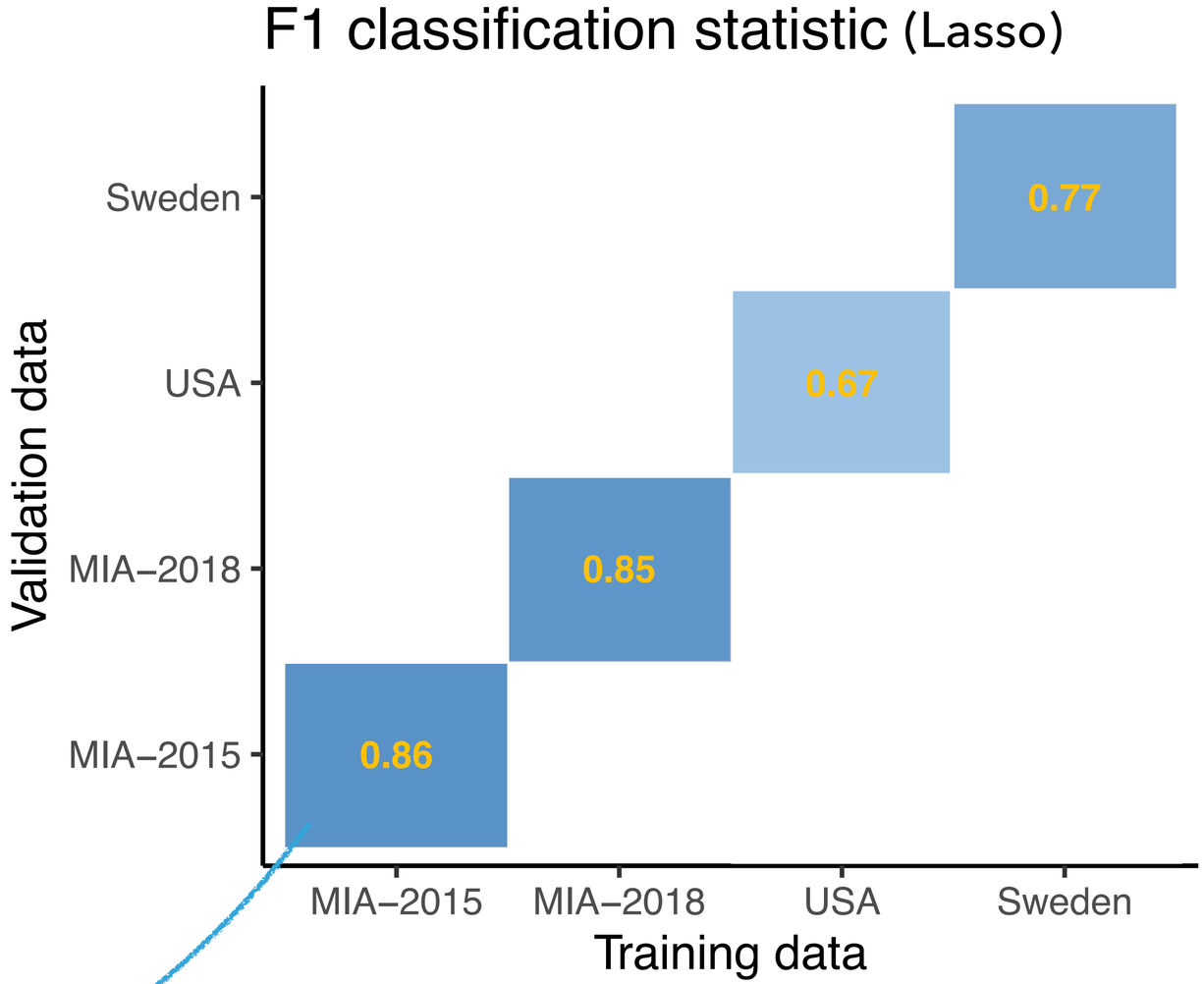


Sample class ● Good ● Poor

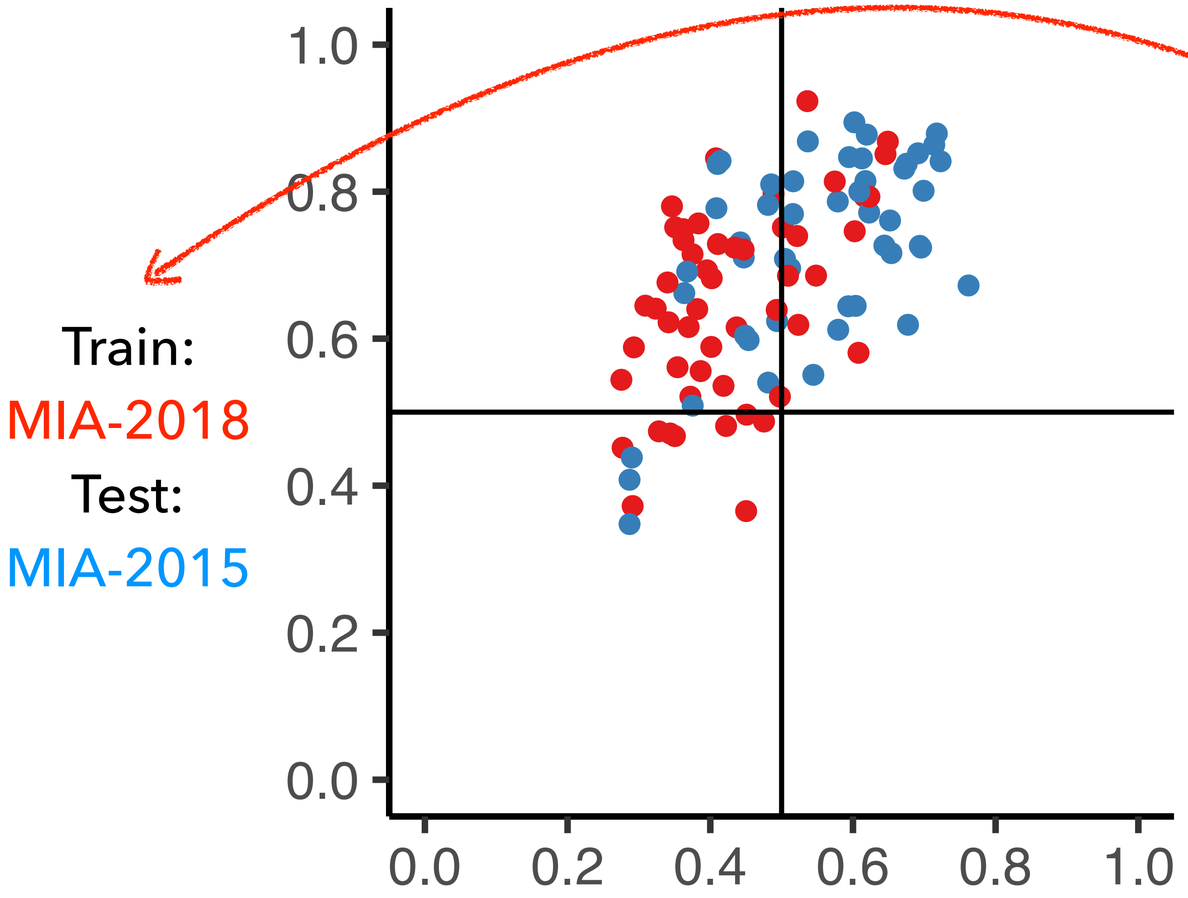
*Predicted values are shown as classification probability

Train: MIA-2015

Test: MIA-2015



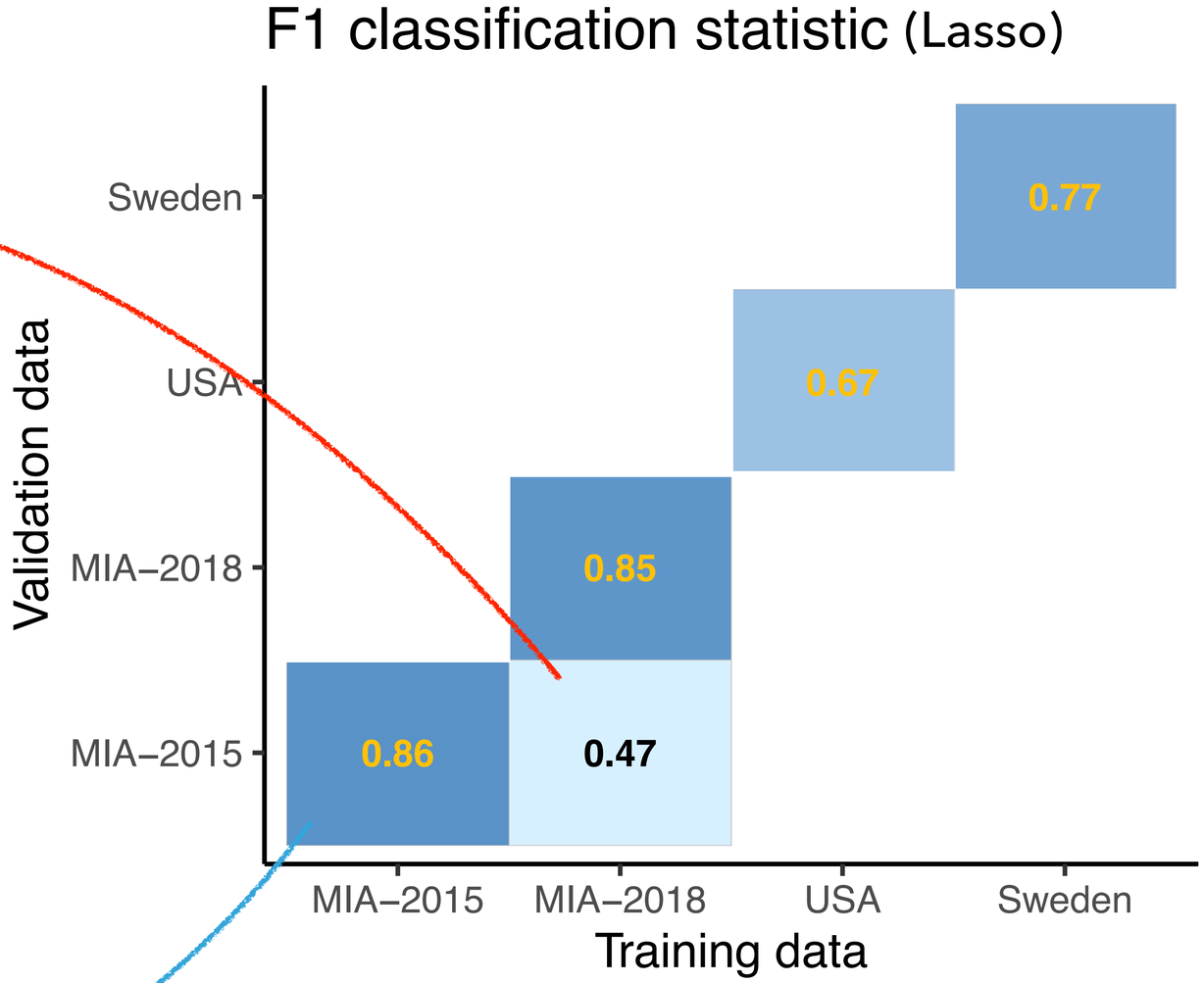
Is log-ratio enough?



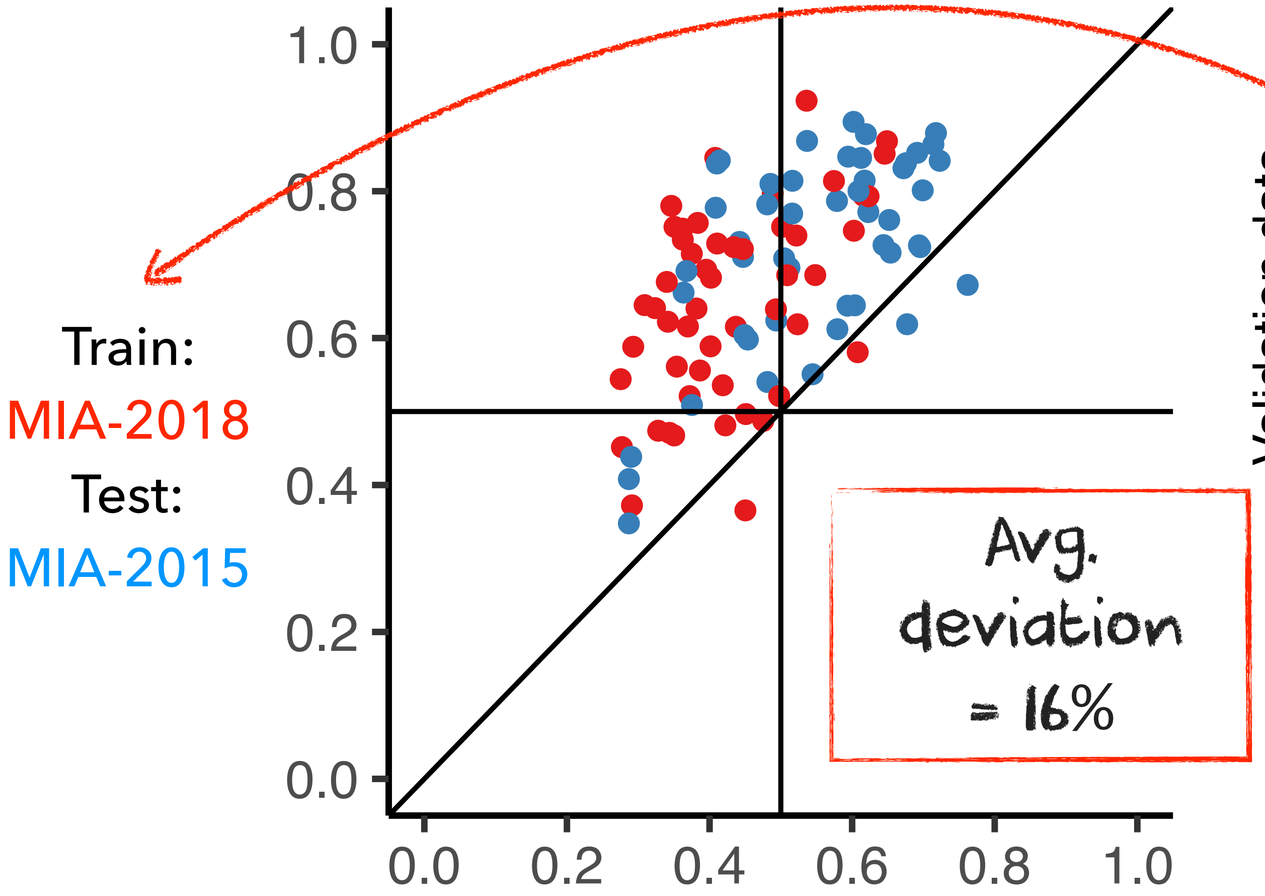
Train: MIA-2015
Test: MIA-2015

Sample class ● Good ● Poor

*Predicted values are shown as classification probability

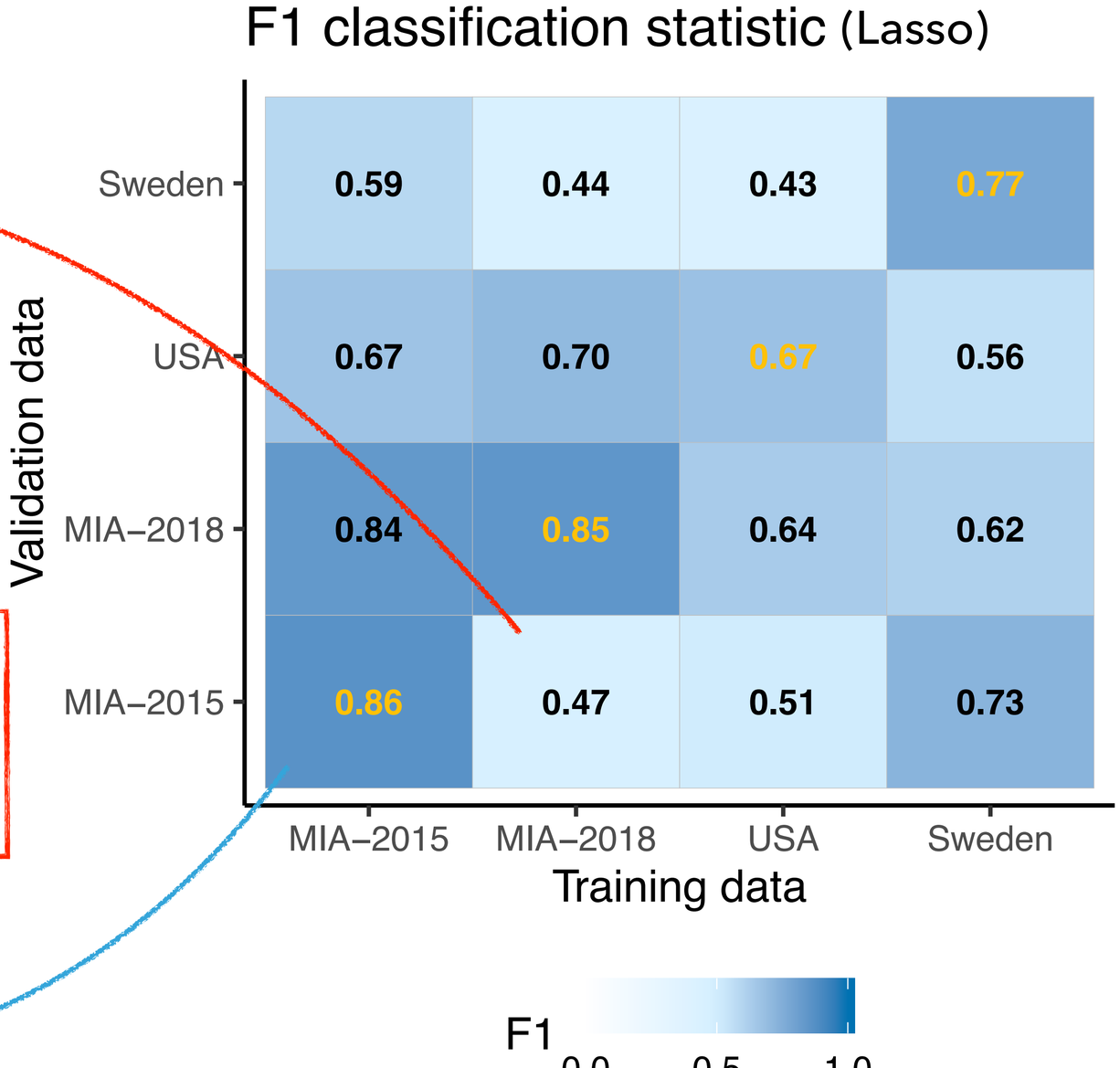


Is log-ratio enough?



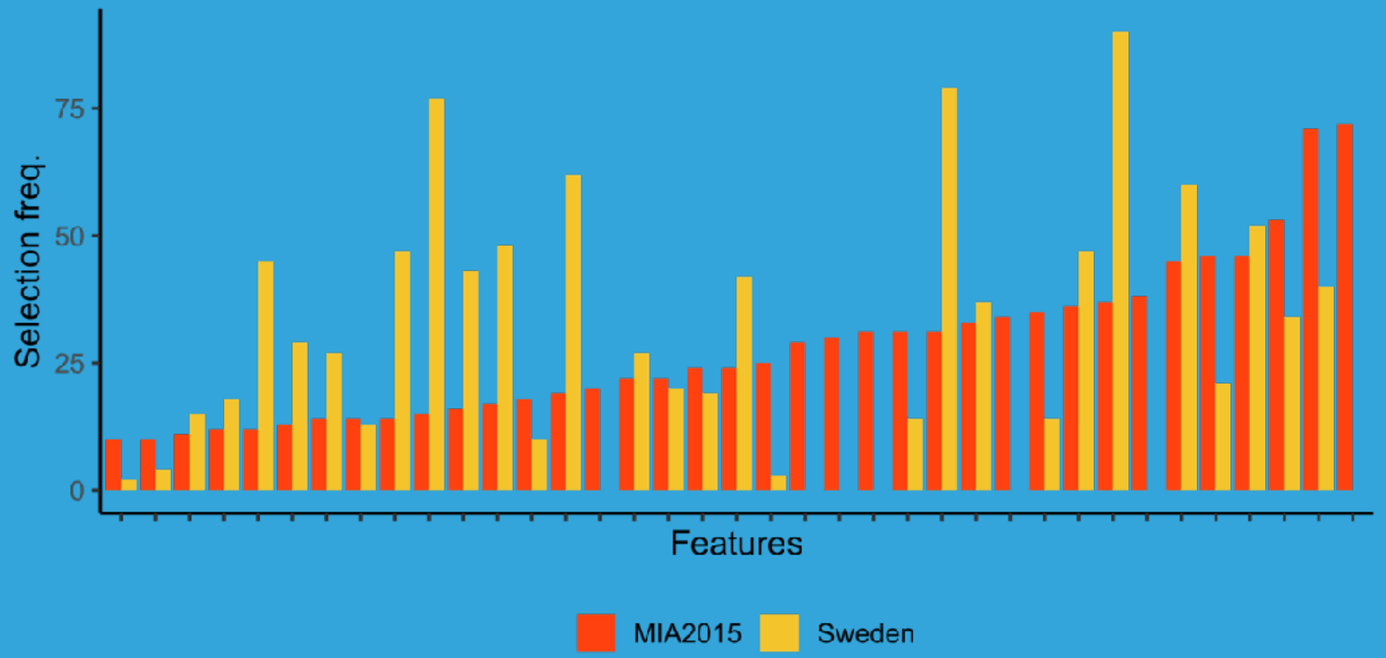
Sample class ● Good ● Poor

*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

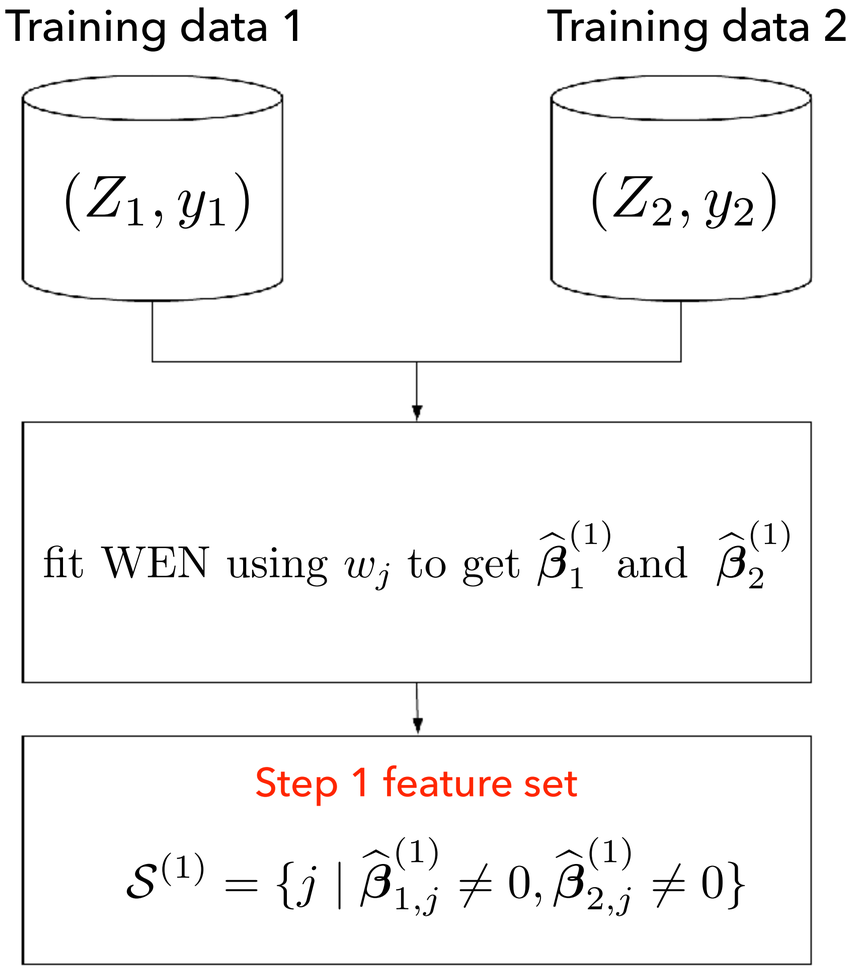
a mix of L1 and L2 penalties

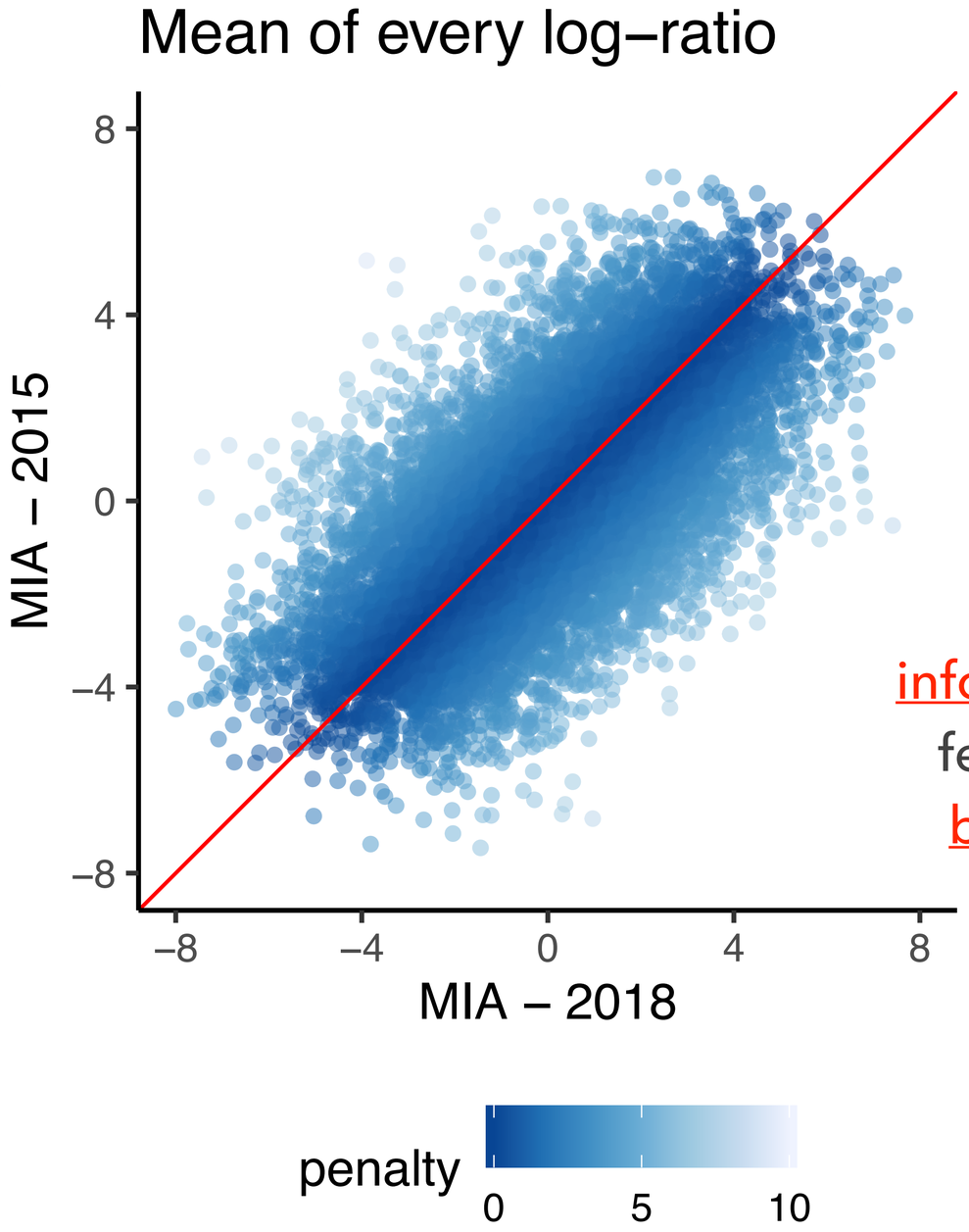
$$\widehat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{Z}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{\binom{p}{2}}}{\operatorname{argmin}} \sum_{i=1}^{n} l\left(y_i, \boldsymbol{z}_i^{\top} \boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{q} w_j \left[\frac{(1-\alpha)}{2} \beta_j^2 + \alpha |\beta_j|\right]$$

modelling on log-ratio features

penalties on each feature, proportional to the stability of features

Step 1: feature selection stability





informative and stable features agreed by both training data

Step 2: feature estimation stability

Step 1 feature set

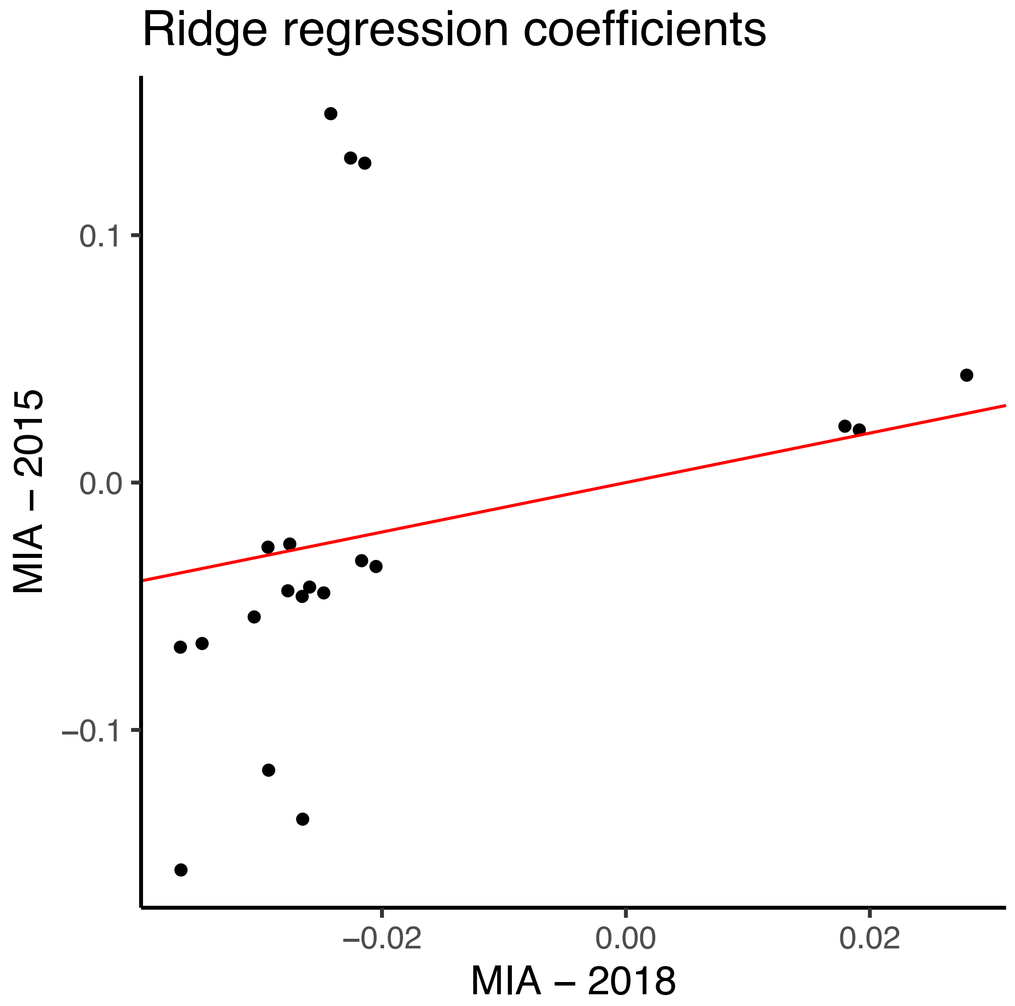
$$\mathcal{S}^{(1)} = \{j \mid \widehat{\boldsymbol{\beta}}^{(1)}_{1,j} \neq 0, \widehat{\boldsymbol{\beta}}^{(1)}_{2,j} \neq 0\}$$

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

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

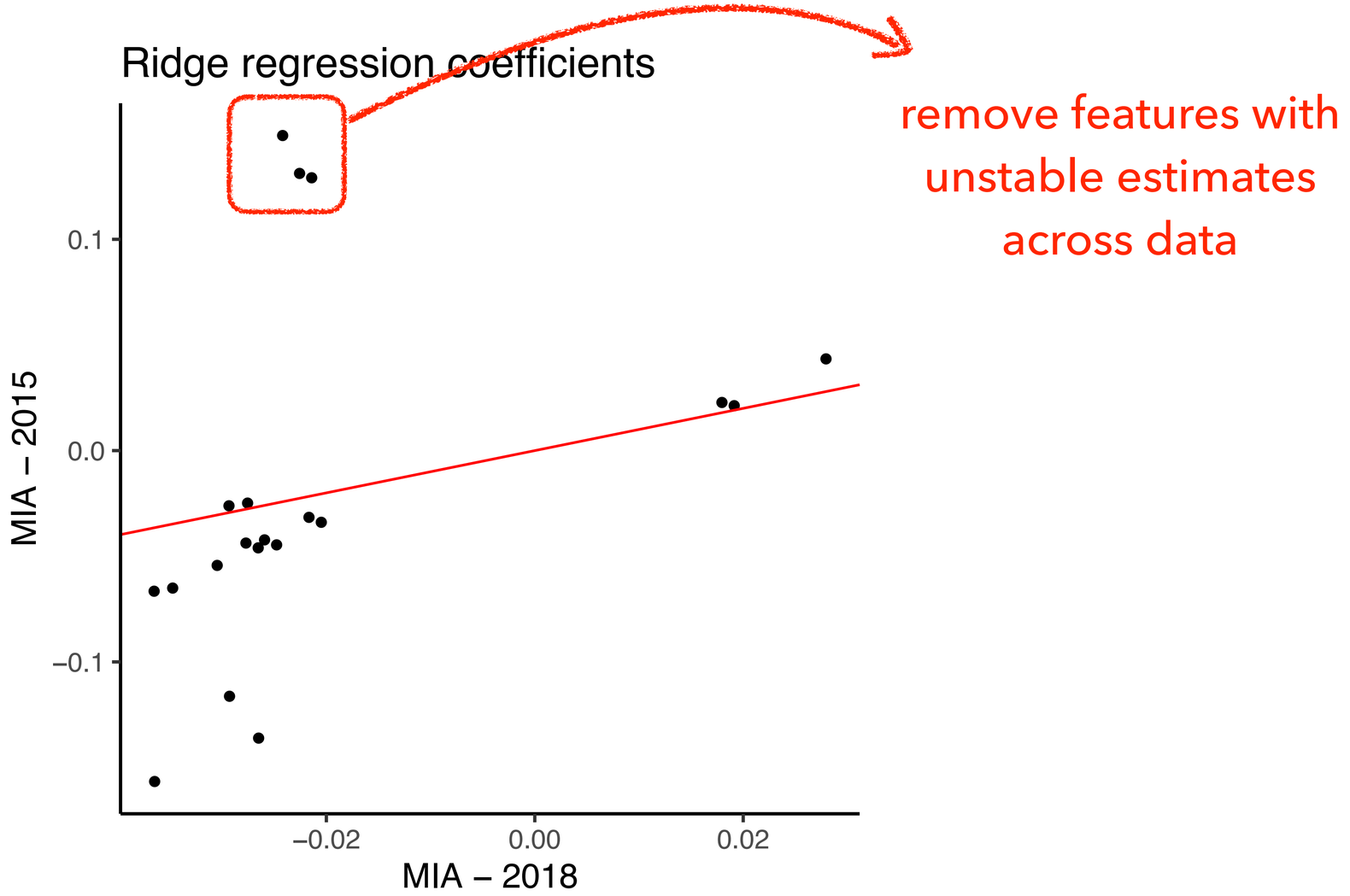
$$\mathcal{S}^{(1)} = \{j \mid \widehat{\boldsymbol{\beta}}^{(1)}_{1,j} \neq 0, \widehat{\boldsymbol{\beta}}^{(1)}_{2,j} \neq 0\}$$

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

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

Step 2 feature set

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



*Not all features are shown

remove features with unstable estimates across data

Step 2: feature estimation stability

Step 1 feature set

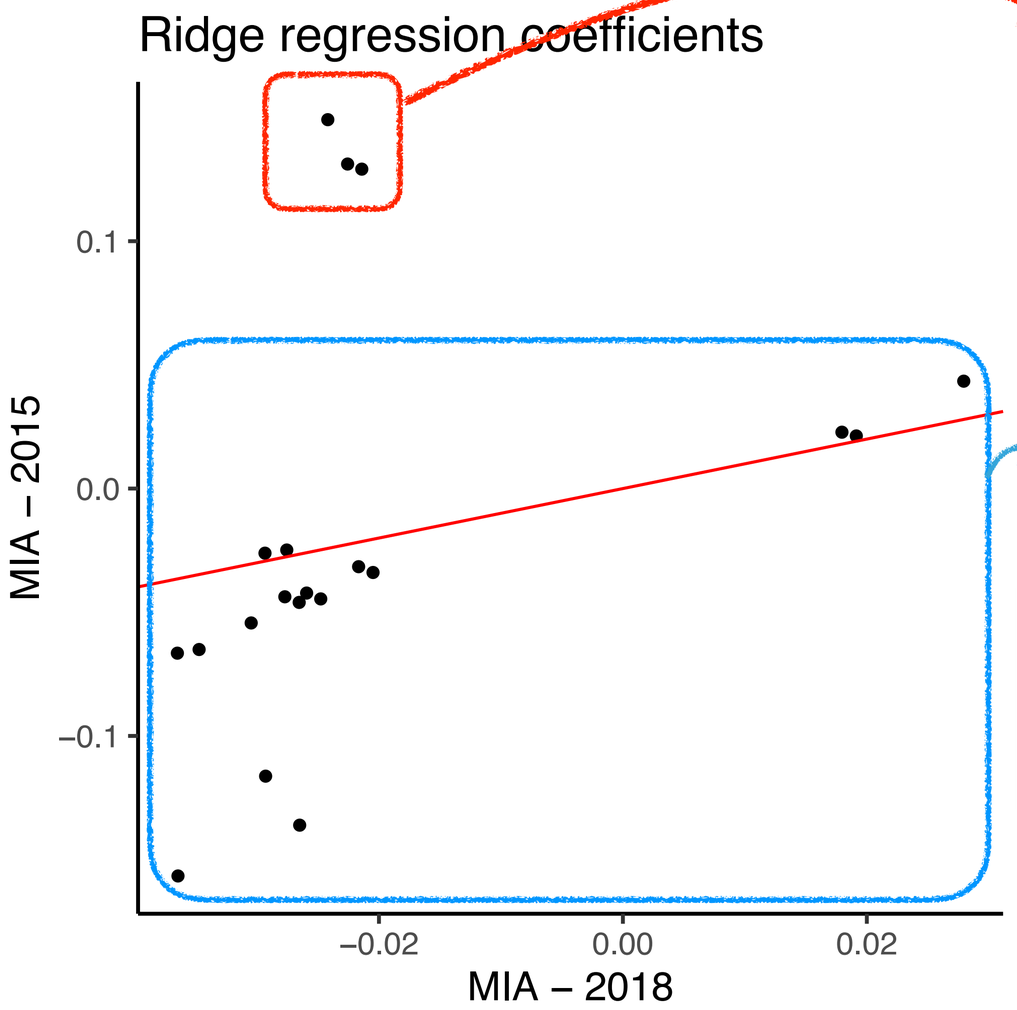
$$\mathcal{S}^{(1)} = \{j \mid \widehat{\boldsymbol{\beta}}^{(1)}_{1,j} \neq 0, \widehat{\boldsymbol{\beta}}^{(1)}_{2,j} \neq 0\}$$

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

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

remove features with unstable estimates across data

features suitable for stable prediction:

Final model

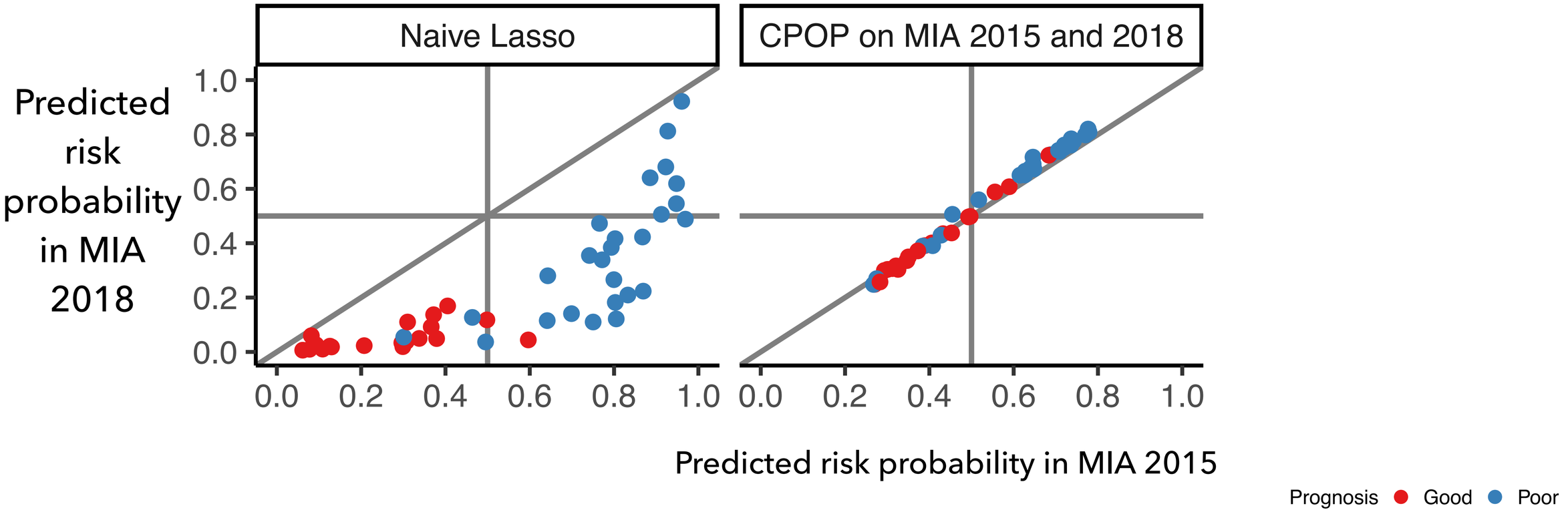
$\widehat{\boldsymbol{\beta}}_1$: fit Ridge on $(\boldsymbol{Z}_{1,\mathcal{S}^{(2)}}, \boldsymbol{y}_1)$

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

test on independent data

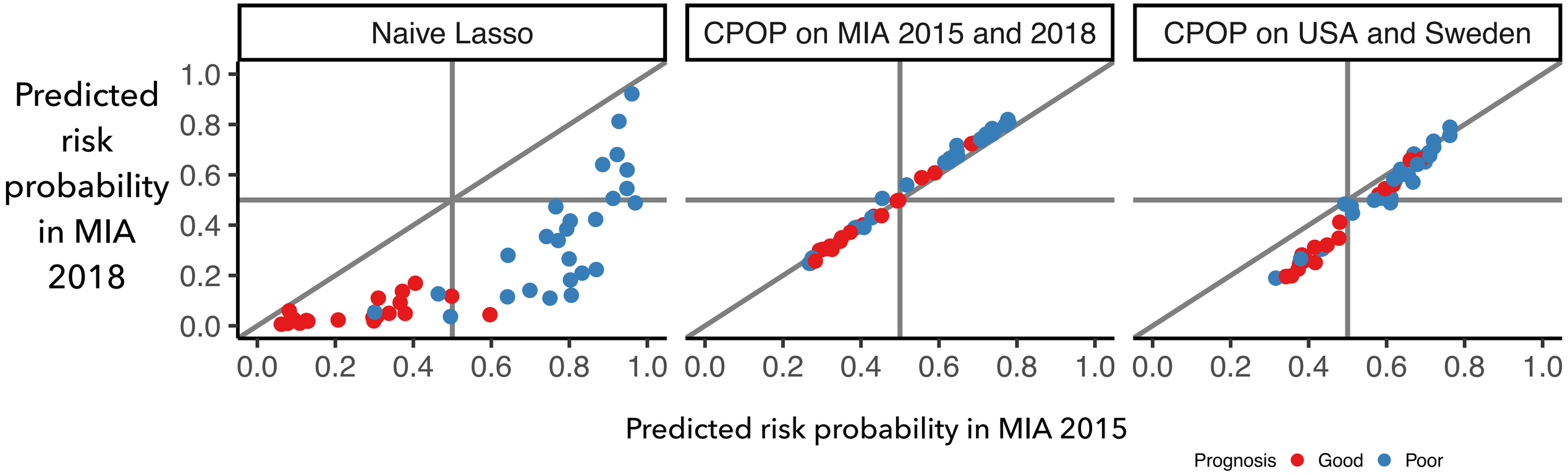
Results

Result 1: Predicted risk probability in MIA 2015 and 2018 data



Small deviation in predicted values across datasets

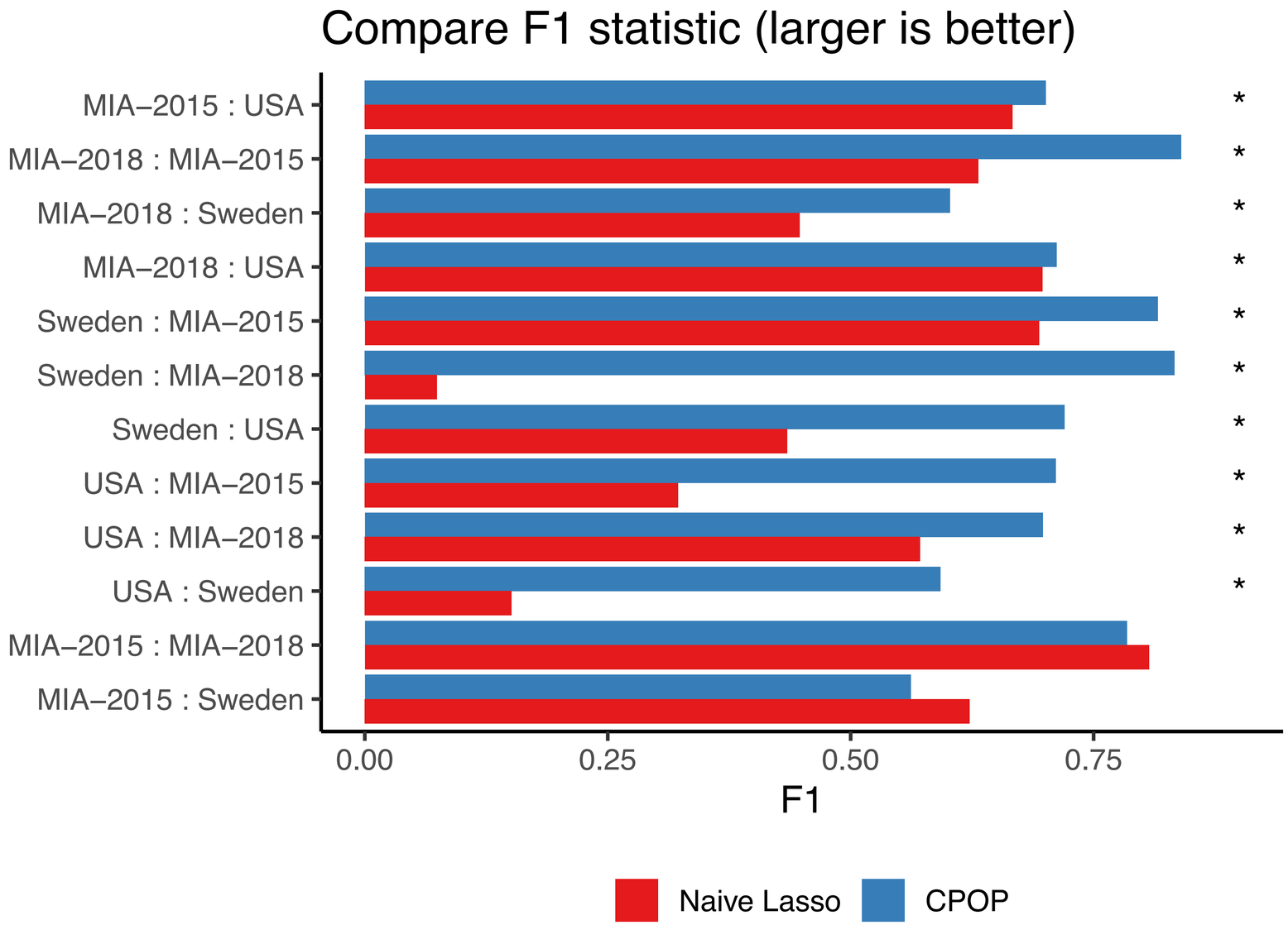
Result 1: Predicted risk probability in MIA 2015 and 2018 data



Small deviation in predicted values across datasets

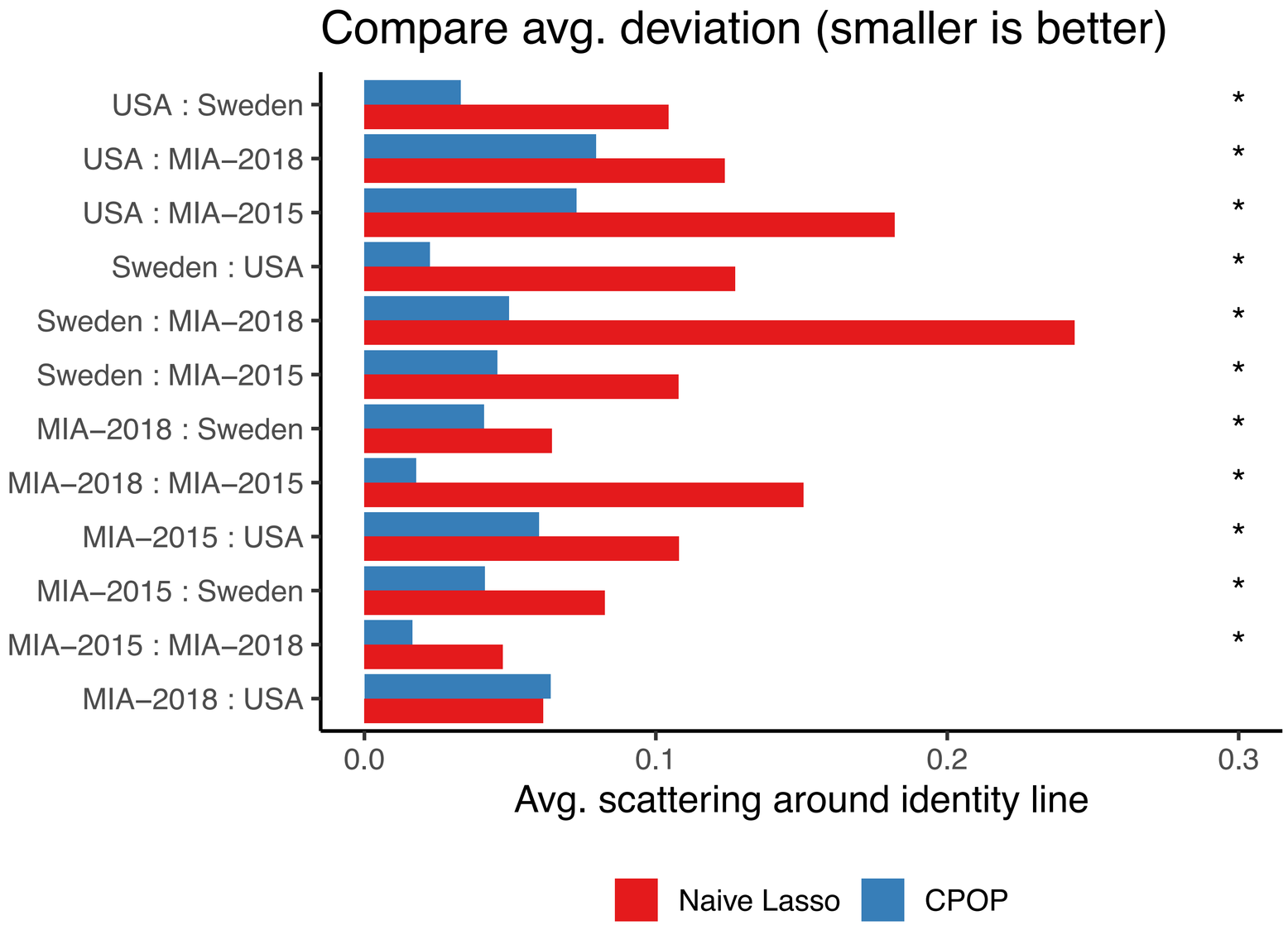
Result 2: four melanoma data

CPOP is highly predictive



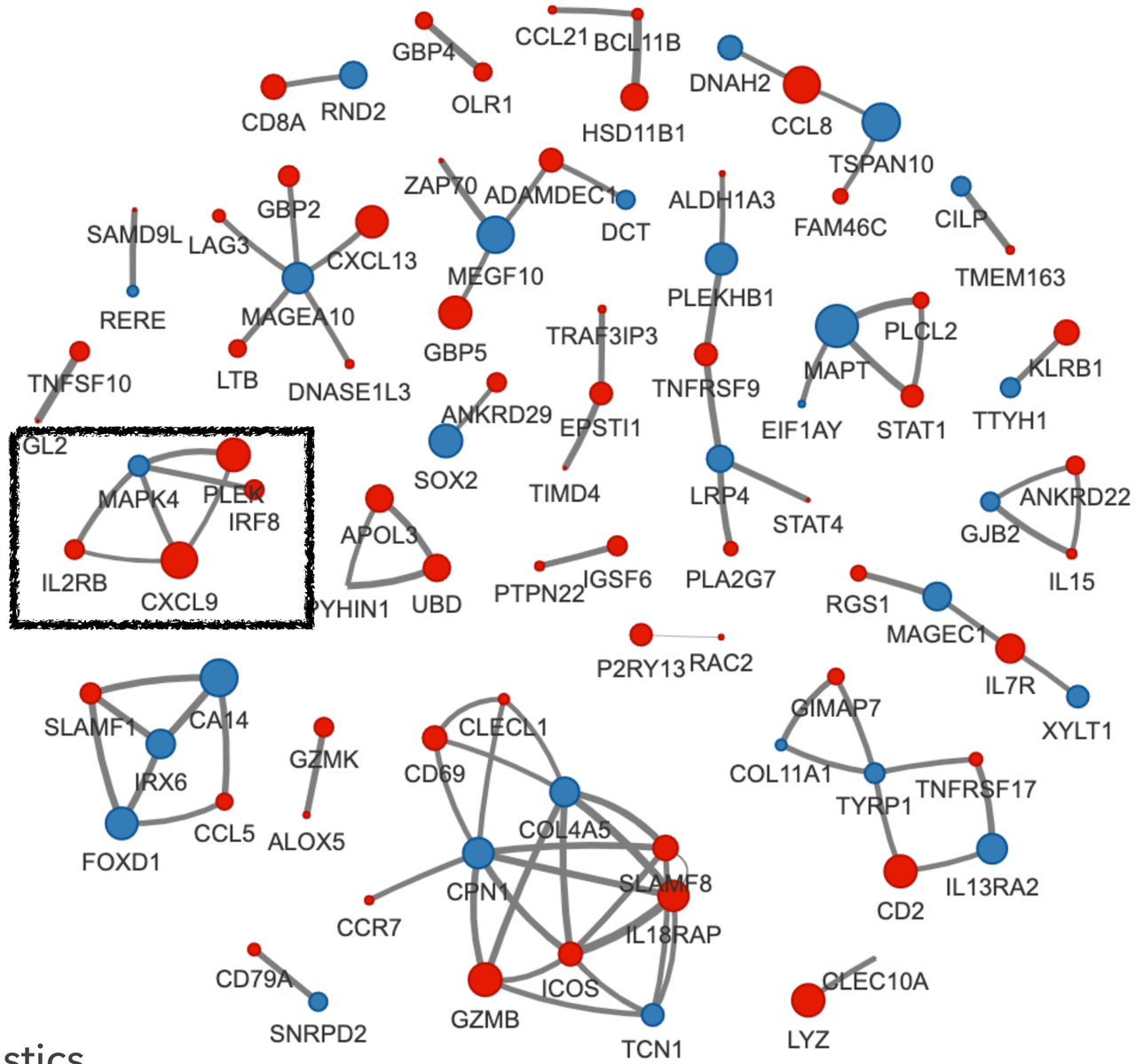
Result 2: four melanoma data

CPOP is highly stable



Result 3: melanoma feature set

CPOP features offer new biological interpretations



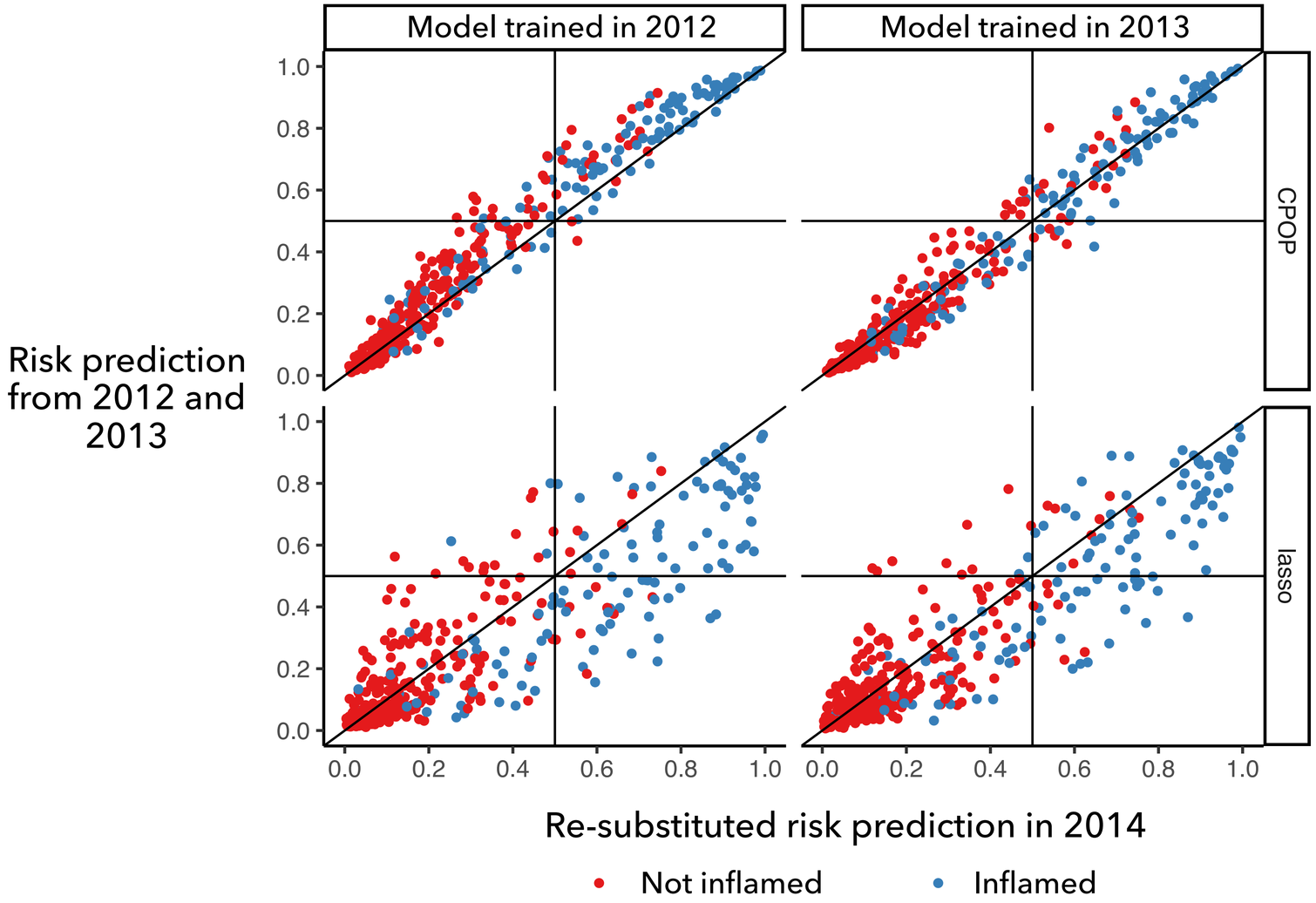
Edges: log-ratio features selected by CPOP

Vertices: genes comprising the log-ratio features

Colour and size: sign and magnitude of a gene's t-statistics

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

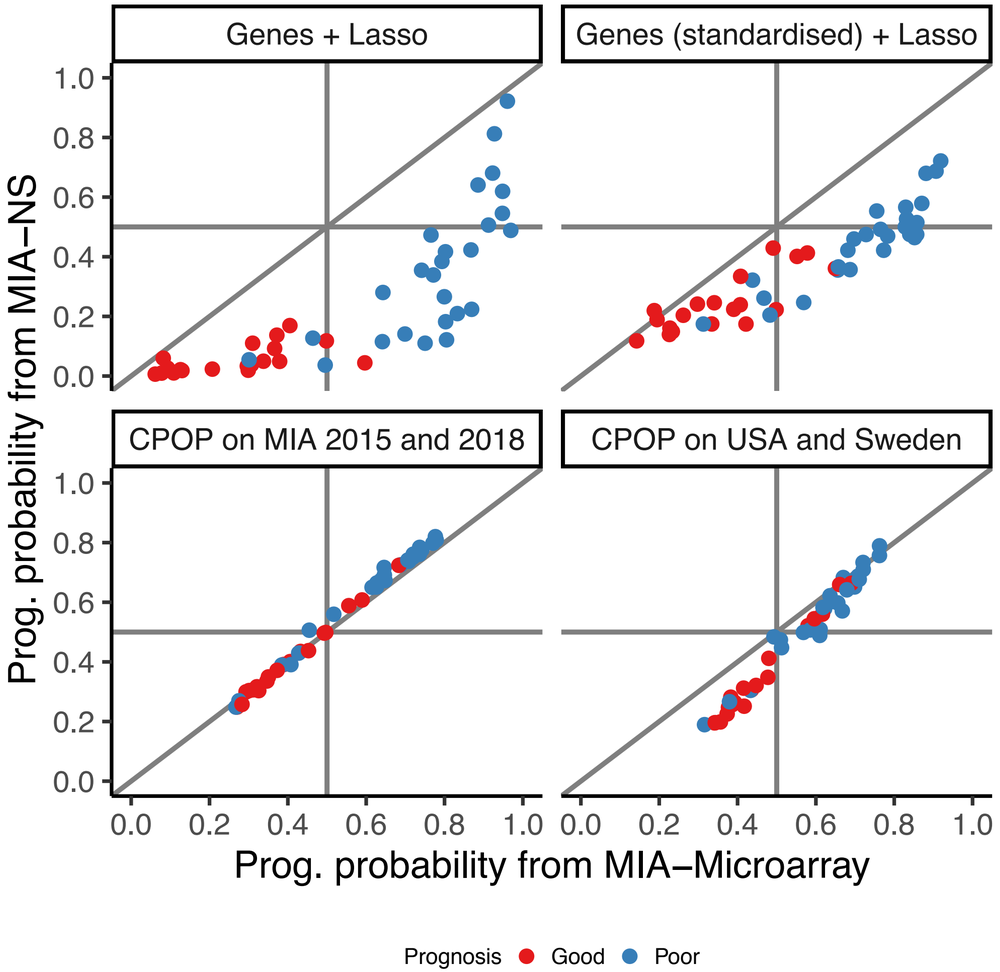
- A flexible and adaptable framework focuses on transferability
- Potential to work under a CRE grant for implementation

Acknowledgement

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

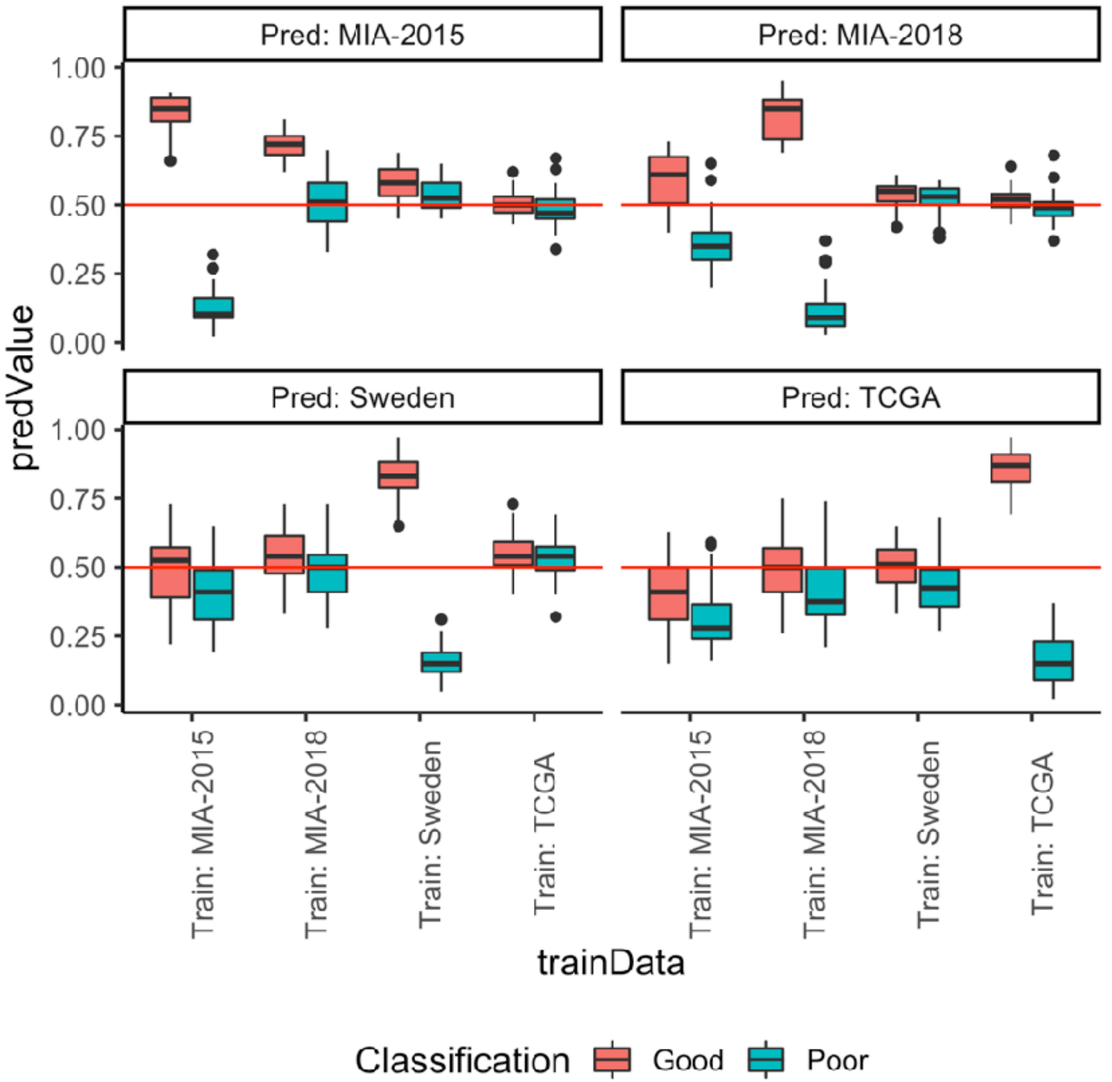
FAQ 1: why don't you just standardise or scale the data by a reference gene?

- Contribution of CPOP is not restricted to these log-ratio
- Normalisation by a "housekeeping" gene doesn't perform well empirically (left)
- Standardisation is not possible in prospective experiments
- Predicted outcome of a single sample depends on other samples that it just happens to be normalised with (McShane et al. 2013)



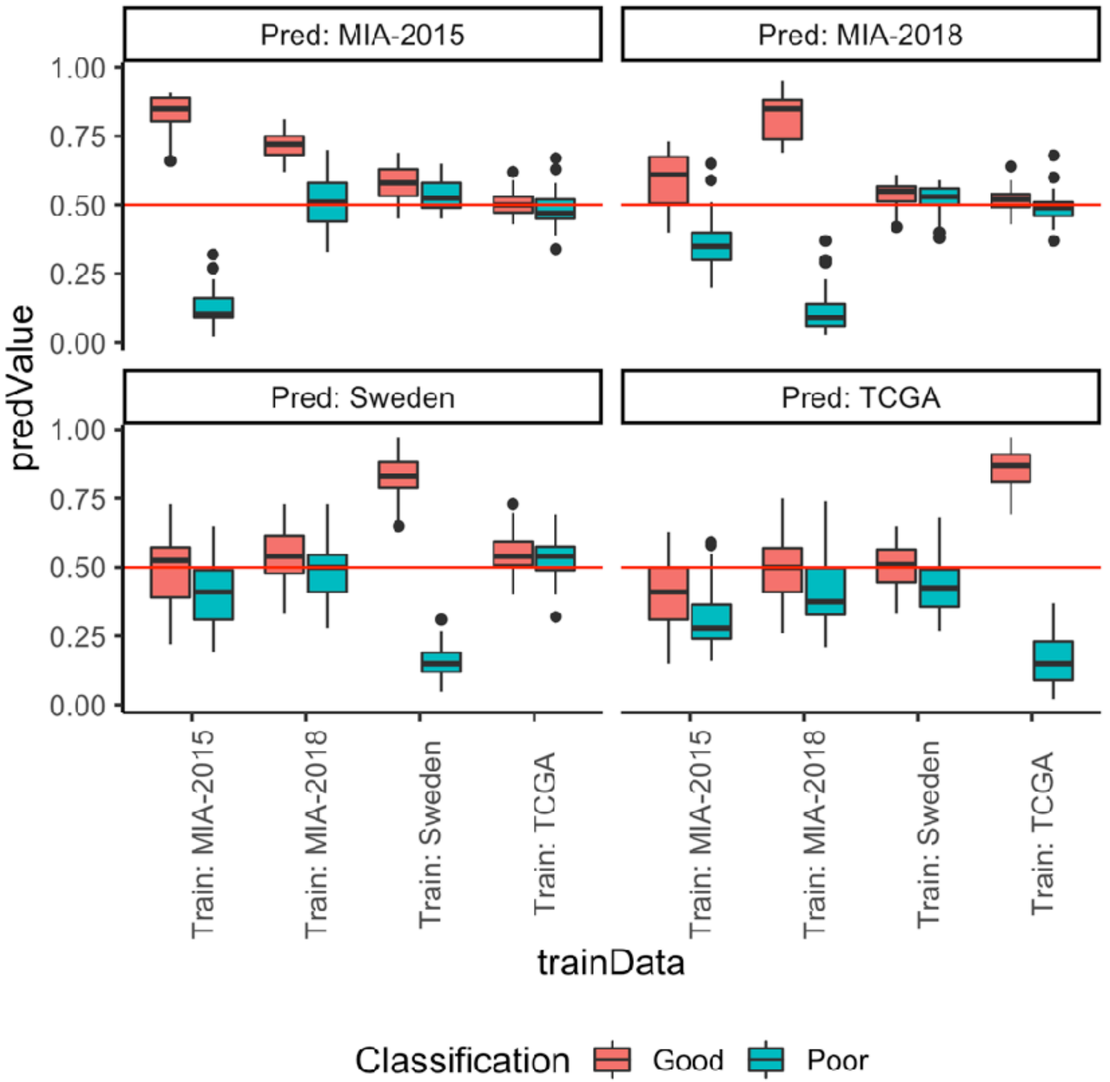
FAQ 2: have you tried [another machine learning method]?

- Yes, but the stabilisation techniques in CPOP is not restricted to regression
- The regression methodologies already offer better transferability results in comparison to, for example, random forest (left)
- Developing CPOP under the EN regression framework allows us to tap into a much richer collection statistical literature that focuses on stability



FAQ 2: have you tried [another machine learning method]?

- The regression model is usually more widely understood/accepted/ established in the medical community
- In terms of future work, we want to extend CPOP to drug sensitivity prediction, which has been established using regression models, e.g. (Garnett et al, Nature, 2012) and (Lee et al, Nature Comm., 2018)



FAQ 3: computational problems with log-ratios

- CPOP is not for biomarker discovery from thousands of genes. In biomedical research, prior to implementation, a gene set is restricted to about a few hundreds

- Variable pre-filtering could be applied before the construction of log-ratios (in the IBD data, 700 genes was converted to 250,000 features)

FAQ 4: platform differences

MIA - 2015: Illumina HumanWG-6 v3.0 expression beadchip, 45 samples

MIA - 2018: Customised NanoString probes, 45 samples

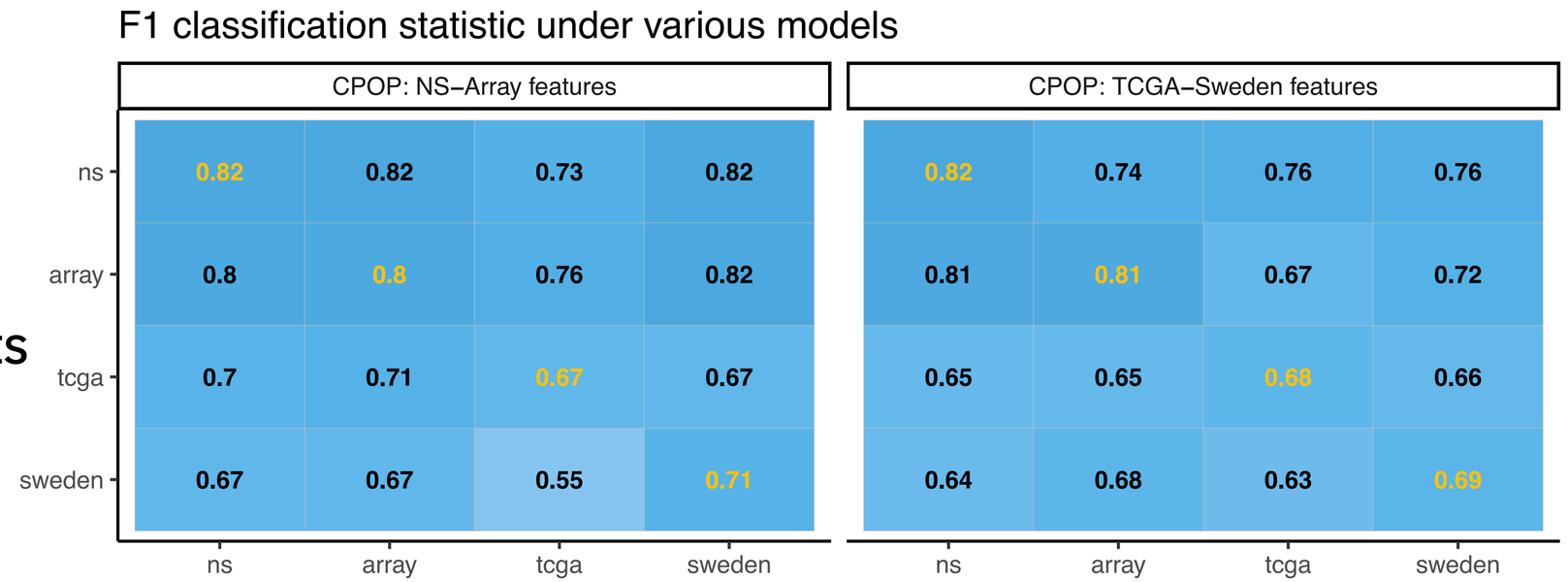
USA: log2-FPKM Illumina RNA-Seq, 99 samples

Sweden: Illumina HumanHT-12 version 4 microarray, 95 samples

- Concordance of log-ratios is much better compare to just gene expression

FAQ 5: weights in CPOP

CPOP: absolute distance weights



CPOP: KS-statistics weights