Reproducibility and Cross-study Replicability of Prognostic Signatures from High Throughput Genomic Data

giovanni_parmigiani@dfci.harvard.edu

2014 Rutgers Statistics Symposium
signatures and prognosis
CuratedOvarianData

23 ovarian cancer microarray studies

ID  Debulk  Status  ...  sampleid  debulking  vital_status  ...
D2640 S  0
GSM123 opt DOD

Machine syntax check

Human double check

Download expression data

Affymetrix platform

Raw data?

(f) RMA re-normalization

Collapse probesets to genes

Probeset  Gene  GSM123  GSM124
204531_s_at  BRCA1  4.0  4.1
211851_x_at  BRCA1  5.0  6.0

Gene  GSM123  GSM124
BRCA1  5.0  6.0

Automatically build documented curatedOvarianData R package
Meta-analysis overview

**Literature review**

- **Prognostic models**
  - 101 candidate papers
  - Five review papers

- **Database of curated gene expression**
  - Standardized clinical annotation and gene ID
  - 23 studies, 2,908 samples

**Inclusion Criteria**

**Prognostic models**

- Training sample size > 40
- Focus on late-stage serous
- Multivariable model
- Continuous risk score
- Claims to predict survival
- Possible to reproduce model

**Database of curated gene expression**

- Sample size > 40
- Primary tumors
- Overall survival available
- Events (deaths) > 15
- Late stage, high grade tumors
- Serous subtype

**14 prediction models implemented**

100 pages documentation

*survHD* Bioconductor package

**10 datasets, 1455 samples**

*curatedOvarianData* Bioconductor package
<table>
<thead>
<tr>
<th>Implemented Models</th>
<th>Validation Statistics for 14 Models in 10 Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dataset Average</strong></td>
<td>1.81 1.47 1.43 1.41 1.39 1.37 1.35 1.14 1.11 1.04</td>
</tr>
<tr>
<td>TCGA11</td>
<td>2.05 2.28 1.58 1.85 1.36 1.64 1.97 <strong>1.94</strong> 1.07 1.07</td>
</tr>
<tr>
<td>Yoshihara12</td>
<td>2.44 <strong>9.65</strong> 1.21 1.8 1.02 <strong>1.77</strong> 1.97 <strong>1.21</strong> 1.35 1.42</td>
</tr>
<tr>
<td>Yoshihara10</td>
<td>2.69 1.38 1.15 <strong>1.93</strong> 1.45 1.57 1.47 1.33 0.7 <strong>7.27</strong></td>
</tr>
<tr>
<td>Kernagis12</td>
<td>2.65 1.39 2.91 <strong>2.08</strong> 1.45 1.39 1.25 1.23 1.32 0.87</td>
</tr>
<tr>
<td>Crijns09</td>
<td>1.22 1.92 1.49 1.51 1.2 1.44 1.28 1.1 <strong>3.04</strong> 1.21</td>
</tr>
<tr>
<td>Bentink12</td>
<td>1.94 1.01 1.89 1.44 1.2 <strong>1.14</strong> 1.62 <strong>1.16</strong> 1.26 1.45</td>
</tr>
<tr>
<td>Bonome08_263genes</td>
<td>1.3 2.73 2 1.32 0.53 <strong>2.01</strong> 1.45 1.17 1.03 0.77</td>
</tr>
<tr>
<td>Mok09</td>
<td>1.54 <strong>1.62</strong> 3.18 1.71 0.89 1.58 1.28 0.98 0.95 1.39</td>
</tr>
<tr>
<td>Bonome08_572genes</td>
<td>0.8 1.89 1.1 1.41 <strong>2.29</strong> <strong>2.27</strong> 1.47 1.07 1.35 0.84</td>
</tr>
<tr>
<td>Sabatier11</td>
<td>1.95 <strong>1.41</strong> 1.17 <strong>1.41</strong> 1.72 1.07 1.19 1.11 1.3 0.73</td>
</tr>
<tr>
<td>Denkert09</td>
<td>2.6 <strong>1.33</strong> 1.31 2.25 1.04 1.29 1.08 1.15 0.79</td>
</tr>
<tr>
<td>Kang12</td>
<td><strong>2.14</strong> 1.19 0.81 <strong>0.85</strong> 1.21 1.46 <strong>1.55</strong> 1.02 0.73</td>
</tr>
<tr>
<td>Konstantinopoulos10</td>
<td>1.34 1.01 0.82 1.07 <strong>2.05</strong> 1.07 1</td>
</tr>
<tr>
<td>Hernandez10</td>
<td><strong>0.68</strong> 0.55 1.07 0.71 0.86 <strong>1.21</strong> <strong>0.87</strong> 1.04 0.9 1.03</td>
</tr>
</tbody>
</table>
meta-analysis of cross-study performance

<table>
<thead>
<tr>
<th>Implemented Models</th>
<th>Validation Statistics for 14 Models in 10 Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset Average</td>
<td>1.81 1.47 1.43 1.41 1.39 1.35 1.14 1.11 1.04</td>
</tr>
<tr>
<td>TCGA11</td>
<td>2.05 2.28 1.58 1.85 1.36 1.64 1.97 1.94 1.07 1.53</td>
</tr>
<tr>
<td>Yoshihara10</td>
<td>2.69 1.38 1.15 1.93 1.45 1.57 1.47 1.33 0.77 7.27</td>
</tr>
<tr>
<td>Kernagis12</td>
<td>2.65 1.39 2.91 2.08 1.45 1.39 1.25 1.23 1.32 0.87</td>
</tr>
<tr>
<td>Crijns09</td>
<td>1.22 1.92 1.49 1.51 1.2 1.44 1.28 1.1 3.04 1.21</td>
</tr>
<tr>
<td>Bentink12</td>
<td>1.94 1.01 1.89 1.44 1.2 1.14 1.62 1.16 1.26 1.45</td>
</tr>
<tr>
<td>Bonome08_263genes</td>
<td>1.3 2.73 2 1.32 0.53 2.01 1.45 1.17 1.03 0.77</td>
</tr>
<tr>
<td>Mok09</td>
<td>1.54 1.82 3.18 1.71 0.89 1.58 1.28 0.98 0.95 1.39</td>
</tr>
<tr>
<td>Bonome08_572genes</td>
<td>0.8 1.89 1.1 1.41 2.29 2.27 1.47 1.07 1.35 0.84</td>
</tr>
<tr>
<td>Sabatier11</td>
<td>1.95 1.15 1.17 1.41 1.72 1.07 1.19 1.11 1.3 0.73</td>
</tr>
<tr>
<td>Denkert09</td>
<td>2.6 0.76 1.33 1.31 2.25 1.04 1.29 1.08 1.15 0.79</td>
</tr>
<tr>
<td>Kang12</td>
<td>2.14 1.19 0.81 0.85 1.21 1.46 1.17 1.55 1.02 0.73</td>
</tr>
<tr>
<td>Konstantinopoulos10</td>
<td>1.34 1.01 0.82 1.07 2.05 1.07 1 1.15 0.97 1.09</td>
</tr>
<tr>
<td>Hernandez2012A</td>
<td>0.68 0.55 1.07 0.71 0.86 1.21 0.79 1.04 0.9 1.03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expression Datasets</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dressman</td>
<td>0.89 1.00 1.12 1.26 1.41 1.58 1.78 2.00</td>
</tr>
<tr>
<td>Yoshihara 2012A</td>
<td></td>
</tr>
<tr>
<td>Konstantinopoulos</td>
<td></td>
</tr>
<tr>
<td>TCGA</td>
<td></td>
</tr>
<tr>
<td>Bentink</td>
<td></td>
</tr>
<tr>
<td>Crijns</td>
<td></td>
</tr>
<tr>
<td>Yoshihara 2010</td>
<td></td>
</tr>
</tbody>
</table>

Legend:
- ■ author training set
- □ author test set
- Author validation
- Summary (95% CI)
- Excl. author test sets
using c-stat instead

Fixed effects, ComBat batch correction
Random effects, ComBat batch correction
Fixed effects, Dressman excluded
Fixed effects, strict sample exclusion

Color Key

Value

0.8 1.2 1.6

1.62 1.62 1.62 1.65 1.65 1.62 1.74 1.74
1.45 1.46 1.45 1.48 1.51 1.46 1.48 1.52
1.45 1.48 1.34 1.38 1.42 1.35 1.37 1.4
1.42 1.42 1.34 1.39 1.4 1.33 1.39 1.42
1.31 1.35 1.26 1.26 1.28 1.28 1.28 1.3 1.34
1.25 1.26 1.3 1.3 1.3 1.3 1.29 1.29
1.2 1.2 1.22 1.24 1.27 1.26 1.25 1.29
1.22 1.22 1.26 1.23 1.24 1.26 1.18 1.2
1.15 1.17 1.13 1.17 1.19 1.13 1.13 1.13
1.17 1.16 1.1 1.15 1.14 1.1 1.17 1.16
1.25 1.25 1.16 1.19 1.19 1.16 1.16 1.16
1.13 1.13 1.08 1.09 1.09 1.08 1.07 1.07
0.94 0.92 0.95 0.93 0.93 0.95 0.94 0.93

TCGA11
Yoshihara12
Kernagis12
Yoshihara10
Bonome08_263genes
Bentink12
Crijns09
Mok09
Bonome08_572genes
Denkert09
Kang12
Sabatier11
Konstantinopoulos10
Hernandez10
sensitivity analysis

(A) Implemented Models

<table>
<thead>
<tr>
<th>Expression Datasets</th>
<th>Dressman</th>
<th>Yoshihara 2012A</th>
<th>T6hills</th>
<th>Bentink</th>
<th>Bonome</th>
<th>Konstantinopoulos</th>
<th>Mok</th>
<th>Yoshihara 2010</th>
<th>TCGA</th>
<th>Crijns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset Average</td>
<td>0.61</td>
<td>0.58</td>
<td>0.57</td>
<td>0.56</td>
<td>0.56</td>
<td>0.55</td>
<td>0.54</td>
<td>0.54</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>TCGA11</td>
<td>0.62</td>
<td>0.69</td>
<td>0.6</td>
<td>0.63</td>
<td>0.61</td>
<td>0.47</td>
<td>0.57</td>
<td>0.6</td>
<td>0.64</td>
<td>0.55</td>
</tr>
<tr>
<td>Yoshihara12</td>
<td>0.63</td>
<td>0.81</td>
<td>0.64</td>
<td>0.6</td>
<td>0.62</td>
<td>0.51</td>
<td>0.5</td>
<td>0.58</td>
<td>0.57</td>
<td>0.55</td>
</tr>
<tr>
<td>Bonome08_263genes</td>
<td>0.57</td>
<td>0.68</td>
<td>0.58</td>
<td>0.6</td>
<td>0.62</td>
<td>0.53</td>
<td>0.6</td>
<td>0.54</td>
<td>0.56</td>
<td>0.52</td>
</tr>
<tr>
<td>Yoshihara10</td>
<td>0.7</td>
<td>0.55</td>
<td>0.62</td>
<td>0.53</td>
<td>0.53</td>
<td>0.54</td>
<td>0.8</td>
<td>0.56</td>
<td>0.56</td>
<td>0.52</td>
</tr>
<tr>
<td>Kernagis12</td>
<td>0.66</td>
<td>0.58</td>
<td>0.63</td>
<td>0.56</td>
<td>0.55</td>
<td>0.55</td>
<td>0.65</td>
<td>0.57</td>
<td>0.55</td>
<td>0.54</td>
</tr>
<tr>
<td>Sabatier11</td>
<td>0.64</td>
<td>0.54</td>
<td>0.56</td>
<td>0.57</td>
<td>0.54</td>
<td>0.62</td>
<td>0.55</td>
<td>0.57</td>
<td>0.56</td>
<td>0.52</td>
</tr>
<tr>
<td>Crijns09</td>
<td>0.5</td>
<td>0.6</td>
<td>0.59</td>
<td>0.55</td>
<td>0.58</td>
<td>0.55</td>
<td>0.56</td>
<td>0.47</td>
<td>0.54</td>
<td>0.67</td>
</tr>
<tr>
<td>Bentink12</td>
<td>0.65</td>
<td>0.56</td>
<td>0.55</td>
<td>0.61</td>
<td>0.57</td>
<td>0.57</td>
<td>0.53</td>
<td>0.53</td>
<td>0.53</td>
<td>0.52</td>
</tr>
<tr>
<td>Bonome08_572genes</td>
<td>0.57</td>
<td>0.6</td>
<td>0.54</td>
<td>0.55</td>
<td>0.64</td>
<td>0.63</td>
<td>0.55</td>
<td>0.5</td>
<td>0.53</td>
<td>0.54</td>
</tr>
<tr>
<td>Mok09</td>
<td>0.53</td>
<td>0.6</td>
<td>0.56</td>
<td>0.57</td>
<td>0.57</td>
<td>0.53</td>
<td>0.69</td>
<td>0.57</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>Kang12</td>
<td>0.63</td>
<td>0.54</td>
<td>0.52</td>
<td>0.54</td>
<td>0.57</td>
<td>0.54</td>
<td>0.49</td>
<td>0.54</td>
<td>0.58</td>
<td>0.52</td>
</tr>
<tr>
<td>Denkert09</td>
<td>0.67</td>
<td>0.52</td>
<td>0.54</td>
<td>0.53</td>
<td>0.53</td>
<td>0.58</td>
<td>0.53</td>
<td>0.51</td>
<td>0.52</td>
<td>0.55</td>
</tr>
<tr>
<td>Hernandez10</td>
<td>0.56</td>
<td>0.61</td>
<td>0.56</td>
<td>0.54</td>
<td>0.53</td>
<td>0.5</td>
<td>0.5</td>
<td>0.54</td>
<td>0.49</td>
<td>0.51</td>
</tr>
<tr>
<td>Konstantinopoulos10</td>
<td>0.57</td>
<td>0.5</td>
<td>0.52</td>
<td>0.48</td>
<td>0.49</td>
<td>0.6</td>
<td>0.5</td>
<td>0.51</td>
<td>0.53</td>
<td>0.5</td>
</tr>
</tbody>
</table>

(B) Summary Validation Statistics

- author training set
- author test set
- model validation
- summary (95% CI)
excl. author test sets
do predictors rank patients similarly?
selection bias in choice of validation study?
A) Methodology for comparing prognostic quality of gene sets to random gene sets

1. Published signature
   - Select one dataset for training: identify good and bad-prognosis genes
   - Determine allowed test datasets (not used for training published signature)
2. 10 random signatures of same size
3. Test prognostic accuracy in every non-training dataset

B) Gene set Improvement over Random Signatures

- Best fit for random signatures
- 95% confidence interval for best fit
- 95% prediction interval for best fit
- Individual random signatures
- Size = IOR of each published signature
- Size = random expected IOR

**Published signature**
Create "Leave-one-dataset-in" matrix of validation C-statistics for published signature

**Random signatures x10**
Create equivalent C' matrices for 10 random signatures of the same size

<table>
<thead>
<tr>
<th>training sets</th>
<th>test sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5</td>
<td>x1</td>
</tr>
<tr>
<td>1 2 3 4 5</td>
<td>x10</td>
</tr>
</tbody>
</table>

*Author training sets are excluded
† cross-validation statistics are not used

**IOR score**
Proportion entries in C matrix larger than corresponding entry in C' matrix
lessons: genomic signatures

- template for meta-analytic signature evaluation
- published ovarian cancer signatures and predictors largely withstand cross-study analysis
- published ovarian cancer signatures and predictors are not very clinically useful
multi-study comparison of classification algorithms

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Adjuvant therapy</th>
<th># patients</th>
<th># ER+</th>
<th>3Q survival [mo.]</th>
<th>Median follow-up [mo.]</th>
<th>Original identifiers</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CAL</td>
<td>chemo, hormonal</td>
<td>118</td>
<td>75</td>
<td>42</td>
<td>82</td>
<td>CAL</td>
<td>Chin et al. (2006)</td>
</tr>
<tr>
<td>2</td>
<td>MNZ</td>
<td>none</td>
<td>200</td>
<td>162</td>
<td>120</td>
<td>94</td>
<td>MAINZ</td>
<td>Schmidt et al. (2008)</td>
</tr>
<tr>
<td>3</td>
<td>MSK</td>
<td>combination</td>
<td>99</td>
<td>57</td>
<td>76</td>
<td>82</td>
<td>MSK</td>
<td>Minn et al. (2005)</td>
</tr>
<tr>
<td>4</td>
<td>ST1</td>
<td>hormonal</td>
<td>512*</td>
<td>507*</td>
<td>114</td>
<td>106</td>
<td>MDA5, TAM, VDX3</td>
<td>Foekens et al. (2006)</td>
</tr>
<tr>
<td>5</td>
<td>ST2</td>
<td>hormonal</td>
<td>517</td>
<td>325</td>
<td>126</td>
<td>121</td>
<td>EXPO, TAM</td>
<td>Symmans et al. (2010)</td>
</tr>
<tr>
<td>6</td>
<td>TRB</td>
<td>none</td>
<td>198</td>
<td>134</td>
<td>143</td>
<td>171</td>
<td>TRANSBIG</td>
<td>Desmedt et al. (2007)</td>
</tr>
<tr>
<td>7</td>
<td>UNT</td>
<td>none</td>
<td>133</td>
<td>86</td>
<td>151</td>
<td>105</td>
<td>UNT</td>
<td>Sotiriou et al. (2006)</td>
</tr>
<tr>
<td>8</td>
<td>VDX</td>
<td>none</td>
<td>344</td>
<td>209</td>
<td>44</td>
<td>107</td>
<td>VDX</td>
<td>Minn et al. (2007)</td>
</tr>
</tbody>
</table>

Cross-study C-statistics, Ridge Regression
Simulations for method comparison

Bernau et al. 2014

Resulting linear predictors rather than on the vectors of regression coefficients, the Euclidean distance between them. However, since our focus is on the differences between the rankings and performance estimates as obtained by CV and CSV respectively.

### 4. RESULTS

Another criterion of interest in practice is the ability of a learning algorithm to recover the true model in terms of the linear predictor when fitted on a dataset drawn from this same model, hence the index \( S_i \) and \( S_{ij} \).

For a given collection of datasets, the CV and CSV procedures considered across \( S_i \) studies, we obtain a measure of the ability of learning algorithm to recover the true model in terms of the linear predictor when fitted on a dataset simulated from another study \( (S_{ij} = 1) \), regardless of their using for training (rows of the Z-matrix) or validation (columns of the Z-matrix), although CV correlation coefficient mentioned above.

\[
I_{ij} = \frac{1}{n} \sum_{k=1}^{n} \left( \hat{y}_{ij}^k - \bar{y}_{ij}^k \right)^2
\]

Where the index \( I_{ij} \) means that (in contrast to \( I \)), whereby \( I_{ij} \) are by definition specific to the considered studies and datasets, since involving vector \( Z \).

\[
I_{ij} = I_{ij} + S_{ij}(1 - S_{ij})
\]

Hence, the corresponding ranking can be seen as "local" or, in other words, specific to the collection of datasets at hand.

\[
\text{Value}_{ij} = \frac{1}{n} \sum_{k=1}^{n} \left( \hat{y}_{ij}^k - \bar{y}_{ij}^k \right)^2
\]

For CAL and MSK, C-statistics for Ridge Regression:

\[
\text{Cor} = 0.896
\]

Ridge Regression: Cor = 0.896

\[
\text{C} = 0.5, 0.6, 0.7, 0.8
\]

The standard way to assess similarity between vectors is to compute, say,

\[
\text{Sim} = \frac{1}{n} \sum_{k=1}^{n} \left( \hat{y}_{ij}^k - \bar{y}_{ij}^k \right)^2
\]

As an illustration, for each pair of studies the second panel of Figure 1 illustrates the high correlation between the performance in cross-study prediction on real and simulated data.

Fig. 1. Comparison of C-Indices in cross-study validation on real (left panel) and simulated data (middle panel) for Ridge Regression. For CAL and MSK, C-indices) than CSV.

The similarity between the two panels is striking, in particular with the diagonal containing the mean C-indices obtained in the other study (with the diagonal containing the mean C-indices obtained).

The average scatter plot illustrates the high correlation between the performance in cross-study prediction on real and simulated data.

All values of the Z-matrix corresponding to these two models build the red ‘bad performance’ cluster in the scatter plot in the right panel. Additionally, this is regardless of their using for training (rows of the Z-matrix) or validation (columns of the Z-matrix), although CV...
Cross-study validation
call the corresponding rankings
local because they are specific to
the collection of datasets at ...

5

55

cross (study) validation

R

4, and also surpasses
Kendall correlation:

CV

studies (CAL and MSK) are removed, the advantage of
more suitable for recovering the global ranking. If the two outlier
in the first one. Finally,

defined in terms of within-study validation. However, the difference
the local ranking
tends to be less correlated with
recovered by

Kendall's correlation between local rankings

5

across

k

Performance differences across algorithms, whether estimated by

Glmnetridge

worst performing algorithms, while

CSV . By both CV and CSV ,
differences between the rank distributions produced by CV and

methods. We observe large differences in the distributions of

method

8-dataset compendia. Table 2 shows the median of these rank

2b shows the distribution of the rankings, across 1000 simulated

Figure 2a shows the distributions of

CV

and

CSV . By both CV and CSV ,

exchange in top global true ranking between

as estimated by

Glmnetlasso

universities.

3.1 Simulated Data

3 RESULTS

3.1 Simulated Data

Figure 2a shows the distributions of

CV

and

CSV . By both CV and CSV ,

exchange in top global true ranking between

as estimated by

Glmnetlasso

universities.

3.1 Simulated Data

Figure 2a shows the distributions of

CV

and

CSV . By both CV and CSV ,

exchange in top global true ranking between

as estimated by

Glmnetlasso

universities.

3.1 Simulated Data

Figure 2a shows the distributions of

CV

and

CSV . By both CV and CSV ,

exchange in top global true ranking between

as estimated by

Glmnetlasso

universities.
with a median rank of 5 over the 1000 simulated datasets, i.e. larger than ranks this algorithm first in all simulated iterations (see Figure the performance differences between the algorithms are relatively in the 1000 simulated datasets. It can be seen from this table that and the 3rd quartile (right), as approximated through computation based on CoxBoost most often ranks 5th based on necessarily receive the same rank with corresponding to the 1000 simulated datasets. (bottom panel) for Fig. 2.

We focus on the local ranking. Performance estimates are also noticeably less variable than resulting rankings. As far as the absolute level of the performance distributions of CSV and CV (top panel) for Fig. 2 representing the simulation result displays the iterationwise averages of the $Z$-matrix. This

Estimated model ranks

Distribution of ranks

Lasso  Ridge  Plusminus  Unicox  SuperPC  CoxBoost

CV

CSV
method for the detection of outlier studies, since it estimates the higher variance observed in Figure 4. Besides, we observe on Figure 4 that:

Due to the presence of outlier studies first quartiles are around 0.5 for CSV, however, these two studies always form a separate cluster with poorer performance. From this point of view, cross-study prediction and within-study prediction are indeed cross-study performance and cross-study estimates averaged over all studies specific values for CV but two values for CSV depending on whether we average the learning algorithm, the relation between cross-study performance and cross-study estimates and row-wise (identical validation study). The correlations vary over the algorithms but most of them are less related than one might expect. In the lower panel of Figure 4, the distributions see Glmnetridge as the clear winner and estimates a better performance for CoxBoost than CSV. CV can be seen that Plusminus and Glmnetridge seem to perform strongly adapt to the specific properties of an individual data set. This problem can be analyzed in greater detail.

This leads to the question whether some algorithms might be considered as specialist algorithms according to the definition given in the introduction. The fact that the ranks of Glmnetridge and SuperPC performs distinctly better in the C-Index scale. The latter observation suggests that this two algorithms strongly adapt to the specific properties of an individual data set. This problem can be analyzed in greater detail.

Although the performance differences are small with respect to the C-Index scale, more noticeable differences are observed with respect to the C-Index estimates. In the upper panel of Figure 4, the distribution of Kendall's correlation between the algorithm rankings shows that cross-study prediction and within-study prediction are indeed strongly related. The correlations vary over the algorithms but the values of the adjusted R² are almost all negative. No algorithm yields noticeably higher values for the adjusted R² than in the study specific values for R adj. (T) = 0.12, R adj. (V) = 0.03.
STAGE 1: Approximate the distribution of the centered array

\[
(Z_{s,v} - E_{p_s,p_v}(Z_{s,v}); \ s, v = 1, \ldots, S)
\]

STAGE 2: Model-based clustering of the studies 1, \ldots, S

| RANDOM PARTITION of \{1, \ldots, S\} |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| 1               | 2               | 3               | 4               | 5               |
| E(Z_{12})       | E(Z_{13})       | E(Z_{14})       | E(Z_{15})       |
| E(Z_{21})       | E(Z_{23})       | E(Z_{24})       | E(Z_{25})       |
| E(Z_{31})       | E(Z_{32})       |                | E(Z_{34})       | E(Z_{35})       |
| E(Z_{41})       | E(Z_{42})       | E(Z_{43})       |                | E(Z_{45})       |
| E(Z_{51})       | E(Z_{52})       | E(Z_{53})       | E(Z_{54})       |
simulations to illustrate clustering

Logistic reg., N=300, 100 covariates and $\|\beta_1\|, \ldots, \|\beta_{10}\|$ strictly positive.

Z: Misclassification rates.

Measurement errors: \{1, 2, 3\} LOW, \{4, 5, 6\} MEDIUM, \{7, 8, 9\} HIGH.
Ridge regression, $\tilde{Z}_{B(s),s}$

Validation datasets:

- 1
- 2
- 3
- 4
- 5
- 6
Model evaluation should be context specific

Simulations to evaluate methodology should be based on modeling of predictive distributions

Multi-study model evaluations are different from single-study and more faithfully represent scientific reproducibility
Prognostic model validation in precision medicine is a meta-analysis problem

The foundation of research on statistical learning should emphasize empirical multi-study reproducibility
Credits

Mike Birrer, Lorenzo Trippa, Curtis Huttenhower
Levi Waldron, Markus Reister, Ben Ganzfried, Christopher Bernau
Benjamin Haibe-Kains, Aedín C. Culhane, Jie Ding, Xin Victoria Wang, Mahnaz Ahmadifar, Svitlana Tyekucheva, Thomas Risch