A comparison study of prediction approaches for multiple training data sets & test data with block-wise missing values

Author: Frederik Ludwigs
Supervisor: Dr. Roman Hornung

Ludwig-Maximilians-University Munich
Institute for Statistics

Master Thesis

XX. August 2020

Table of Contents

Block-Wise Missingness

Block-wise missingness is a special type of missingness that is common in practice, particular in the context of multi-omics data [1]

weight	*height*	*income*	*education*	g_1	$\cdots$	g_{100}	Y	
65.4	187	2.536	*Upper*				1	**Fold1**
83.9	192	1.342	*Lower*				0	
67.4	167	5.332	*Upper*				1	
		743	*Lower*	-0.42	$\cdots$	1.43	1	**Fold2**
		2.125	*Lower*	0.52	$\cdots$	-1.37	0	
105.2	175			-1.53	$\cdots$	2.01	0	**Fold3**
71.5	173			0.93	$\cdots$	0.53	0	
73.0	169			0.31	$\cdots$	-0.07	1	
Block 1		**Block 2**		**Block 3**				

Table: A data set with block-wise missingness - consisting of three feature-blocks, three folds and a binary target variable 'Y'.

Random Forest Model (RF)

- Introduced by Breiman in 2001 [2]
- Ensemble method that uses the decision tree as a base learner
- Modified bagging algorithm to construct decorrelated decision trees
- A prediction of a RF equals the average of the predictions from all its decision trees
- The RF can be evaluated internally with the OOB-error
  $\rightarrow$ "almost identical to that obtained by N-fold cross-validation" [[3], p. 593]

Complete-Case Approach

Idea: Process the training data - with regard to the test data -, such that it does not contain any missing values afterwards:

1. Remove all folds from the training data that miss at least one feature-block from the test data
2. Remove all feature-blocks from the training data that are not available for the test data
   → On this data a RF can be trained regularly & create predictions for the test data then!

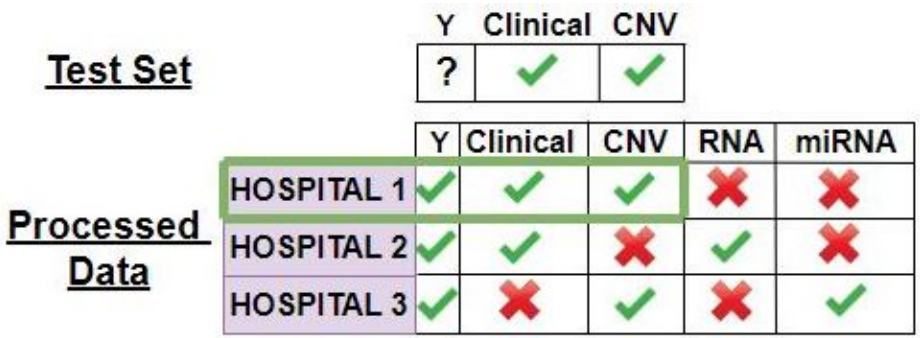

Figure: The 'Complete-Case' processing of the training data according to the available feature-blocks in the test-set.

Single-Block Approach

Idea: Use a single feature-block of the train data - that the train & test data have in common - to train a RF and create predictions on the test data then:

1. Train a RF on the single feature-blocks that test and train data have in common (remove observations with missing values from the blocks)
2. Each fitted RF can create predictions for the test data then
   $\rightarrow$ can result in multiple predictions for the test data

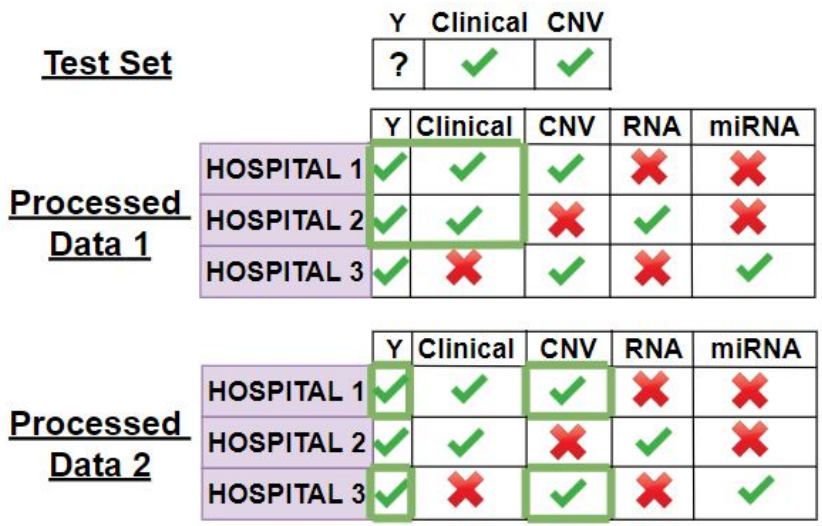

Figure: 'Single-Block' processing of the training data so a random forest model can be regularly trained with each of these processed data sets.

Imputation Approach

Idea: Impute the missing values in the train data and fit a RF on the feature-blocks that the train & test data have in common:

1. Impute missing values in the train data with the 'missForest' approach
   $\rightarrow$ train data is completely observed then
2. On the feature-blocks that the train & test data have in common a RF can be trained regularly and create predictions for the test data then!

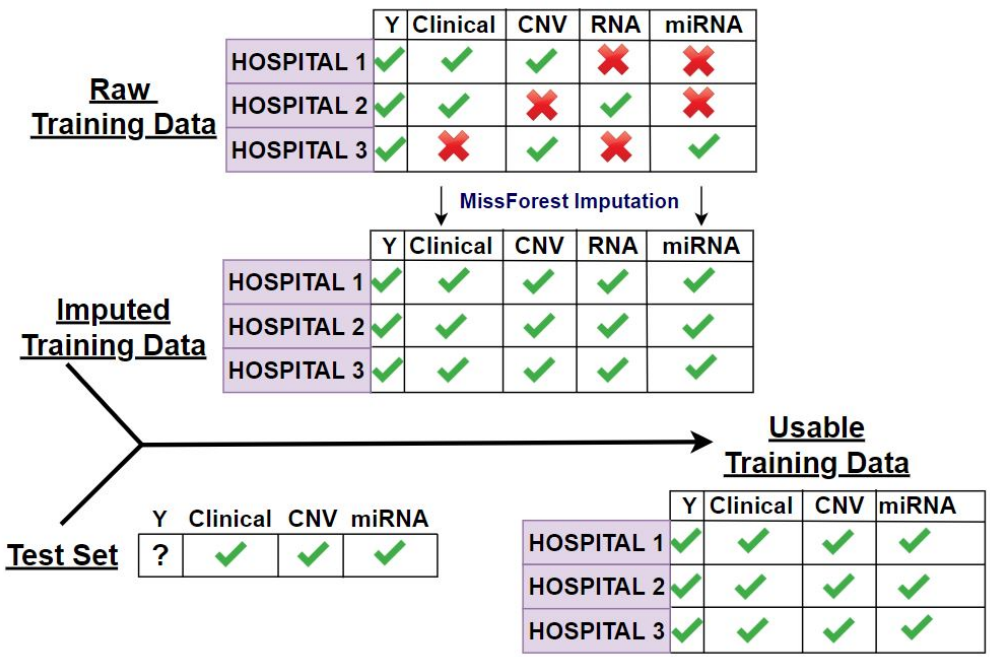

Figure: 'Imputation' approach to deal with block-wise missingness

Block-Wise Approach

Idea: Fit a separate RF on each feature-block of the train data. To predict on the test data, the predictions of the block-wise fitted models are aggregated.

1. Fit a separate RF on the observed parts of each feature-block in the training data
2. To predict on the test data, each block-wise RF is asked for a prediction - only those RFs that were fitted on feature-block that is available for the test-set can do so
3. These block-wise predictions can be aggregated in a (un)weighted way

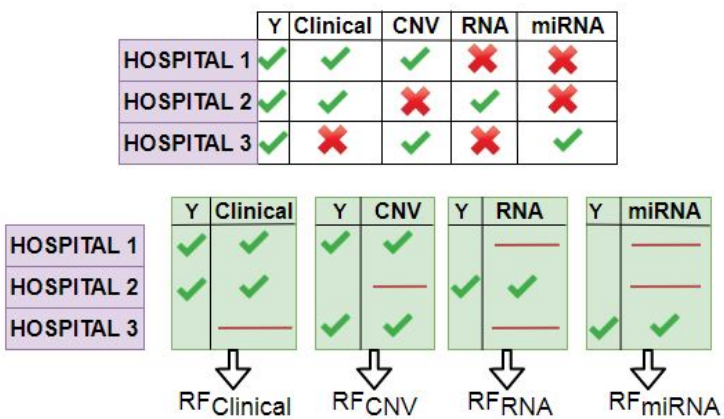

Figure: Training of random forest models with the 'Block-Wise' approach.

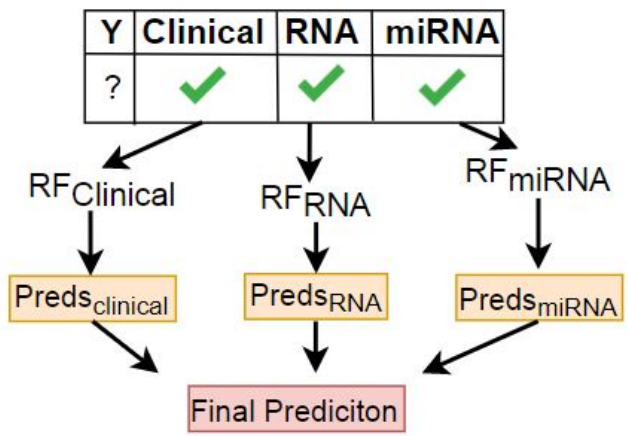

Figure: Prediction on test data with the 'Block-Wise' approach.

Fold-Wise Approach

Idea: Fit a separate RF on the of each fold of the train data. To predict on the test data, the predictions of the fold-wise fitted models are aggregated.

1. Fit a separate RF on the observed parts of each fold in the training data
2. To predict on the test data, each fold-wise RF is asked for a prediction - only those RFs that were fitted on a fold that contains at least one of the blocks from the test data can do so

   The single decision trees of a RF might be pruned for this
3. The fold-wise predictions can be aggregated in a (un)weighted way

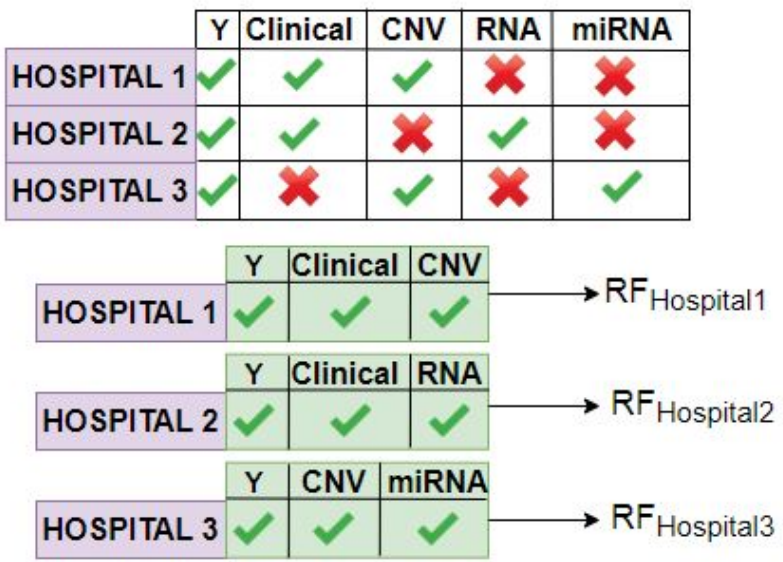

Figure: Training of random forest models with the 'Fold-Wise' approach.

Fold-Wise Approach

Idea: Fit a separate RF on the of each fold of the train data. To predict on the test data, the predictions of the fold-wise fitted models are aggregated.

1. Fit a separate RF on the observed parts of each fold in the training data
2. To predict on the test data, each fold-wise RF is asked for a prediction - only those RFs that were fitted on a fold that contains at least one of the blocks from the test data can do so
   The single decision trees of a RF might be pruned for this
3. The fold-wise predictions can be aggregated in a (un)weighted way

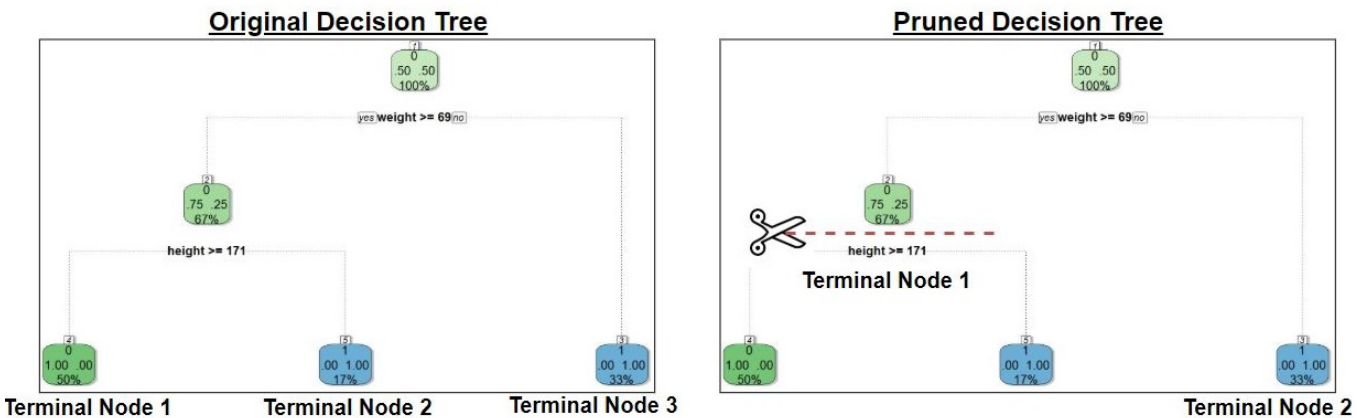

Figure: The pruning of a single decision tree.

Fold-Wise Approach

Idea: Fit a separate RF on the of each fold of the train data. To predict on the test data, the predictions of the fold-wise fitted models are aggregated.

1. Fit a separate RF on the observed parts of each fold in the training data
2. To predict on the test data, each fold-wise RF is asked for a prediction - only those RFs that were fitted on a fold that contains at least one of the blocks from the test data can do so
   The single decision trees of a RF might be pruned for this
3. The fold-wise predictions can be aggregated in a (un)weighted way

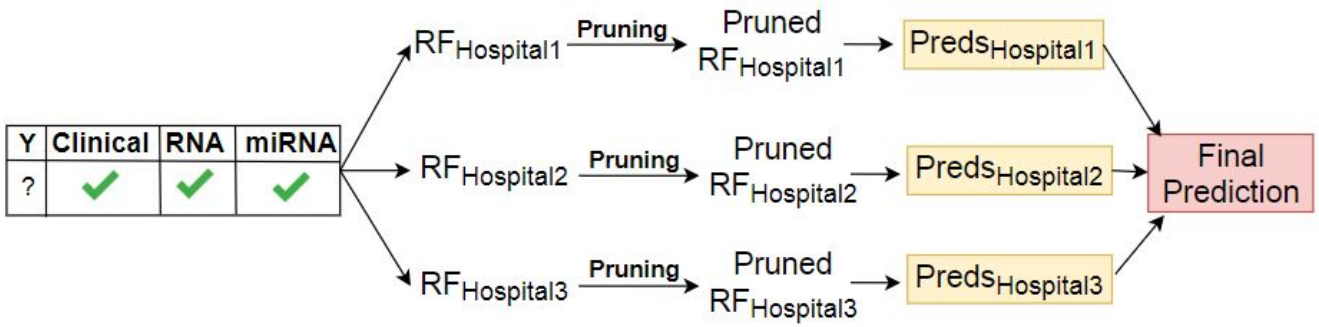

Figure: The prediction on test data with the 'Fold-Wise' approach.

Metrics

Used to "evaluate the performance of a statistical learning method [...] - measure how well its predictions actually match the observed data" [[4], p. 29].

'F-1 Score' represents the harmonic mean of the precision and recall

$$\textbf{F-1 Score} = 2 * \frac{(\frac{\textbf{TP}}{\textbf{TP+FP}}) * (\frac{\textbf{TP}}{\textbf{TP+FN}})}{(\frac{\textbf{TP}}{\textbf{TP+FP}}) + (\frac{\textbf{TP}}{\textbf{TP+FN}})} \tag{1}$$

'Balanced Accuracy' represents the average of the class-wise accuracies

$$\textbf{Balanced Accuracy} = \frac{\frac{\textbf{TP}}{\textbf{TP+FN}} + \frac{\textbf{TN}}{\textbf{TN+FP}}}{2} \tag{2}$$

'Matthews correlation coefficient' can be seen as a "discretization of the Pearson correlation for binary variables" [[5], p. 1]

$$\textbf{MMC} = \frac{\textbf{TP} * \textbf{TN} - \textbf{FP} * \textbf{FN}}{\sqrt{(\textbf{TP} + \textbf{FP}) * (\textbf{TP} + \textbf{FN}) * (\textbf{TN} + \textbf{FP}) * (\textbf{TN} + \textbf{FN})}} \tag{3}$$

TCGA data

- Received 21 processed data sets from Dr. Hornung
- The covariate 'gender' is used as a binary response variable
  - Even though it "is not a clinically meaningful outcome[...], it features major advantages for a purely methodological investigation" [[6], p. 5]
  - 7 data sets do not contain a 'gender' covariate and had to be removed
    $\rightarrow$ A total of 14 data sets remained.
- Each of the 14 DFs contain the same five feature-blocks, whereby the dimensions were reduced to reduce the computational effort

Feature-Block	$\varnothing$ **covariates in original**	$\varnothing$ **covariates in reduced**
Clinical	3.5	3.5
miRNA	770	385
Mutation	16218	1616
CNV	57964	2898
RNA	23559	3555

Table: Average amount of covariates in each block over the 14 TCGA data sets.

TCGA data

TCGA data sets are fully observed!!

→ Induce block-wise missingness so they can be used for the investigation of the approaches that can deal with block-wise missingness!

	Y	Clinical	CNV	RNA	Mutation	miRNA
Fold 1	✓	✓	✓	✗	✗	✗
Fold 2	✓	✓	✗	✓	✗	✗
Fold 3	✓	✓	✗	✗	✓	✗
Fold 4	✓	✓	✗	✗	✗	✓

	Y	Clinical	miRNA	RNA	Mutation	CNV
Fold 1	✓	✓	✓	✓	✓	✓
Fold 2	✓	✓	✗	✓	✓	✓
Fold 3	✓	✓	✗	✗	✓	✓
Fold 4	✓	✓	✗	✗	✗	✓

	Y	Clinical	CNV	RNA	Mutation	miRNA
Fold 1	✓	✓	✓	✓	✗	✓
Fold 2	✓	✓	✓	✓	✗	✗
Fold 3	✓	✓	✗	✓	✓	✓
Fold 4	✓	✓	✗	✓	✓	✓

	Y	Clinical	RNA & miRNA	Mutation & CNV
Fold 1	✓	✓	✓	✗
Fold 2	✓	✓	✗	✓

Table: The four various block-wise missingness patterns for the TCGA data sets.

TCGA data

Algorithm 1: Evaluation of the approaches with the TCGA data

Input : D ← TCGA data set with n observations & p features
App ← Approach
$Patt$ ← Pattern of block-wise missingness

1. Split the fully observed data set D into five equally sized folds

2. **for** $k \leftarrow 1$ **to** 5 **do**

- **2.1** Use fold k as test-set and the remaining four folds as train-set;
- **2.2** Induce the block-wise missingness according to $Patt$ into the train-set;
- **2.3** Evaluate the predictive performance of the approach App for different combinations of observed feature-blocks in the test-set;
  - **2.3.1** Evaluation on the fully observed test-set;
  - **2.3.2** Evaluation on test-sets with one missing feature-block;
  - **2.3.3** Evaluation on test-sets with two missing feature-blocks;
  - **2.3.4** Evaluation on test-sets with three missing feature-blocks;
  - **2.3.5** Evaluation on test-sets with a single feature-block;

Clinical asthma data

- The 'clinical asthma data' is a real-world data set with block-wise missingness
- It was provided by the group of Prof. Dr. med. Bianca Schaub at the 'paediatric clinic Dr. von Haunersches Kinderspital'
- Has a binary target variable that is defined as the presence of asthma *(521 observations: 265 with a negative & 256 with a positive response)*
- Consists of six different blocks, whereby the amount of observations in the blocks inversely reflects the effort of generating the data

ID	Feature-Block	Number of observations	Number of variables
1	Questionaire	521	44
2	Clinical routine diagnostics	385	16
3	Allergen sensitization	472	19
4	Cytokine expression data	149	29
5	Gene expression data I	66	82
6	Gene expression data II	46	84

Table: The number of observations and variables for the different blocks in the clinical asthma data.

Clinical asthma data

Algorithm 2: Evaluation of the approaches with the clinical asthma data

Input : D $\leftarrow$ clinical asthma data with block-wise missingness
$App \leftarrow$ Approach

1. Split the clinical asthma data D into five equally sized folds

2. for $k \leftarrow 1$ **to** 5 **do**

- **2.1** Use fold k as test-set and the remaining four folds as train-set;
- **2.2** Extract the unique patterns of block-wise missingness in the test-set - $test_{patterns}$;
- **2.3 for** $pattern_{current} \in test_{patterns}$ **do**
  - **2.3.1** Fit App on the train-set, such that it can be used for the prediction on $pattern_{current}$;
  - **2.3.2** Generate predictions for the test-observations with block-wise missingness according to $current_{pattern}$
- **2.4** Evaluate the predictive performance of the approach App by comparing the predicted classes with the true classes;

TCGA data - Comparison - Pattern 1

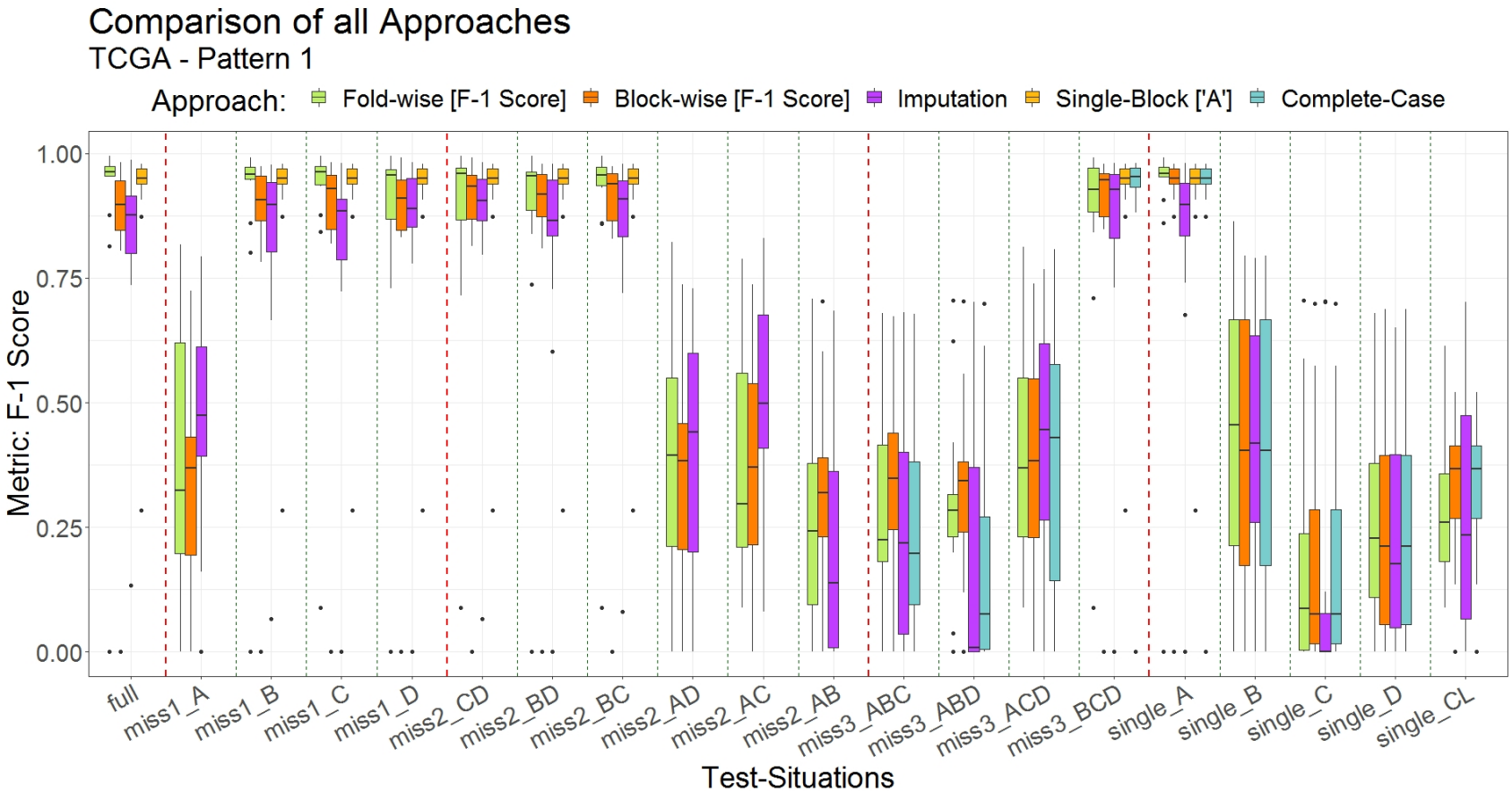

Figure: Comparison of the different approaches on the TCGA data with induced block-wise missingness according to pattern 1.

→ Fold-Wise approach is the best in 11 test-situations
→ Imputation approach is the best in 4 test-situations
→ Block-Wise approach is the best in 3 test-situations

TCGA data - Comparison - Pattern 2

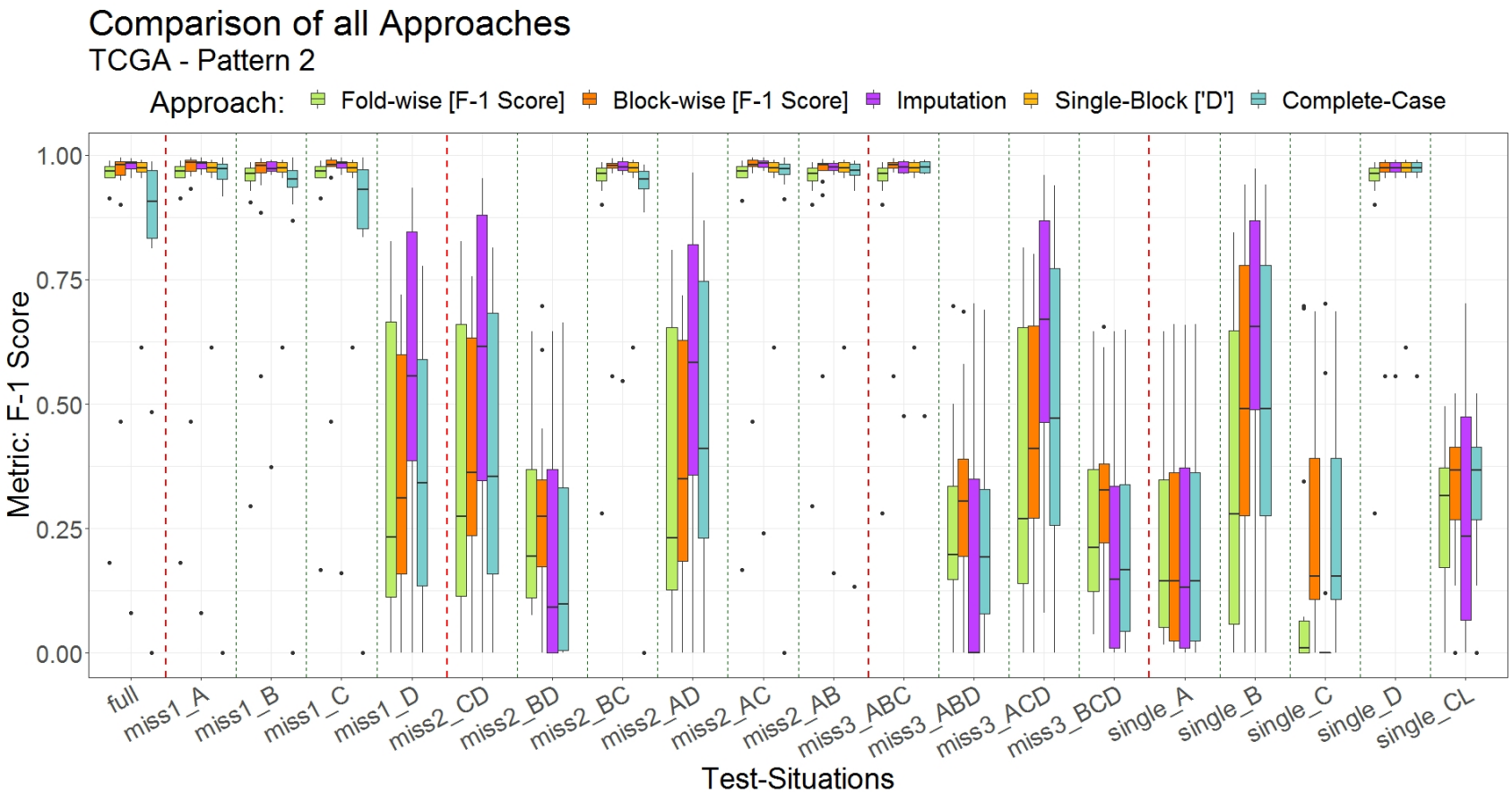

Figure: Comparison of the different approaches on the TCGA data with induced block-wise missingness according to pattern 2.

→ Imputation approach is the best in 8 test-situations
→ Block-Wise approach is the best in 8 test-situations

TCGA data - Comparison - Pattern 3

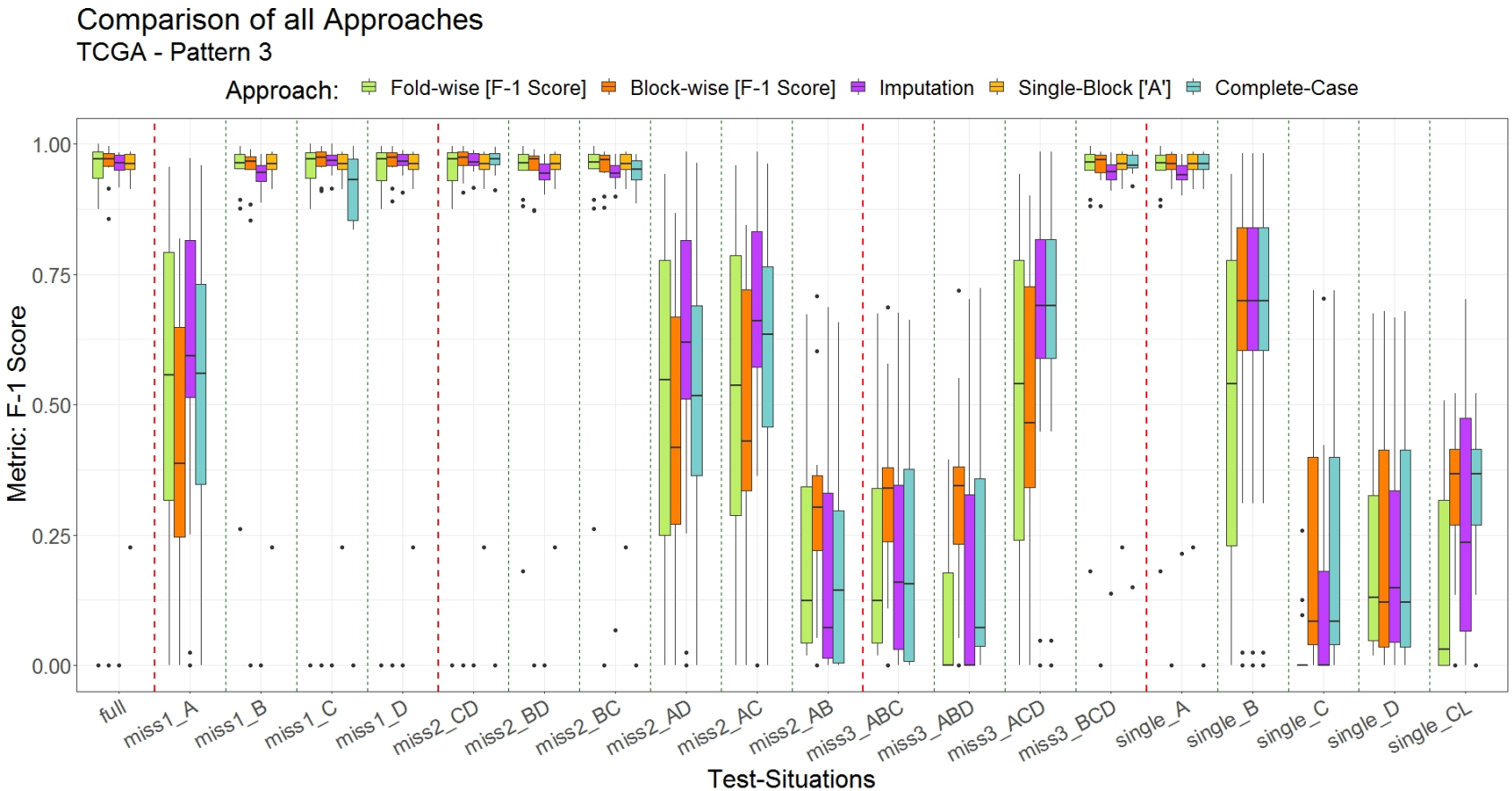

Figure: Comparison of the different approaches on the TCGA data with induced block-wise missingness according to pattern 3.

→ Block-Wise approach is the best in 10 test-situations
→ Imputation approach is the best in 4 test-situations
→ Fold-Wise approach is the best in 2 test-situations

TCGA data - Comparison - Pattern 4

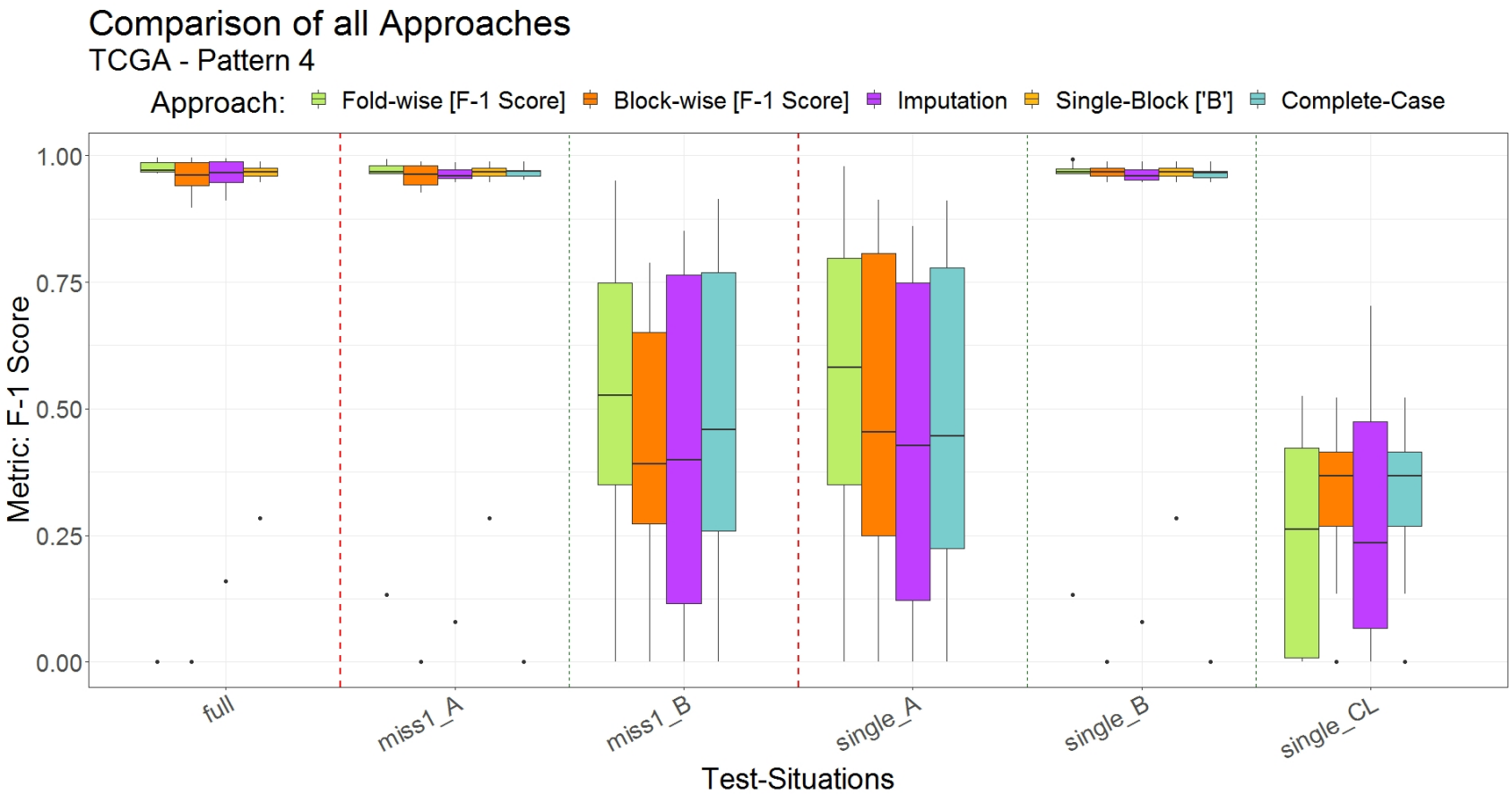

Figure: Comparison of the different approaches on the TCGA data with induced block-wise missingness according to pattern 4.

→ Fold-Wise approach is the best in 4 test-situations
→ Complete-Case approach in just a single case

Clinical asthma data - Hagenberg's Methods

[1] The thesis from Hagenberg [7] introduced two adaptions of the priority-Lasso, such that it can directly deal with block-wise missingness.

- 'PL - ignore, intercept'
- 'PL - ignore, zero'
- 'PL - impute, max. n'
- 'PL - impute, max. blocks'

The diverse priority-Lasso approaches were evaluated with different priorities for the single feature-blocks in the training data and different subsets of blocks from test data used for the prediction

[2] Additionally, Hagenberg has used the mdd-sPLS method [8] as a reference approach

Clinical asthma data - Results

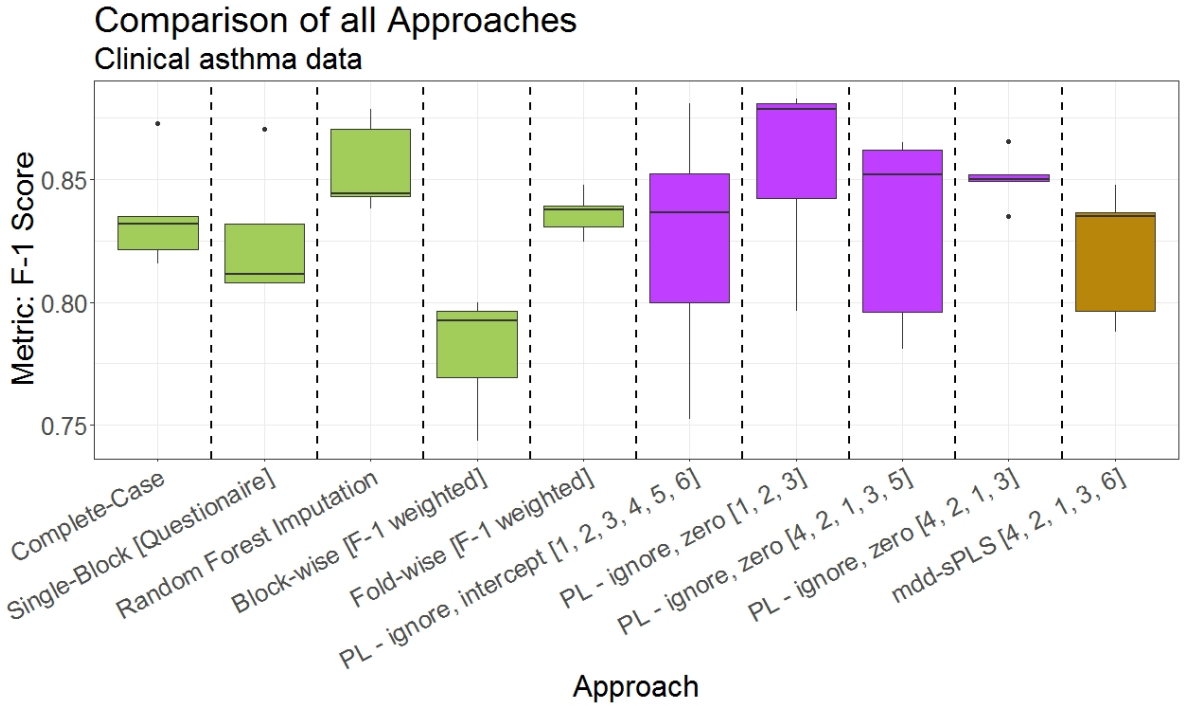

Figure: Comparison of the random forest adaptions, the priority-Lasso adaptions and the mdd-sPLS method on the clinical asthma data.

$\rightarrow$ Block-Wise $<$ Single-Block $<$ Complete-Case $<$ mdd-sPLS $<$ 'PL - ignore, intercept [1, 2, 3, 4, 5, 6]' $<$ Fold-Wise $<$ Imputation $<$ 'PL - ignore, zero [4, 2, 1, 3]' $<$ 'PL - ignore, zero [4, 2, 1, 3, 5]' $<$ 'PL - ignore, zero [1, 2, 3]'

Discussion and Conclusion

Random Forest based approaches:

- Single-Block & Complete-Case are the simplest adaptions
  $\rightarrow$ worst predictive performance among the RF approaches
- Block- & Fold-Wise approach have the best predictive performance with the 'F-1 Score' as weight metric
- Performance of the Block- & Fold-Wise approach is opposing
  $\rightarrow$ in settings, where one performs good, the other performs bad
- Performance of the Imputation approach is comparable to the performance of the Block- & Fold-Wise approach
- The Imputation & Fold-Wise approach are extremely slow compared to the other RF approaches

Discussion and Conclusion

Hagenberg's approaches:

- The PL adaptions were evaluated with different block priorities and various subset of blocks used for the prediction
  $\rightarrow$ worst performance with 'naive' block priority & all feature-blocks for the prediction
- 'PL-impute' always outperformed by 'PL-ignore'
- 'mdd-SPLS' always outperformed by 'PL-ignore'
- 'PL-ignore' are the best among Hagenberg's approaches
  $\rightarrow$ no big difference between 'PL-ignore, intercept' & 'PL-ignore, zero'

Discussion and Conclusion

Comparison of the approaches:

- Comparison of the approaches based on the clinical asthma data
- Imputation & Fold-Wise are the best RF approaches on the clinical asthma data → outperform mdd-sPLS & 'PL-ignore, intercept' with the naive block-priority
- 'PL-ignore' with other block priorities and/ or only a subset of blocks for the prediction outperform the Imputation & Fold-Wise approach
- A 'selection bias' might have been introduced, as much more variants of the priority-Lasso have been tried out than with the RF approaches

⇒ **Choice of the approach depends on the prior knowledge of the data!**

- User knows which feature-blocks might be more important than others
  → 'ignore' priority-Lasso adaptions can be recommended
- User does not know which feature-blocks might be more important than others
  → random forest approaches seem to be the better choice
  (especially Fold- & Block-Wise, as these use an OOB metric to estimate the importance of the diverse folds/ feature-blocks)

Discussion and Conclusion

Outlook:

- Comparison of mdd-sPLS, PL & RF approaches on further data-sets for a better generalisation of the results
- Implementation of the Fold-Wise approach directly in 'C' / 'Java' for speed improvement
- Combining the idea of the 'Block Forest' article [9] with the Fold-Wise approach
  $\rightarrow$ Incorporation of the block structure for the split point selection

References I


Roman Hornung et al. "Random forests for multiple training data sets with varying covariate sets". manuscript - unpublished yet. - in prep.


Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.


Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.


Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.


Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric". In: *PloS one* 12.6 (2017).


Nicole Schüller et al. "Improved outcome prediction across data sources through robust parameter tuning". In: (2019).

References II


Jonas Hagenberg. "Penalized regression approaches for prognostic modelling using multi-omics data with block-wise missing values". manuscript - unpublished yet. - in prep.


Hadrien Lorenzo, Jérôme Saracco, and Rodolphe Thiébaut. "Supervised Learning for Multi-Block Incomplete Data". In: *arXiv preprint arXiv:1901.04380* (2019).


Roman Hornung and Marvin N Wright. "Block Forests: random forests for blocks of clinical and omics covariate data". In: *BMC bioinformatics* 20.1 (2019), p. 358.

Attachment I

There are two main reasons for block-wise missingness in 'Multi-Omics' data:

1. Collection of omics data is more expensive and complex as for regular clinical data
   $\rightarrow$ can not always be collected for all participants of a study
2. Collection of data sets from different sources - e.g. various hospitals

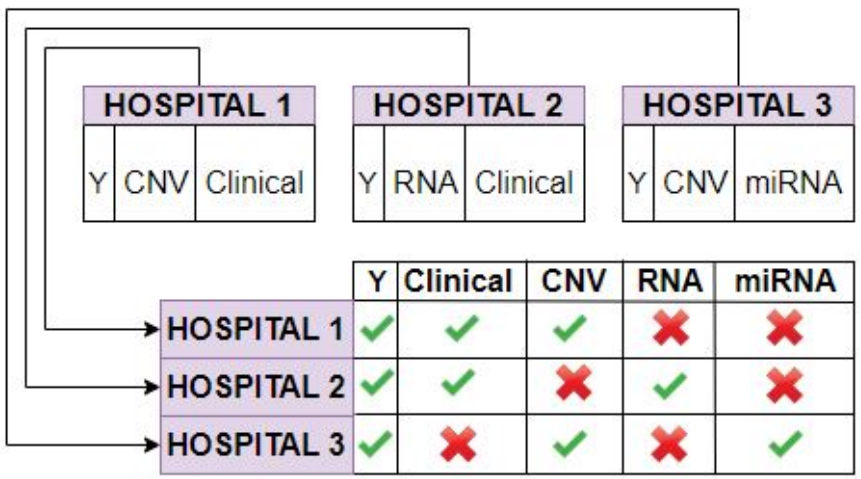

Figure: Block-wise missingness, when concatenating data from diverse sources.

Attachment II

Algorithm 3: Growing a random forest model

Input : D $\leftarrow$ data with n observations & p features
M $\leftarrow$ number of trees in the forest
n_{min} $\leftarrow$ 'MinSplit' argument of a decision tree
mtry $\leftarrow$ number of variables to draw at each split

for $m \leftarrow 1$ **to** M **do**

1. Draw a bootstrap sample **Z*** of size 'n' from 'D';

2. Based on **Z*** grow a decision tree, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached;

2.1 Randomly draw 'mtry' of the 'p' available variables;

2.2 Pick the best splitting point among the 'mtry' variables;

2.3 Split the node into two daughter nodes;

Attachment III

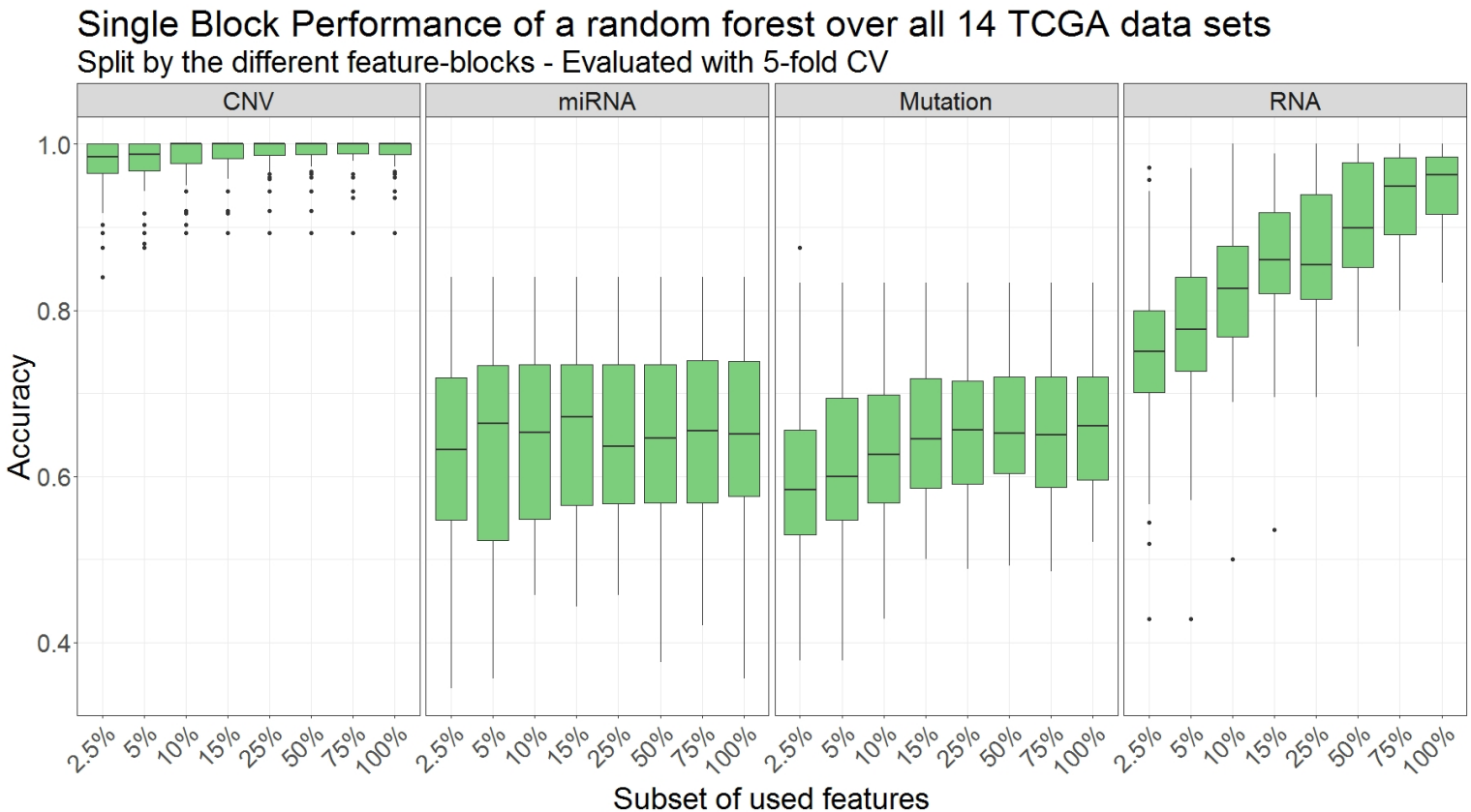

Figure: The accuracy of a random forest model evaluated on the single omics feature-blocks for a range of possible subsets on all 14 TCGA data sets.

Attachment IV

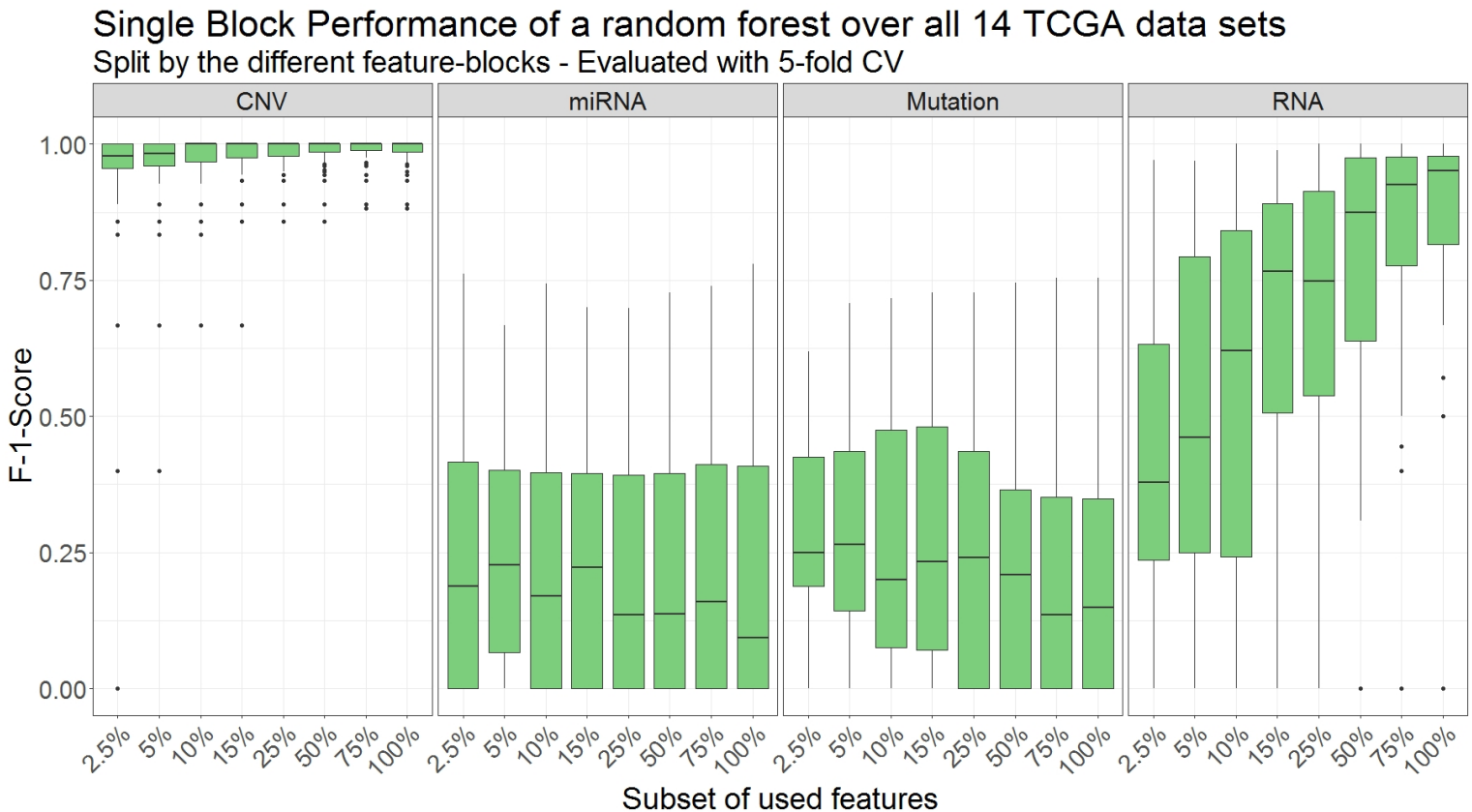

Figure: The F-1 score of a random forest model evaluated on the single omics feature-blocks for a range of possible subsets on all 14 TCGA data sets.

Attachment V

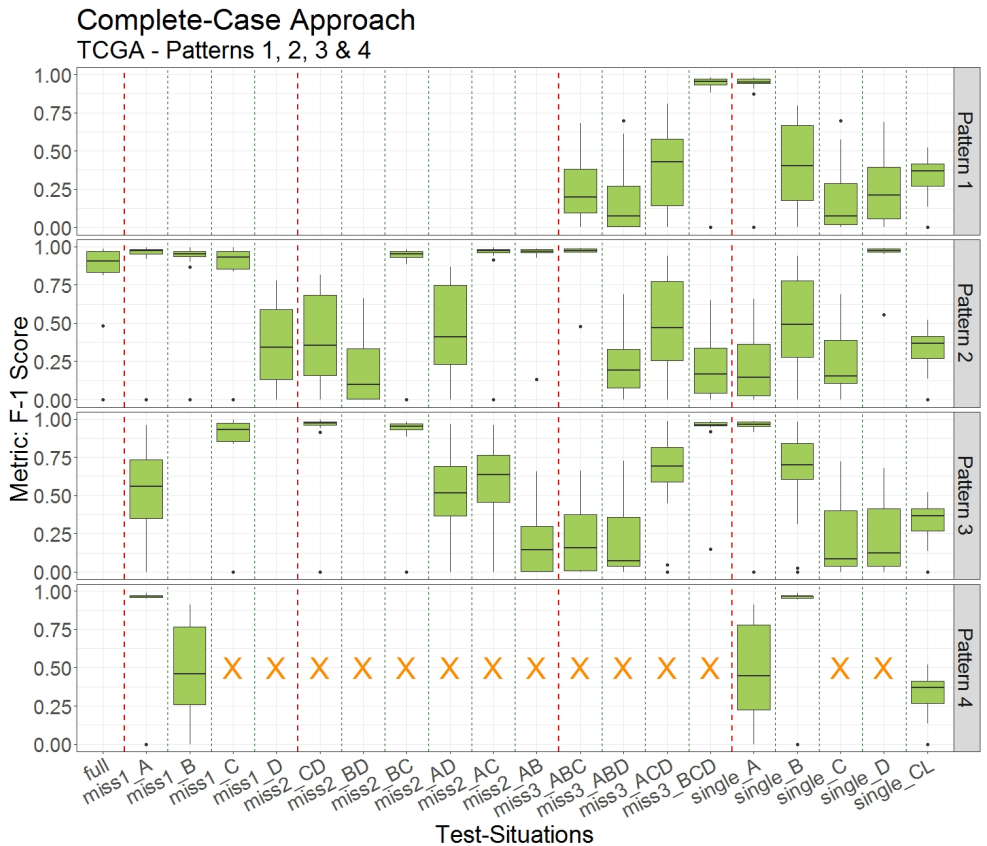

Figure: Results of the 'Complete-Case' approach on the TCGA data with induced block-wise missingness according to pattern 1, 2, 3 & 4. Test-Situations that are not available for 'Pattern 4' are marked with an orange cross.

Attachment VI

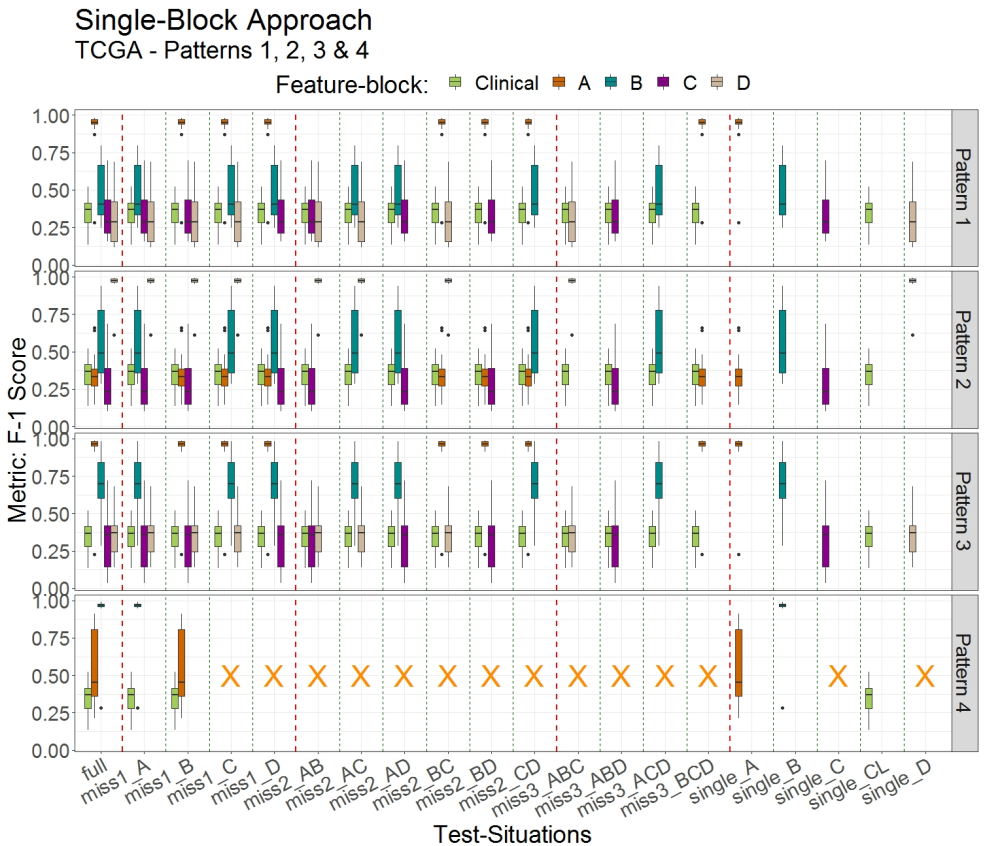

Figure: Results of the 'Single-Block' approach on the TCGA data with induced block-wise missingness according to pattern 1, 2, 3 & 4. Test-Situations that are not available for 'Pattern 4' are marked with an orange cross.

Attachment VII

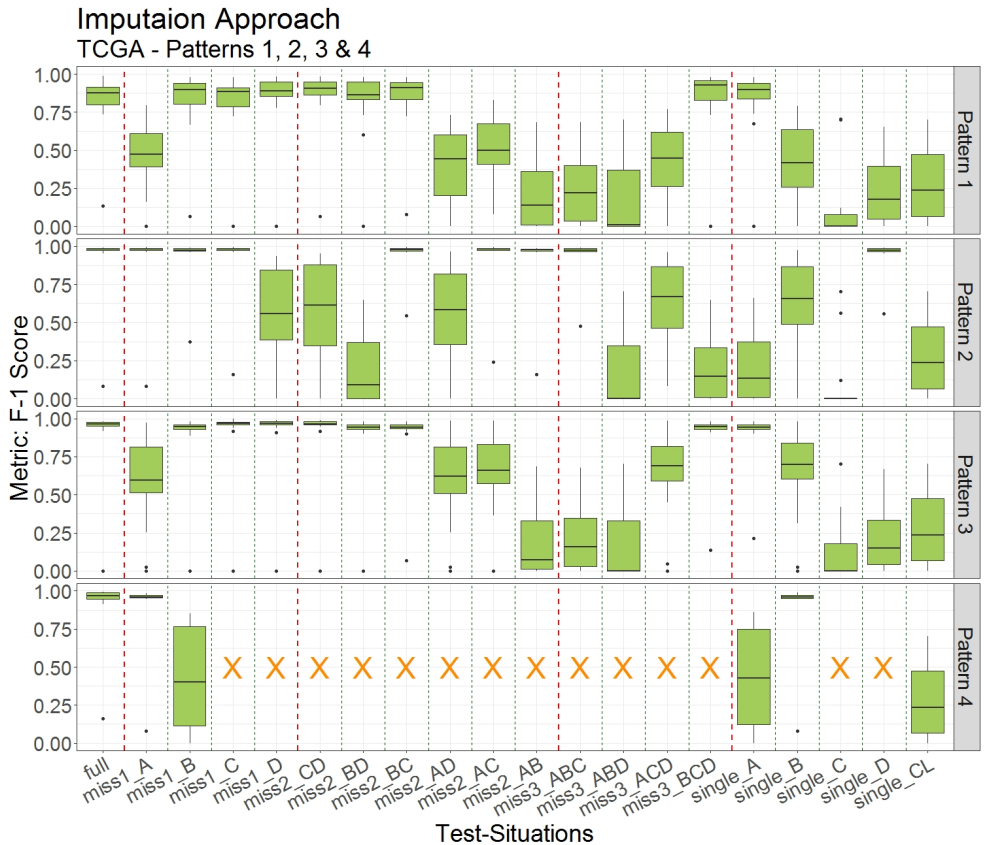

Figure: Results of the 'Imputation' approach on the TCGA data with induced block-wise missingness according to pattern 1, 2, 3 & 4. Test-Situations that are not available for 'Pattern 4' are marked with an orange cross.

Attachment VIII

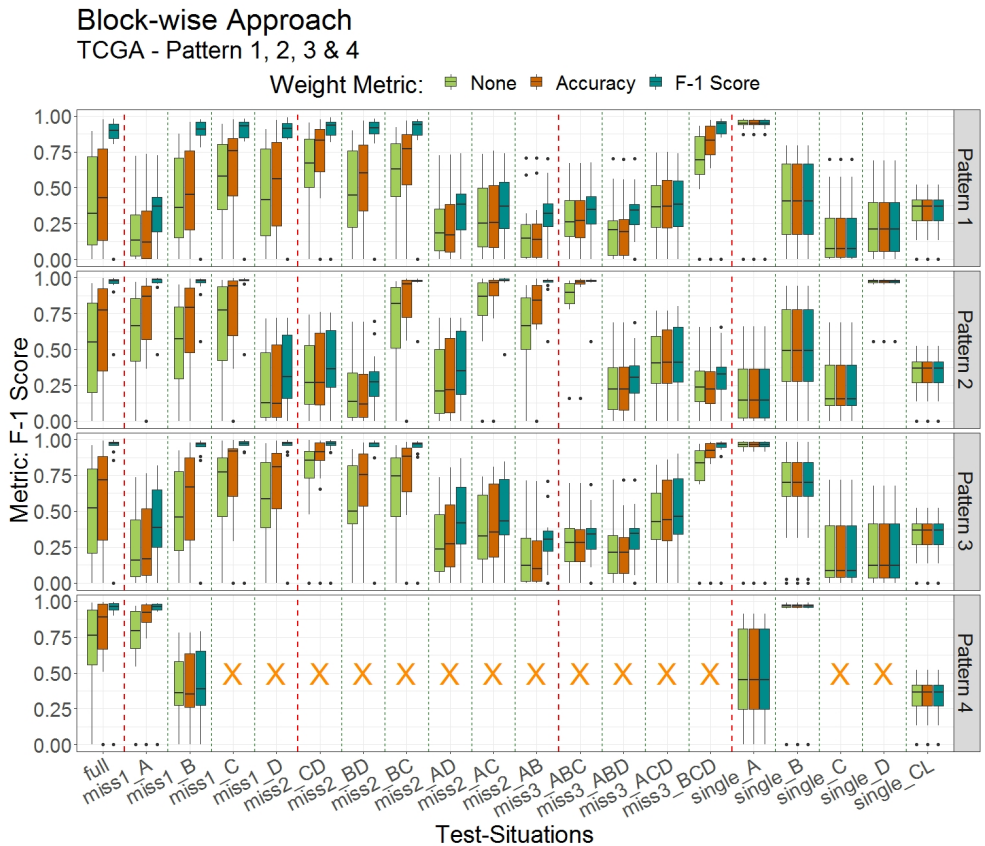

Figure: Results of the 'Block-Wise' approach on the TCGA data with induced block-wise missingness according to pattern 1, 2, 3 & 4. Test-Situations that are not available for 'Pattern 4' are marked with an orange cross.

Attachment IX

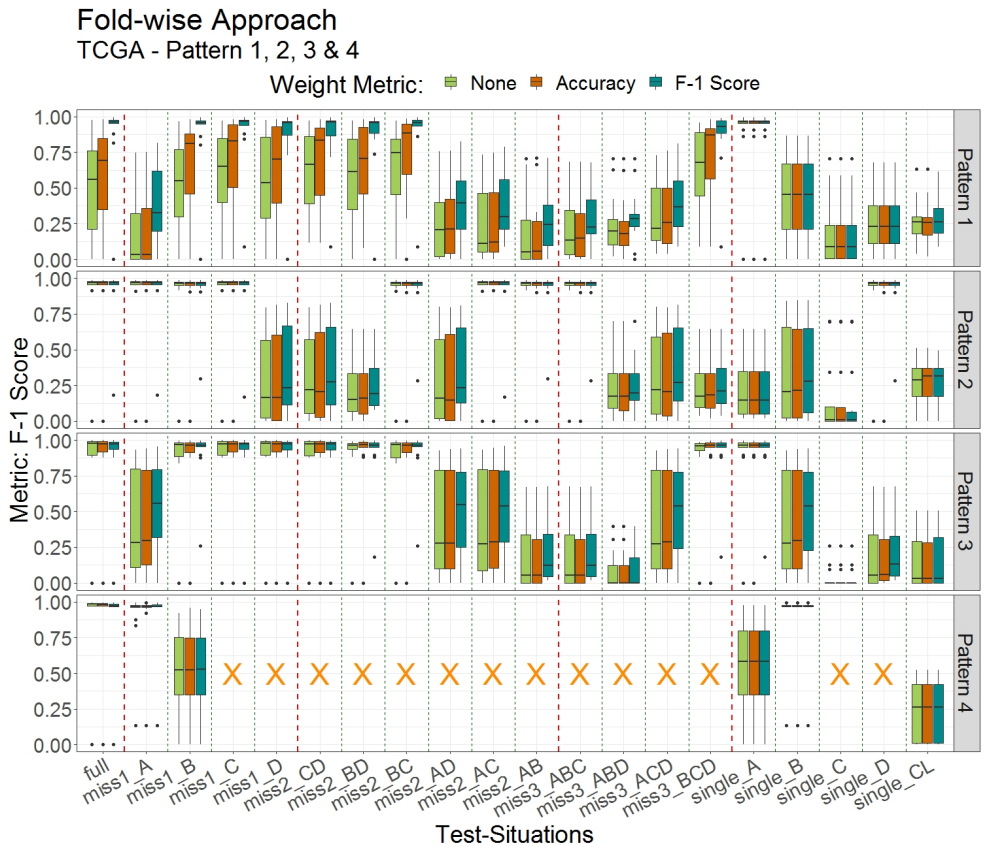

Figure: Results of the 'Fold-Wise' approach on the TCGA data with induced block-wise missingness according to pattern 1, 2, 3 & 4. Test-Situations that are not available for 'Pattern 4' are marked with an orange cross.

Attachment X

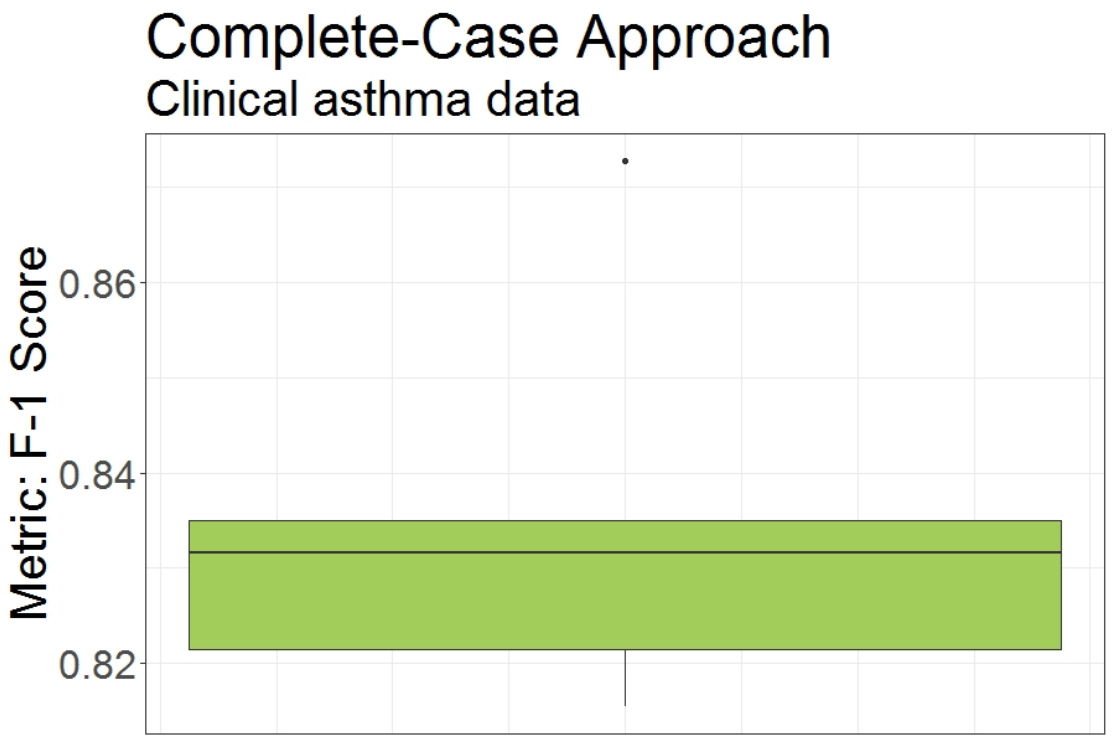

Figure: Results of the 'Complete-Case' approach on the clinical asthma data

Attachment XI

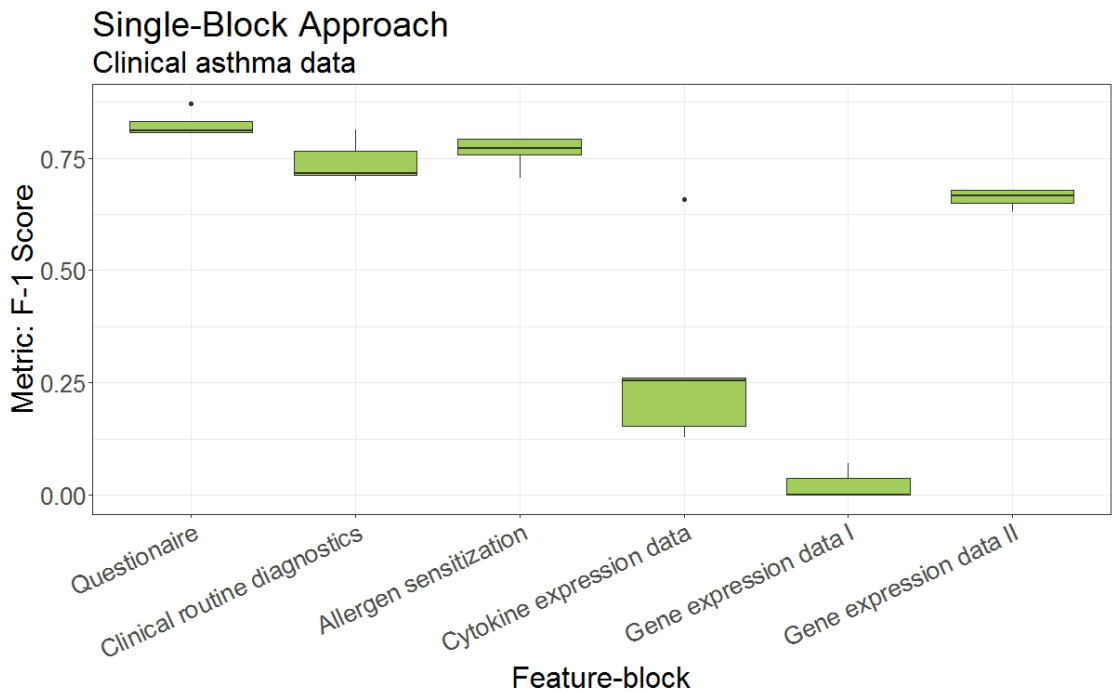

Figure: Results of the 'Single-Block' approach on the clinical asthma data

Attachment XII

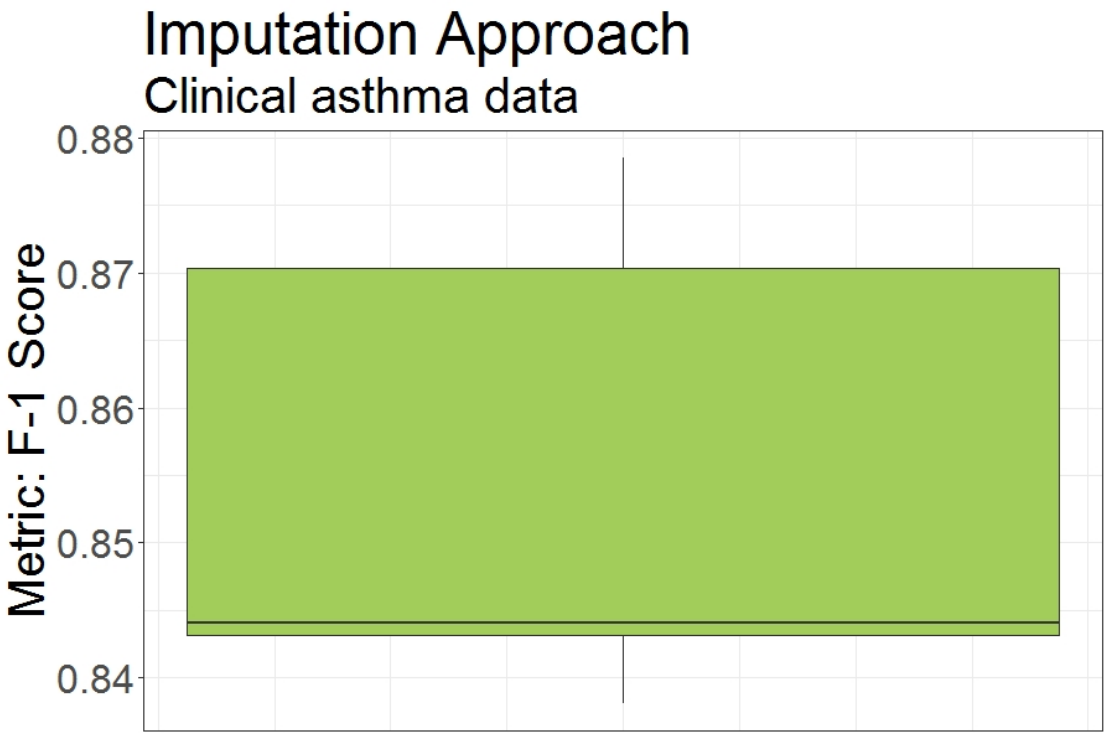

Figure: Results of the 'Imputation' approach on the clinical asthma data

Attachment XIII

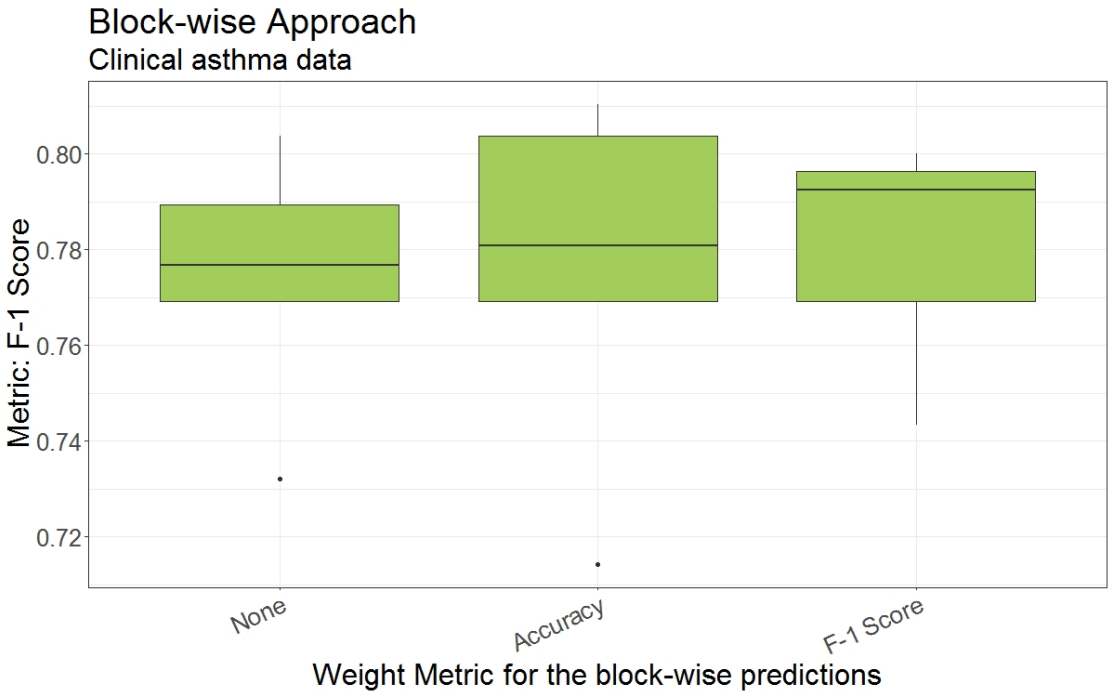

Figure: Results of the 'Block-Wise' approach on the clinical asthma data

Attachment XIV

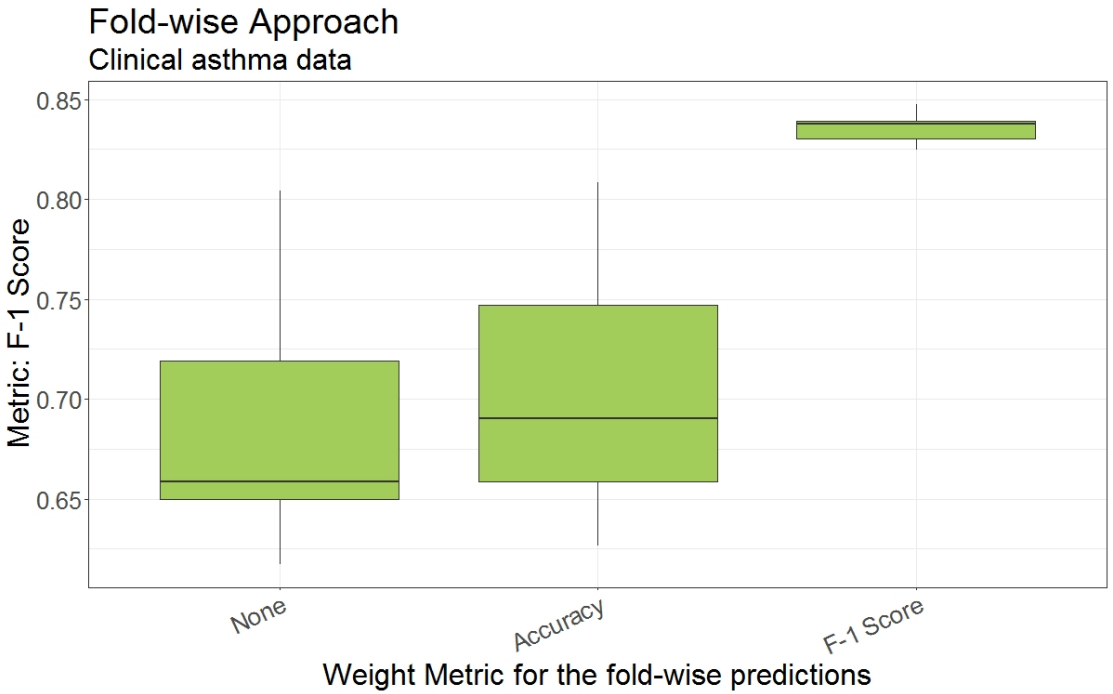

Figure: Results of the 'Fold-Wise' approach on the clinical asthma data

Attachment XV

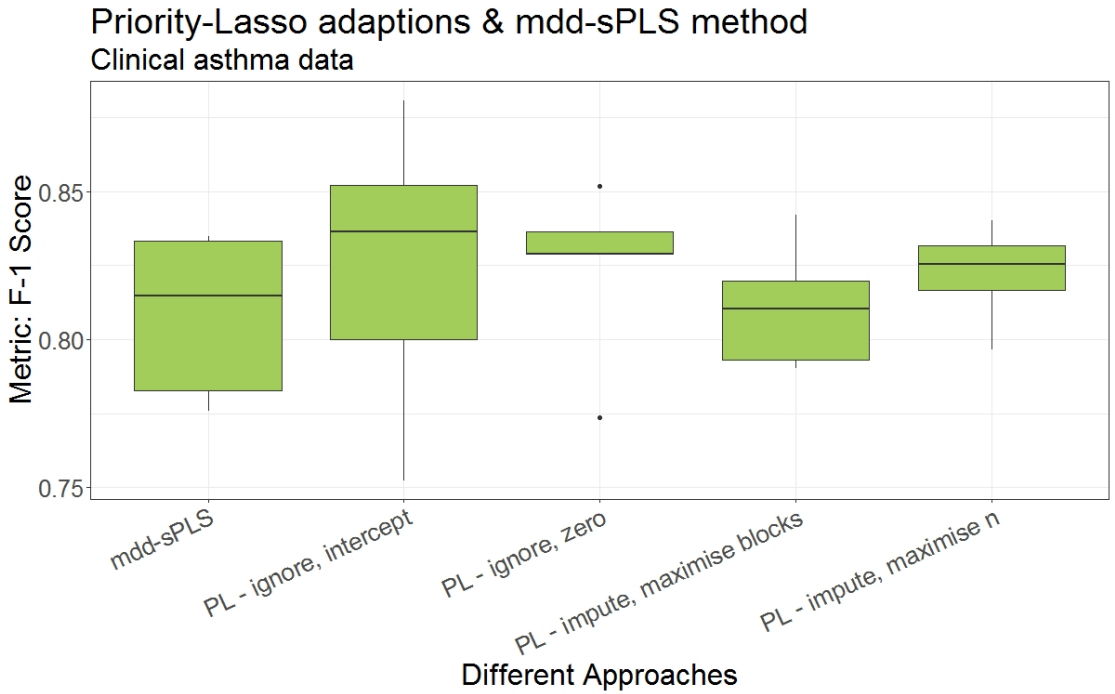

Figure: Results of the priority-Lasso adaptions and mdd-sPLS with the naive block order on the clinical asthma data

Attachment XVI

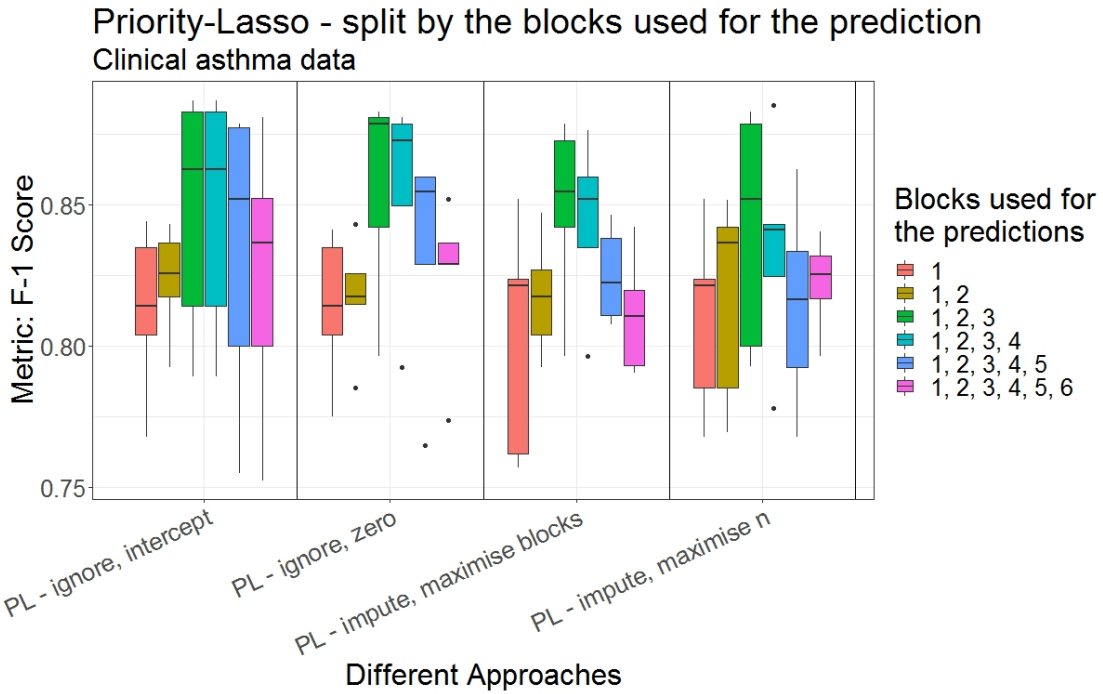

Figure: Results of the priority-Lasso adaptions with the naive block order on the clinical asthma data - using only a subset of the blocks for the prediction

Attachment XVII

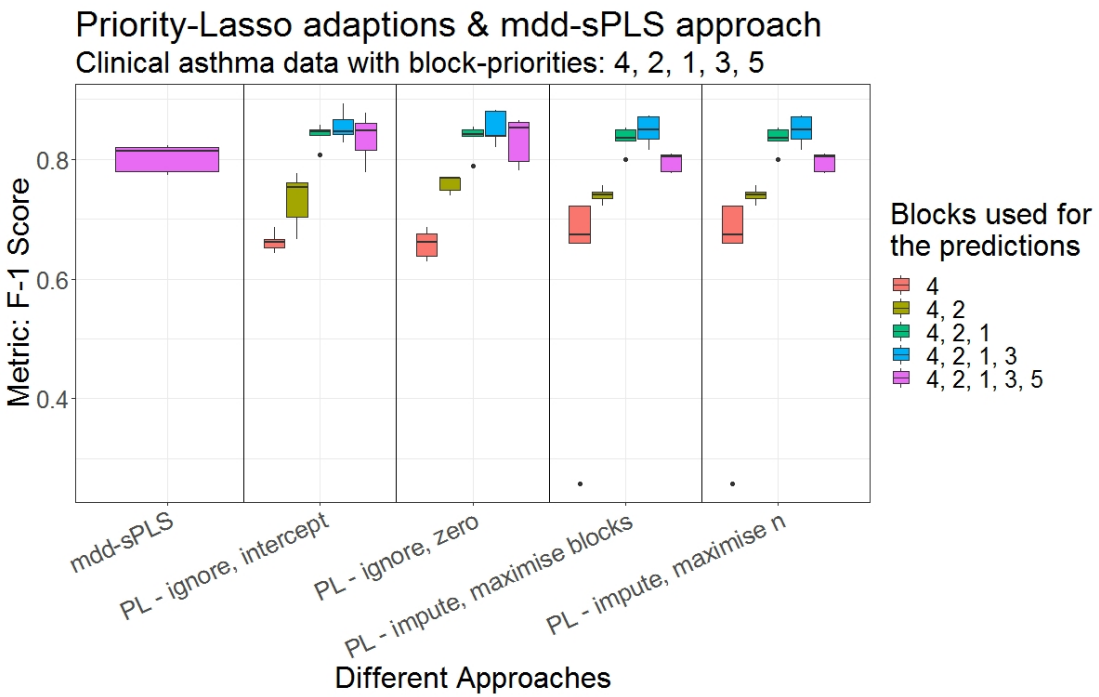

Figure: Results of the priority-Lasso adaptions with adjusted block order (4, 2, 3, 1, 5) on the clinical asthma data - using only a subset of the blocks for the prediction - and the mdd-sPLS method

Attachment XVIII

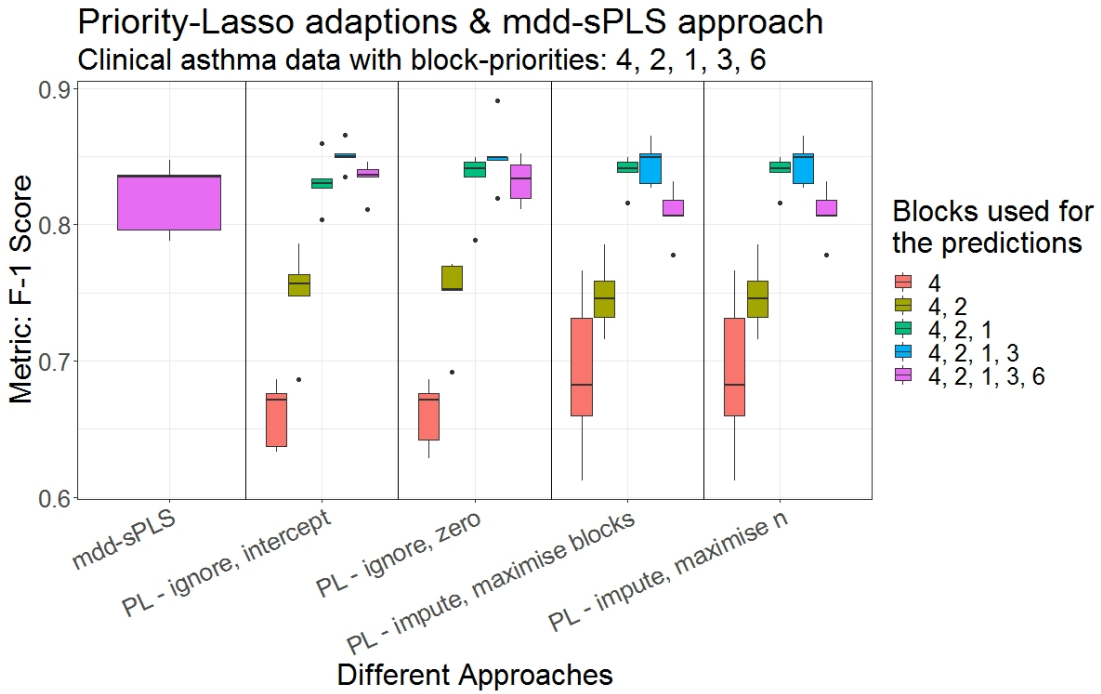

Figure: Results of the priority-Lasso adaptions with adjusted block order (4, 2, 3, 1, 6) on the clinical asthma data - using only a subset of the blocks for the prediction - and the mdd-sPLS method