



**HAL**  
open science

## Impact of joint Dimension Reduction methods for survival prediction - Extension of a multi-omics benchmark study

Vincent Le Goff, Vincent Guillemot, Cathy Philippe, Gwendoline Mendes, Jean-François Deleuze, Edith Le Floch, Arnaud Gloaguen

### ► To cite this version:

Vincent Le Goff, Vincent Guillemot, Cathy Philippe, Gwendoline Mendes, Jean-François Deleuze, et al.. Impact of joint Dimension Reduction methods for survival prediction - Extension of a multi-omics benchmark study. Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), Jun 2024, Toulouse, France. cea-04734003

**HAL Id: cea-04734003**

**<https://cea.hal.science/cea-04734003v1>**

Submitted on 13 Oct 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Vincent LE GOFF<sup>1</sup>\*, Vincent GUILLEMOT<sup>2</sup>, Cathy PHILIPPE<sup>3</sup>, Gwendoline MENDES<sup>4</sup>, Jean-François DELEUZE<sup>1</sup>, Edith LE FLOCH<sup>1</sup>, Arnaud GLOAGUEN<sup>1</sup>

<sup>1</sup> Centre National de Recherche en Génomique Humaine (CNRGH), CEA, Université Paris-Saclay, Evry, France

<sup>2</sup> Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France

<sup>3</sup> NeuroSpin, CEA Saclay, Université Paris-Saclay, France

<sup>4</sup> CentraleSupélec, Gif-sur-Yvette, France

\* vincent.legoff@cnrgh.fr

## Introduction

This work aims to expand on a previous study [1] that compares the **survival predictions** of **13 different supervised multi-omics methods** on **datasets from 18 cancer types** from The Cancer Genome Atlas (TCGA).

This comparison is based on the analysis of **4 types of molecular data** (mRNA, miRNA, CNV, Mutations) and **clinical data**. Methods are divided into 3 groups: **reference models** (Kaplan-Meier and a Cox regression on clinical data only), **"naive" models**, unable to distinguish between each omics data, and **"structured" models**.

### Objectives:

- Building on the conclusion in [1] that structured methods perform better than naive ones, we include methods not only capable of exploiting the group structure of omics data but also of extracting links between them: **Joint Dimension Reduction (JDR)** methods. We test these methods in **unsupervised and supervised** settings whenever possible.
- Then, to go further on the evaluation of the importance of clinical data **we compare all of these methods (naive/structured/JDR) without clinical data** to see if molecular data can lead to similar prediction performance as with the addition of clinical data.

## JDR methods

### RGCCA/SGCCA

Regularized Generalized Canonical Correlation Analysis [2]

$$\text{argmax}_{\mathbf{w}_1, \dots, \mathbf{w}_L} \sum_{k,l=1}^L c_{kl} g(\text{cov}(\mathbf{X}_k \mathbf{w}_k, \mathbf{X}_l \mathbf{w}_l))$$

$$\text{s.t. } \begin{cases} (1-\tau_l) \text{var}(\mathbf{X}_l \mathbf{w}_l) + \tau_l \|\mathbf{w}_l\|_2^2 = 1, l = 1, \dots, L \\ \|\mathbf{w}_l\|_2^2 = 1 \text{ and } \|\mathbf{w}_l\|_1 \leq s_l \end{cases}, l = 1, \dots, L$$

### JIVE

Joint and Individual Variation Explained [3]

$$\mathbf{X}_l^T \approx \mathbf{U}_l \mathbf{S}_l + \mathbf{W}_l \mathbf{S}_l \text{ s.t. } \begin{cases} \mathbf{S}_l \mathbf{S}_l^T = 0 \\ \text{rank}(\mathbf{S}_l) = r_j \\ \text{rank}(\mathbf{S}_l) = r_a \end{cases}$$

### IntNMF

Integrative Non-Negative Matrix Factorization [4]

$$\mathbf{X}_l \approx \mathbf{W}_l \mathbf{H}_l^T \text{ s.t. } \mathbf{W}_l, \mathbf{H}_l \geq 0, l = 1, \dots, L$$

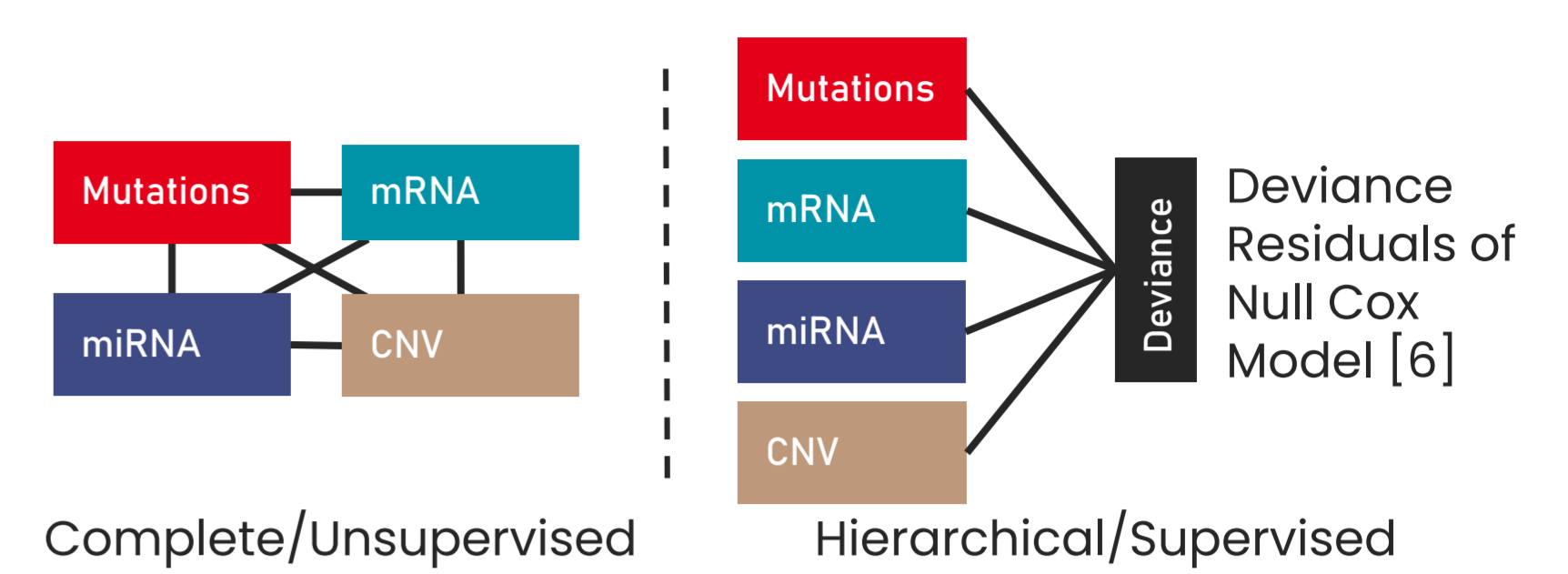
### MOFA

Multi-Omics Factor Analysis [5]

$$\mathbf{X}_l \approx \mathbf{Z} \mathbf{W}_l^T$$

fit through Bayesian methods

### RGCCA Unsupervised & Supervised



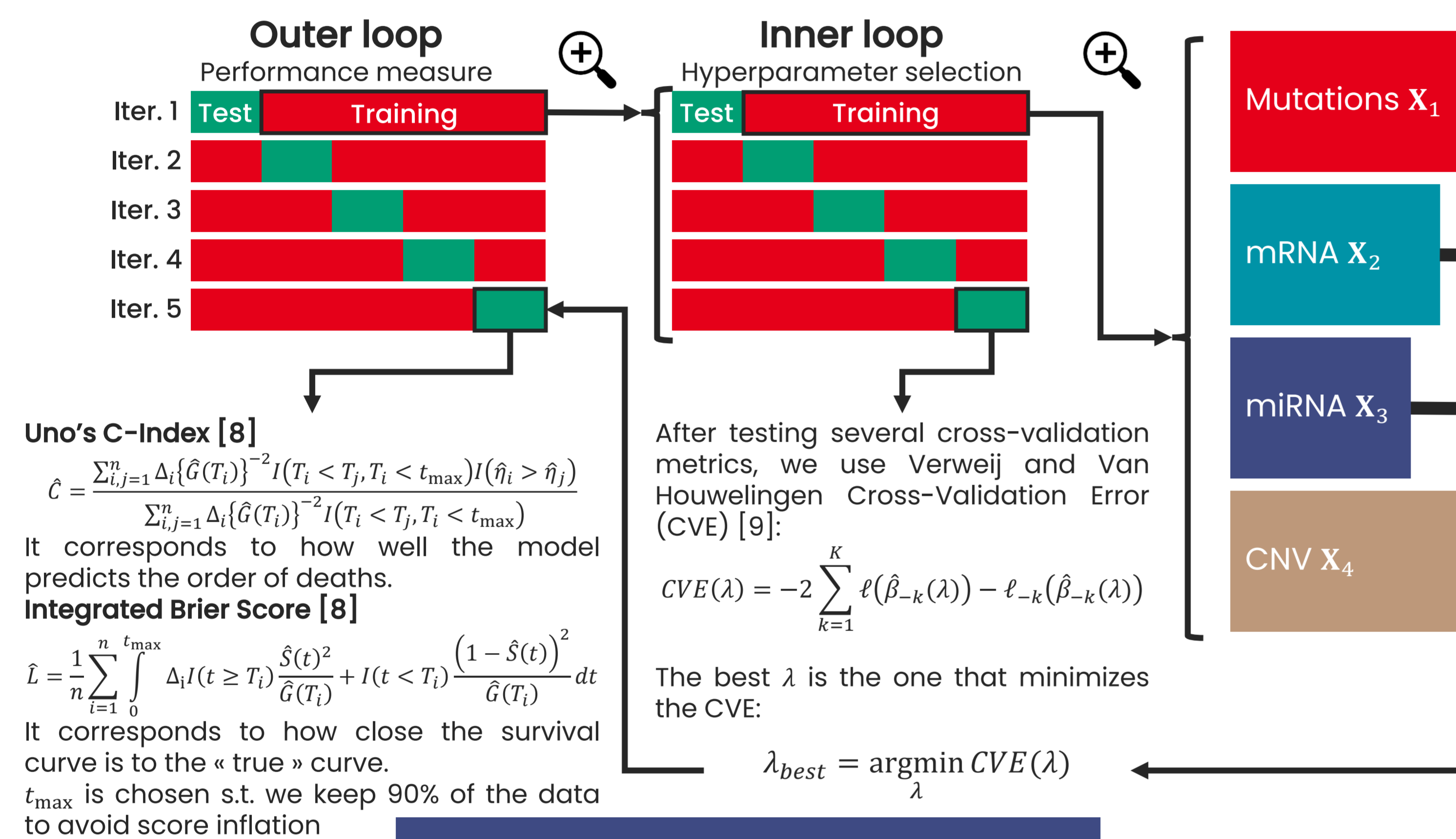
### Supervised JIVE [7]

$$\text{argmin}_{\mathbf{U}_1, \mathbf{S}_1, \mathbf{W}_1, \dots, \mathbf{U}_L, \mathbf{S}_L, \mathbf{W}_L} (1-\eta) \left\| \begin{bmatrix} \mathbf{X}_1^T \\ \vdots \\ \mathbf{X}_L^T \end{bmatrix} - \mathbf{U}_j - \begin{bmatrix} \mathbf{W}_1 \mathbf{S}_1 \\ \vdots \\ \mathbf{W}_L \mathbf{S}_L \end{bmatrix} \right\|_F^2 + \eta \left\| \mathbf{y}^T - \mathbf{S}_j^T \boldsymbol{\theta}_1 - \sum_{l=1}^L \mathbf{S}_l^T \boldsymbol{\theta}_{2l} \right\|_2^2$$

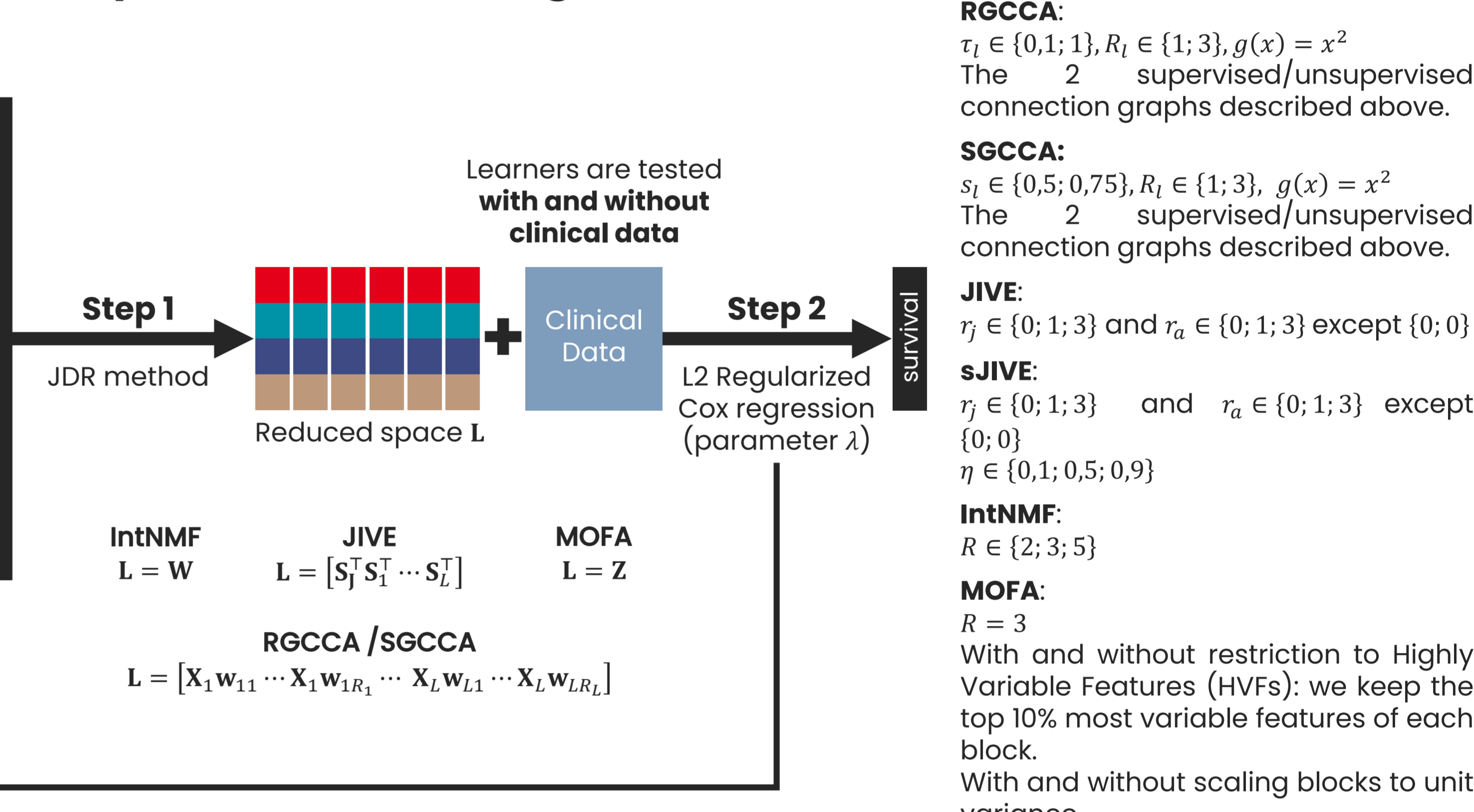
y: Deviance Residuals of Null Cox Model

## Survival Prediction with JDR

### Nested Cross-Validation



### Sequential Modeling



### Tested Parameters

**RGCCA**:  $\tau_l \in \{0, 1\}$ ;  $R_l \in \{1, 3\}$ ;  $g(x) = x^2$

The 2 supervised/unsupervised connection graphs described above.

**SGCCA**:  $s_l \in \{0, 5; 0, 75\}$ ;  $R_l \in \{1, 3\}$ ;  $g(x) = x^2$

The 2 supervised/unsupervised connection graphs described above.

**JIVE**:  $r_j \in \{0, 1; 3\}$  and  $r_a \in \{0, 1; 3\}$  except  $\{0, 0\}$

**sJIVE**:  $r_j \in \{0, 1; 3\}$  and  $r_a \in \{0, 1; 3\}$  except  $\{0, 0\}$

$\eta \in \{0, 1; 0, 5; 0, 9\}$

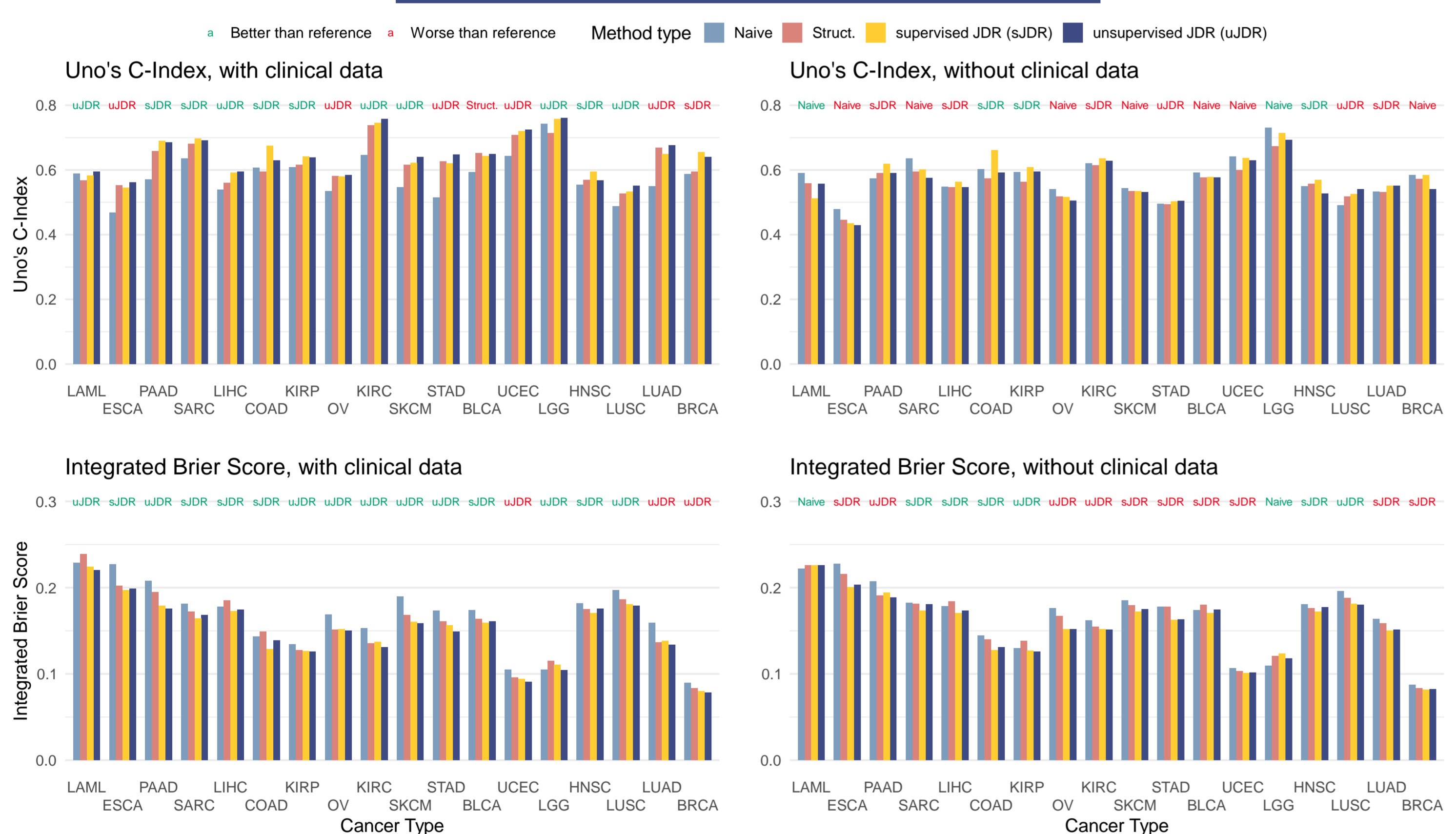
**IntNMF**:  $R \in \{2; 3; 5\}$

**MOFA**:  $R = 3$

With and without restriction to Highly Variable Features (HVF): we keep the top 10% most variable features of each block.

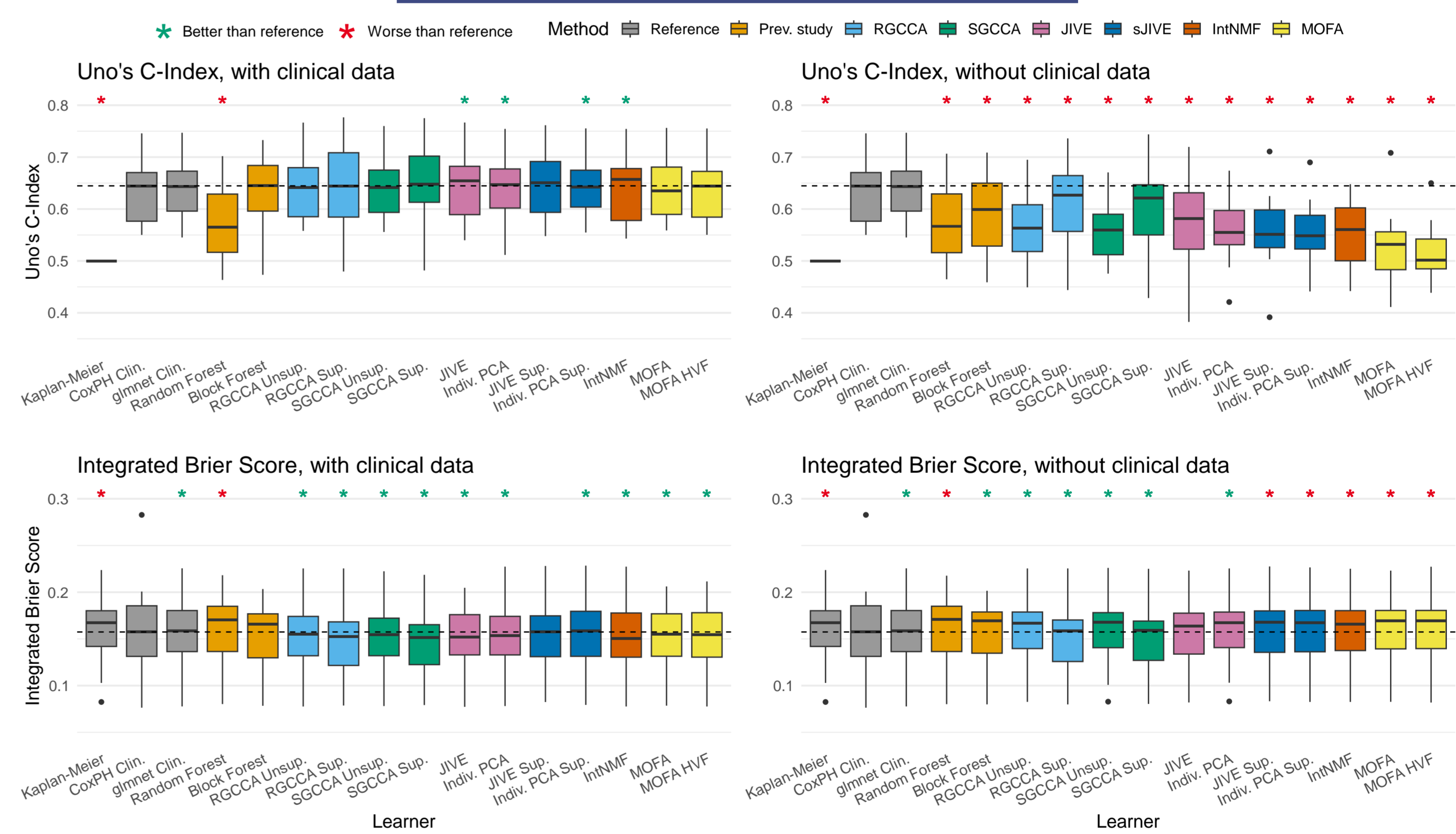
With and without scaling blocks to unit variance.

## Results by cancer



Barplots for each class of methods and each cancer type. All tested parameter combinations were averaged. The best class of method for each dataset is written above, and colored green if the average performance of the winning class is better than the mean performance of CoxPH Clin. We consider individual PCA (JIVE with  $r_j = 0$ ) as structured and joint PCA (JIVE with  $r_a = 0$ ) as a JDR method. Dataset are ordered by number of observations.

## Results by method



Boxplots of each tested method. Several parameter combinations were tested for each JDR methods, only the best performing ones are represented here. P-values were computed using a paired Wilcoxon test between the folds of the considered methods and those of CoxPH Clin. Combinations kept: RGCCA Sup./Unsup.:  $R_l = 1, \tau_l = 0, 1$ ; SGCCA Sup./Unsup.:  $R_l = 1, s_l = 0, 5$ ; JIVE:  $r_j = 3, r_a = 1$ ; PCA:  $r_j = 0, r_a = 1$ ; JIVE Sup.:  $r_j = 3, r_a = 1, \eta = 0, 9$ ; PCA Sup.:  $r_j = 0, r_a = 1, \eta = 0, 9$ ; IntNMF:  $R = 3$ ; MOFA: 3 factors, scaled MOFA HVF: 3 factors, scaled.

## Conclusion

### With clinical data

- JDR methods** (supervised and unsupervised) **improve both C-Index and IBS**.
- IntNMF has the best median performance** across datasets.

### Without clinical data

- C-index and IBS are inferior across all methods.
- JDR are the best performing non reference methods.
- Supervised RGCCA and SGCCA** have the highest performance.
- Supervising JIVE does not seem to improve performance**, as is the case with RGCCA/SGCCA.

## Perspectives

- Add **more JDR methods** in the benchmark (MCIA, iCluster, Scikit-Fusion).
- Use an **automatic procedure** to select the **best performing parameters** for each method.
- Include methods capable of extracting **Common, Local and Distinct (CLD)** components, in line with the good performance of JIVE.
- Design a new method** based on the best performing ones of the benchmark to exploit biological *a priori* information.

Dataset	Best method with Clinical Data		Best method without Clinical Data	
	C-Index	IBS	C-Index	IBS
BRCA	RGCCA Sup.	CoxPH Clin.	CoxPH Clin.	CoxPH Clin.
LUAD	sJIVE	CoxPH Clin.	CoxPH Clin.	CoxPH Clin.
LUSC	Joint PCA *	sJIVE	Joint PCA	Priority LASSO *
HNSC	RGCCA Sup. *	SGCCA Sup. *	RGCCA Sup. *	SGCCA Sup. *
LGG	LASSO *	MOFA HVF *	LASSO	MOFA HVF *
UCEC	CoxPH Clin.	CoxPH Clin.	CoxPH Clin.	CoxPH Clin.
BLCA	IntNMF *	SGCCA Sup. *	RGCCA Sup.	SGCCA Sup. *
STAD	sJIVE *	IntNMF *	Glimnet Clin. *	CoxPH Clin.
SKCM	JIVE *	sJIVE *	Glimnet Clin.	Glimnet Clin.
KIRC	Joint PCA *	Joint PCA *	Glimnet Clin.	CoxPH Clin.
OV	MOFA HVF	MOFA HVF	IntNMF	IntNMF
KIRP	IntNMF *	Block Forest	RGCCA Unsup.	Individual PCA
COAD	RGCCA Sup. *	RGCCA Sup. *	RGCCA Sup. *	RGCCA Sup. *
LHNC	Individual PCA *	Ranger *	RGCCA Sup.	Ranger *
SARC	RGCCA Sup. *	SGCCA Sup. *	LASSO	RGCCA Sup. *
PAAD	sJIVE *	SGCCA Unsup. *	Glimnet Clin. *	SGCCA Unsup. *
ESCA	SGCCA Unsup.	RGCCA Sup.	CoxPH Clin.	RGCCA Sup.
LAML	Block Forest *	MOFA HVF *	Ranger *	Ranger *