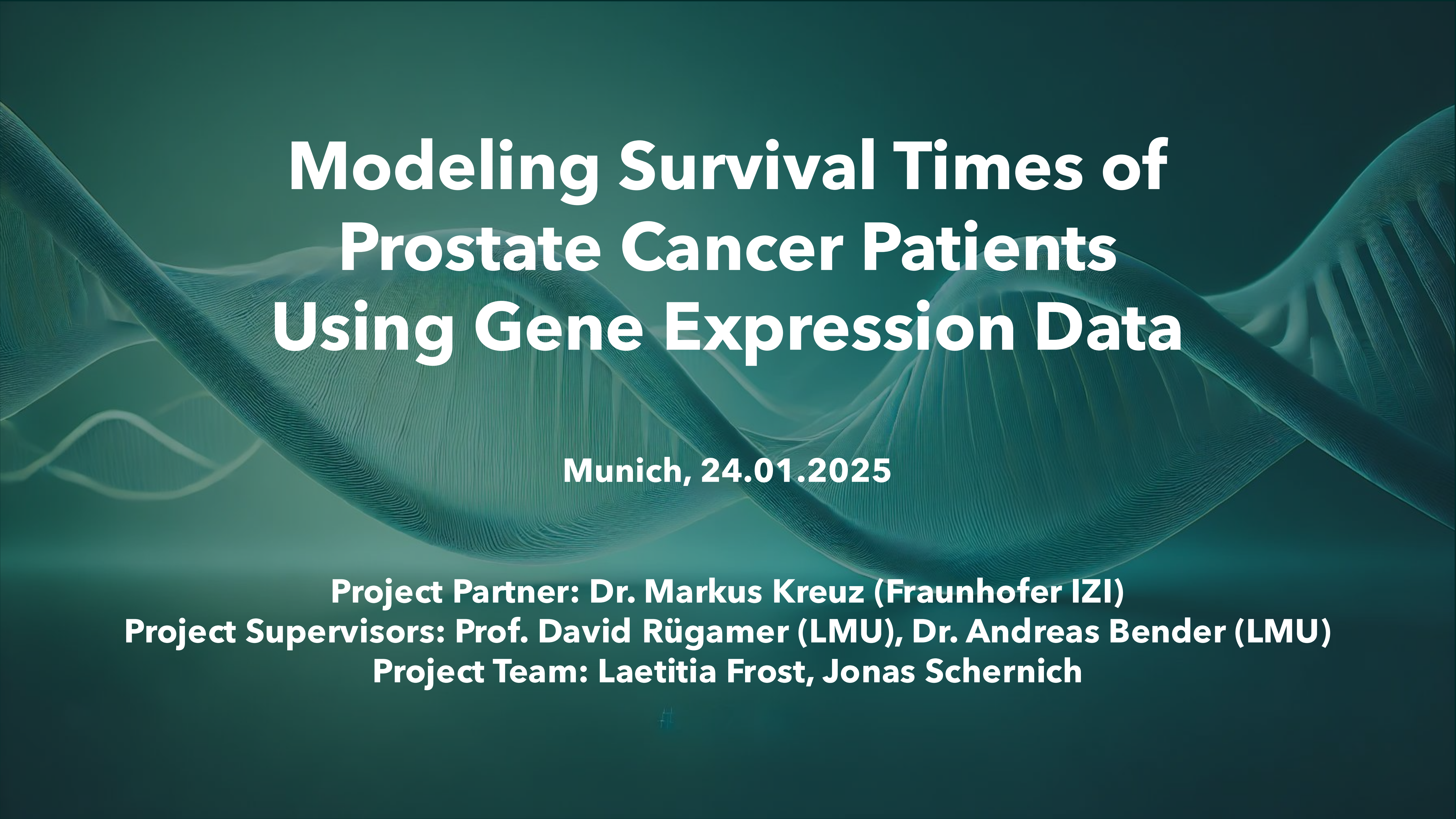


# Modeling Survival Times of Prostate Cancer Patients Using Gene Expression Data

Munich, 24.01.2025

Project Partner: Dr. Markus Kreuz (Fraunhofer IZI)
Project Supervisors: Prof. David Rügamer (LMU), Dr. Andreas Bender (LMU)
Project Team: Laetitia Frost, Jonas Schernich

# Intro

- Prostate cancer (PCa) as most prevalent solid cancer among men in Western Countries
- Highly variable clinical manifestation, resulting in very different optimal types of treatment e.g.:
    - Aggressive types of PCa: Radical prostatectomy (RP)
    - Low risk types of PCa: Active surveillance

→ Precise risk stratification wrt. to a patient's disease outcome vital for adequate treatment decisions

- Approaches for diagnosis and risk stratification i.a. include:
    - Blood markers e.g. prostate-specific antigen levels (PSA)
    - Biopsy of prostate tissue:
        - Gleason Score a metric for aggressiveness of cancer
        - Biomarkers e.g via RNA-sequencing

→ Ongoing research on the inclusion of biomarkers in the prediction of PCa outcomes

# Existing research I

ProstaTrend [1]: RNA expression-based score for the prediction of PCa prognosis

- Fresh-frozen tissue specimens from 4 cohorts (223 patients) after RP
- Prognostic endpoints:
  - Death of disease (DOD): Time from RP to death from PCa
  - Biochemical recurrence (BCR): Time from RP to PSA level ≥ 0.2 ng/ml
- Construction of ProstaTrend:
  - Meta-analysis based on univariate Cox regression models per gene and cohort (~ 1400 genes)
  - Combination of these genes by a weighted median approach into a prognostic score:
    - Risk groups: PTS > 0: Increased risk; PTS < 0: Reduced risk

# Existing research II

ProstaTrend-ffpe [2]: Extension of ProstaTrend to formalin-fixed paraffin-embedded (FFPE) biopsy specimens:

- FFPE-conserved tissue specimens from two cohorts resulting in 176 patients
- Primary prognostic endpoint: Biochemical recurrence (BCR)
- Construction of ProstaTrend-fppe:
    - Filtered genes from ProstaTrend susceptible to FFPE-associated degradation
    - Kept only genes that showed consistent prognostic effects

    → 204 selected genes

# Consulting Goals

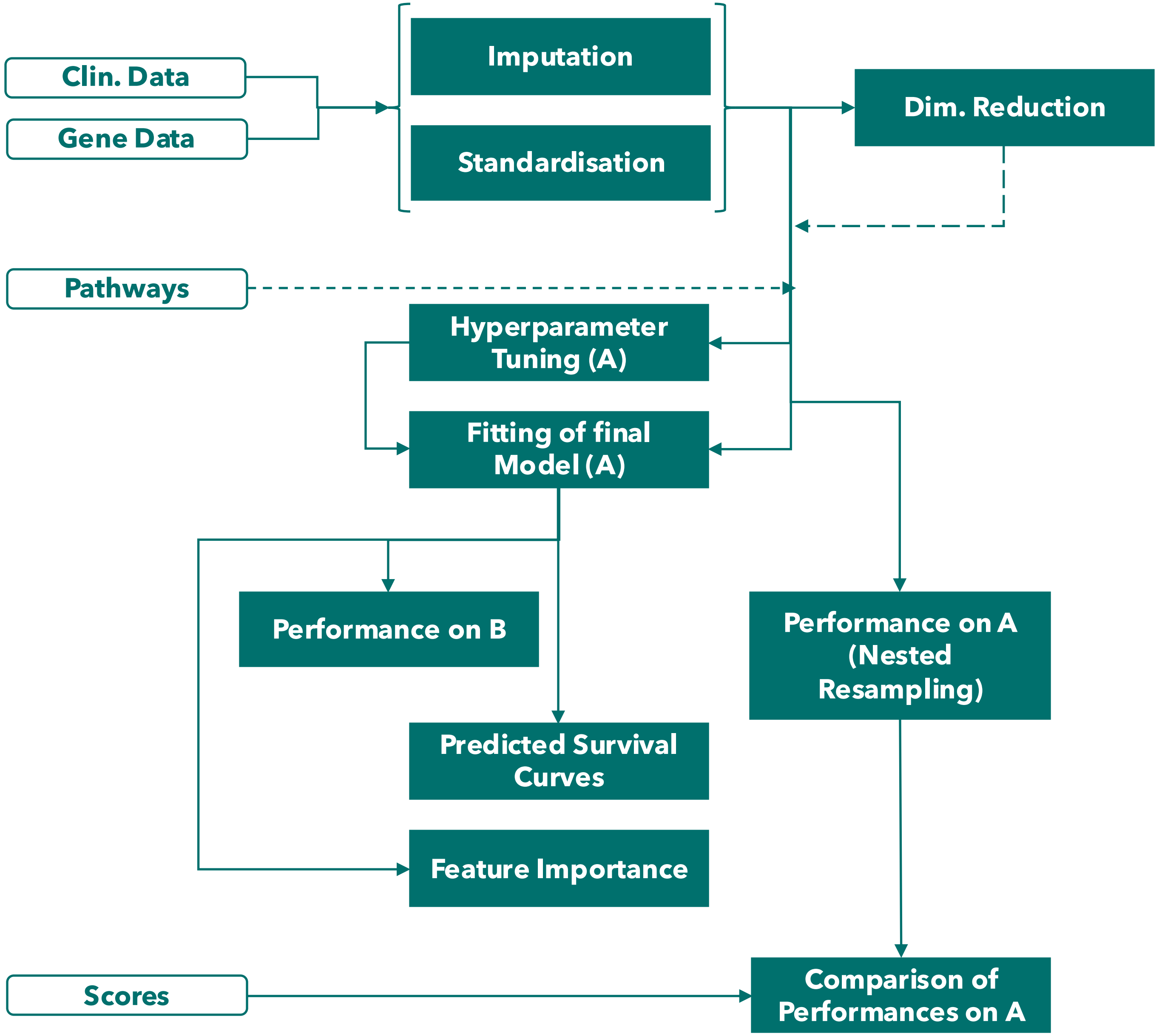
- Model the time to BCR using patients' time-to-event, clinical and gene expression data as potential features using various machine learning models
- Comparison of performance across applied models
- Comparison of model performances and the ProstaTrend-ffpe score

Additionally:

- Evaluate the performance w.r.t to the different data origins
- Evaluate the value of incorporating genetic data into risk assessment
- Selected features across the different models and connection to genes used in ProstTrend-ffpe

# Outline

- Data & Preprocessing
- Modelling Process
- Models
- Results
- Summary and Critique


Clin. Data
Gene Data
Imputation
Standardisation
Dim. Reduction
Pathways
Hyperparameter Tuning (A)
Fitting of final Model (A)
Performance on B
Performance on A (Nested Resampling)
Predicted Survival Curves
Feature Importance
Scores
Comparison of Performances on A

## Data & Preprocessing

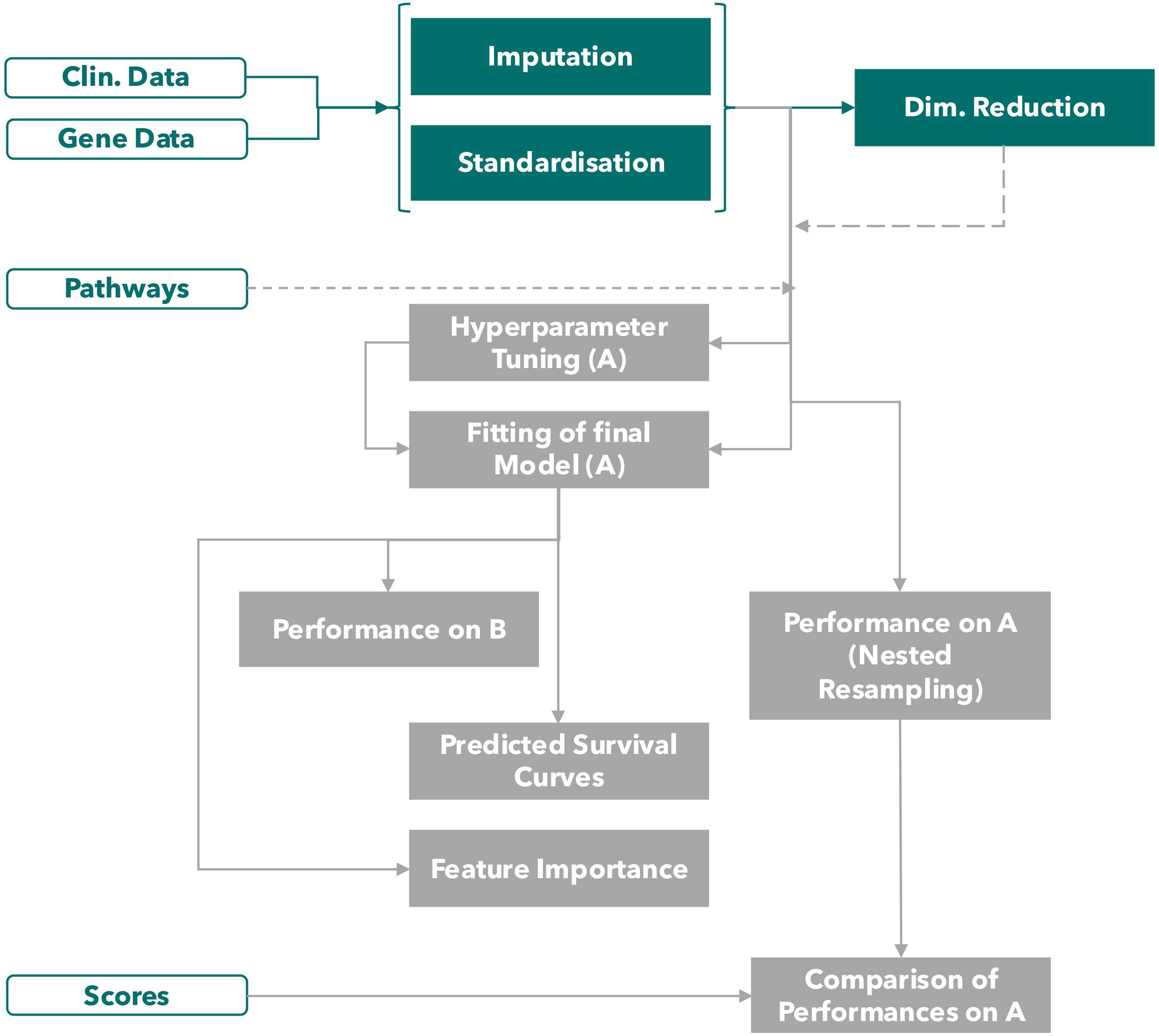
- Data Overview
- Clinical Data
- Pathways
- Gene Expression Data
- Resulting Data Sets

## Modelling Process

## Models

## Results

## Summary and Critique


Clin. Data
Gene Data
Imputation
Standardisation
Dim. Reduction
Pathways
Hyperparameter Tuning (A)
Fitting of final Model (A)
Performance on B
Performance on A (Nested Resampling)
Predicted Survival Curves
Feature Importance
Scores
Comparison of Performances on A

# Data Overview – Cohort Overview

| Group | Cohort | Name | Year | Location | Num. Patients |
|---|---|---|---|---|---|
| A | Cohort 1 | Atlanta | 2014 | USA | 100 |
| A | Cohort 2 | Belfast | 2018 | UK | 248 |
| A | Cohort 3 | CamCap | 2016 | UK | 112 |
| A | Cohort 4 | CancerMap | 2017 | UK | 133 |
| A | Cohort 5 | CPC | 2017 | Canada | 73 |
| A | Cohort 6 | CPGEA | 2020 | China | 120 |
| A | Cohort 7 | DKFZ | 2018 | Germany | 82 |
| A | Cohort 8 | MSKCC | 2010 | USA | 131 |
| A | Cohort 9 | Stockholm | 2016 | Sweden | 92 |
| B | Cohort 10 | TCGA | 2015 | USA | 332 |
| B | Cohort 11 | UKD2 | 2020 | Germany | 164 |

Total:
1091

Total:
496

1587

- Group A: Used for training of models
- Group B: Used for the construction of ProstaTrend-ffpe

# Data Overview – Data Sets

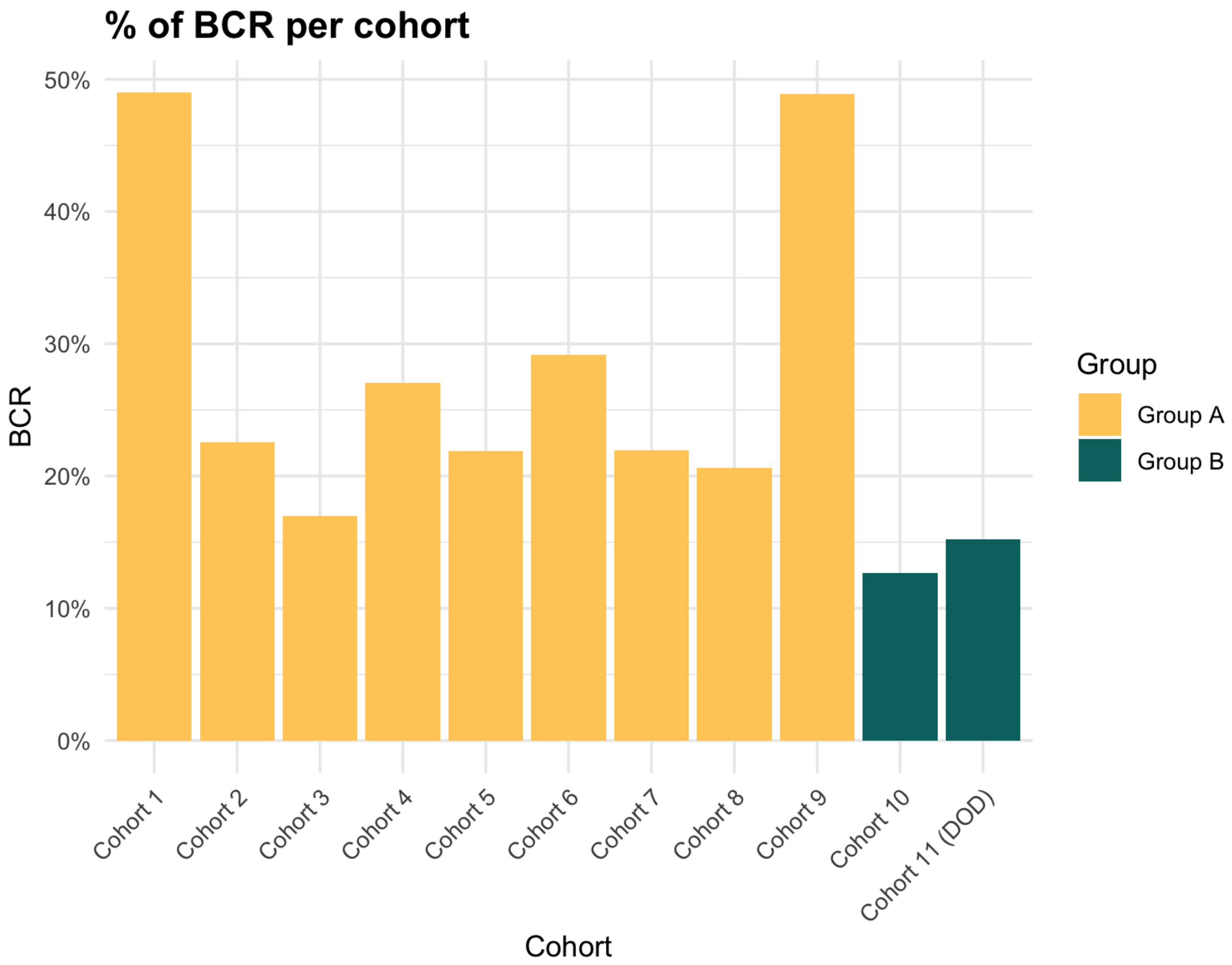
- Clinical Data
  - Target variables: BCR Status, Month to BCR
  - Used features:
    - Age
    - Tissue preservation method
    - Gleason score
    - Preoperative PSA

    → Selected based on medical sources
- RNA gene expression data
- ProstaTrend-ffpe scores resulting from previous research
- Pathways

# Clinical Data – Target Variable

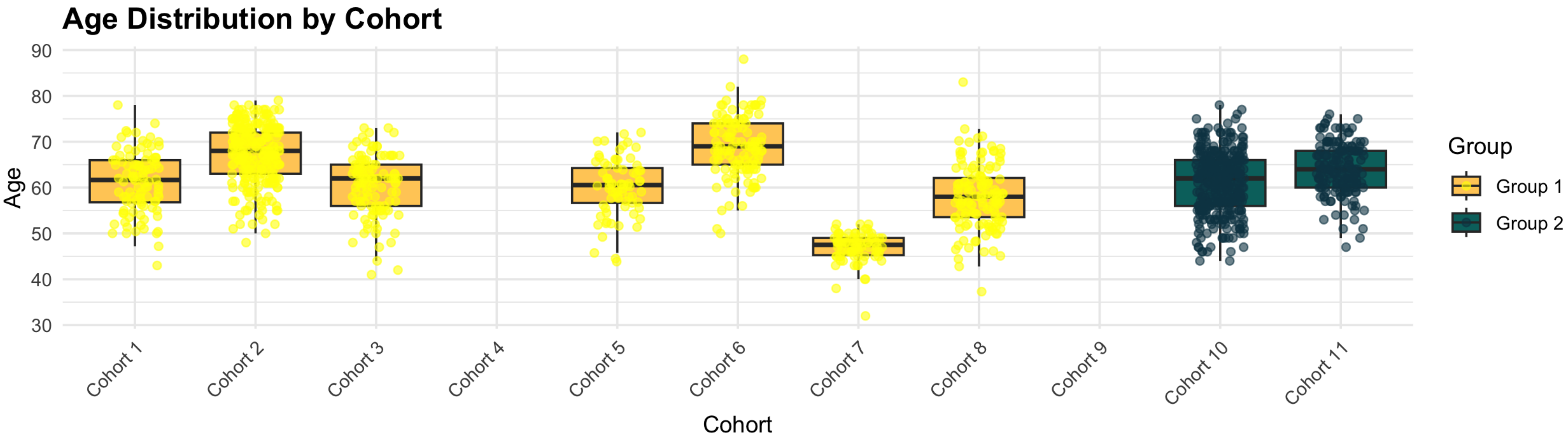
- Right censored data
- BCR status: Indicator on event of BCR
- Number of Month to BCR
- Preprocessing:
  - Month to BCR: 0 replaced by 0.0001
  - Month to BCR and BCR Status: Missing for Cohort 11,

  → replaced by Month to DOD and DOD Status



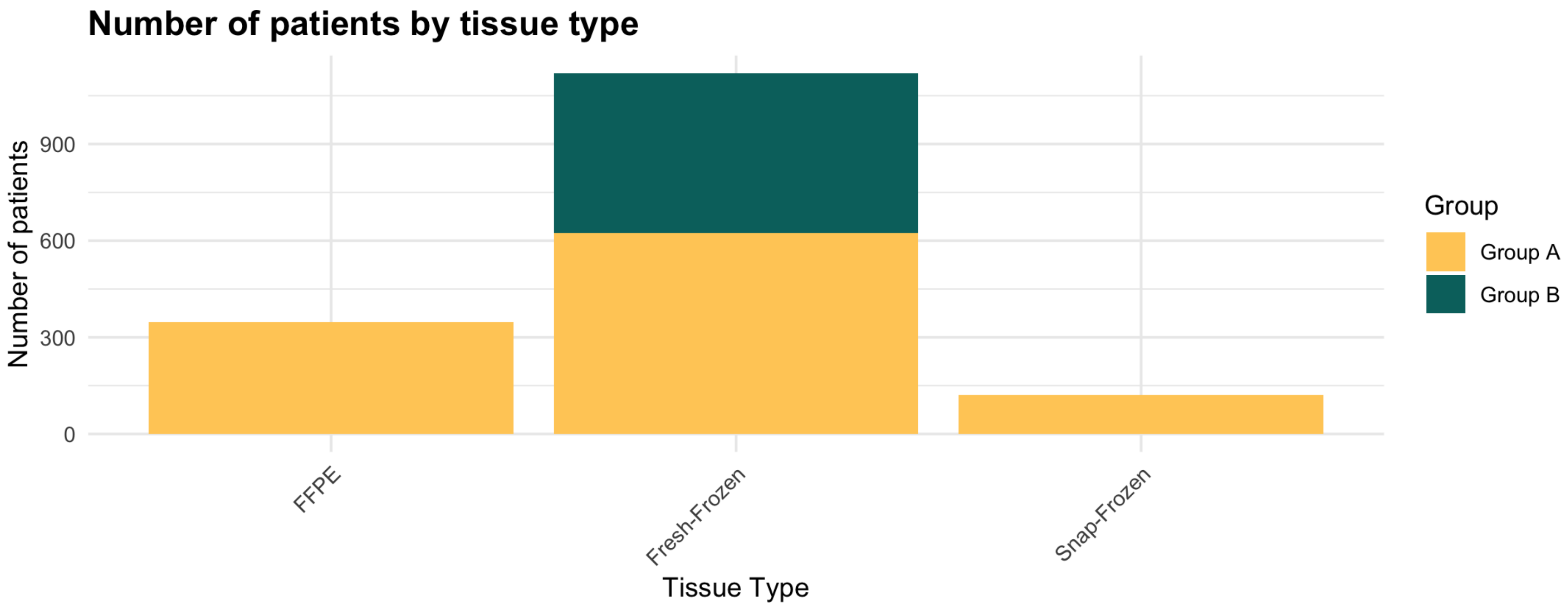
# Clinical Data - Age

- Age of the Patients at the time of the cancer diagnosis
- Medical research relates age to recovery prognosis [13]



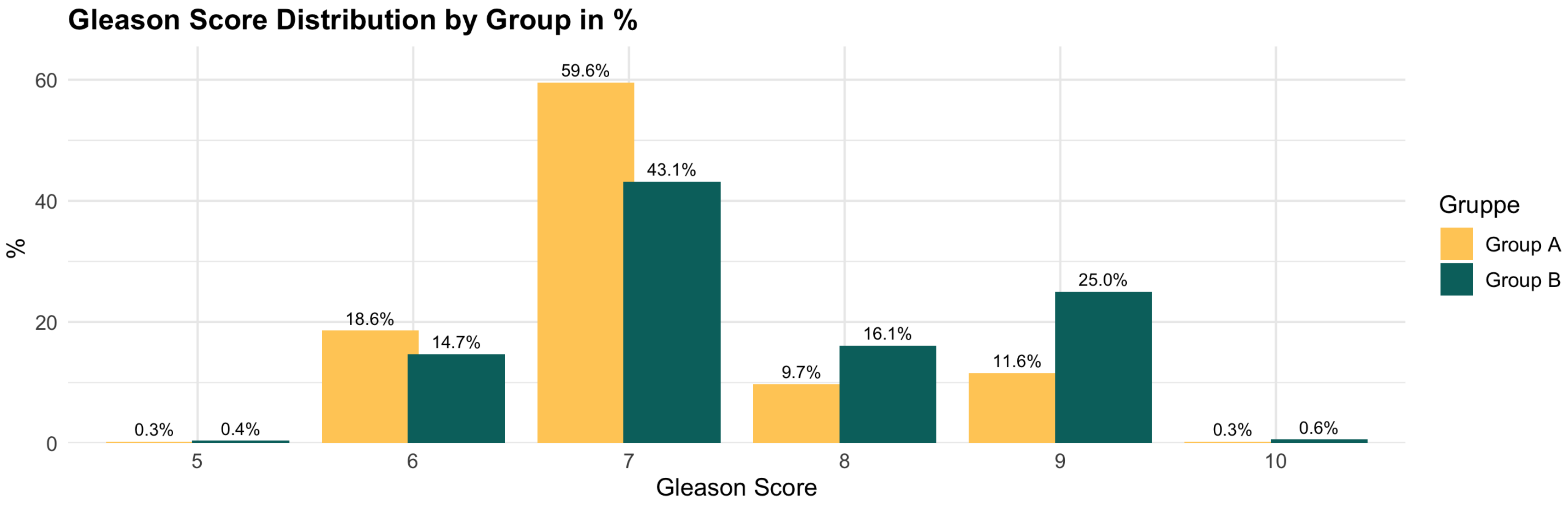
# Clinical Data – Tissue Preservation

- Indicates the tissue preservation method
  - Types: Fresh-Frozen, Formalin-Fixed Paraffin-Embedded and Snap-Frozen
- Sample quality can vary depending on the preservation method [14]



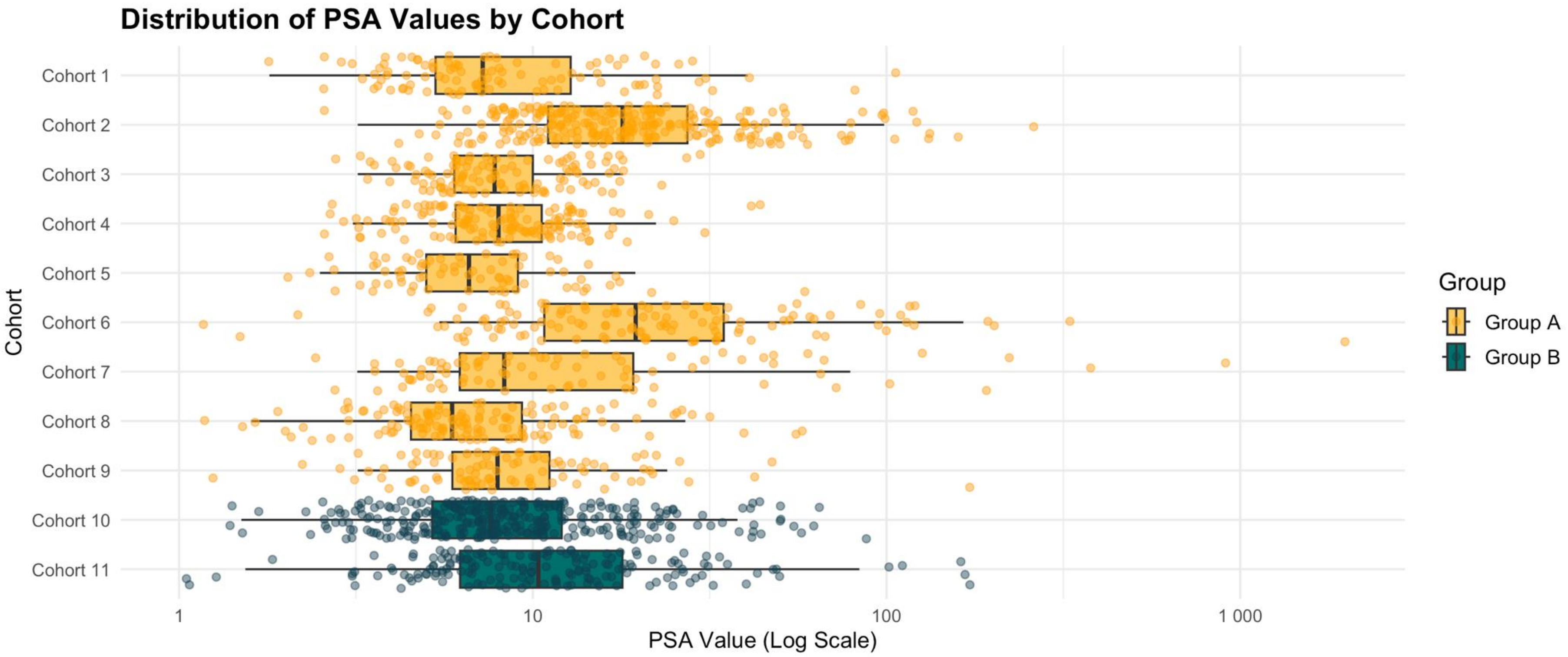
# Clinical Data - Gleason Score

- Metric measure for aggressiveness for PCa
  - Low score → Low aggressiveness
  - High score → High aggressiveness
- Commonly used predictor for PCa outcomes [15]



# Clinical Data - Preoperative PSA

- Preoperative values of the Prostate-specific antigen (PSA)
- Relevant blood marker for the diagnosis of PCa: High levels associated with worse outcomes [16]



# Clinical Data - Imputation

- Missing values:

| Feature | # of missing values | Fully missing |
|---|---|---|
| Age | 225 | 4, 9 |
| Tissue | - | - |
| Gleason Score | 3 | - |
| Preoperative PSA | 37 | - |

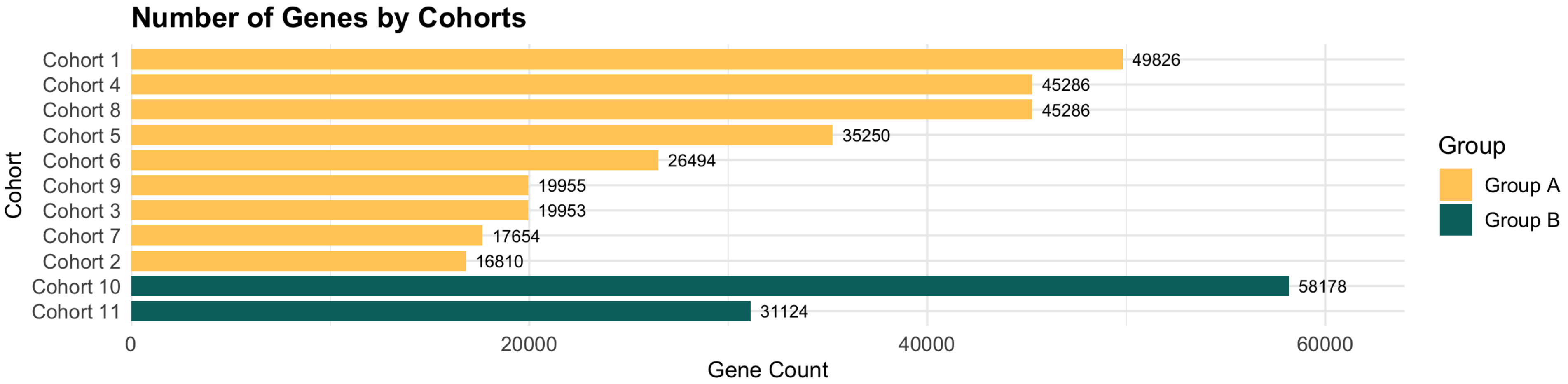
- Imputation via median value based on all cohorts in Group A

# Pathways

- Represent known biological processes or functional relations between genes e.g.:
    - Genes that affect the response to certain types of therapy e.g. radiotherapy
- Each pathway is comprised of several genes
- One gene can be part of multiple pathways
- Inclusion of pathway-based information can increase the predictive accuracy of models [4]
- Obtained data:
    - Identified Pathways: 143 pathways associated with PCa
    - Total Genes: 6,094 genes linked to these pathways

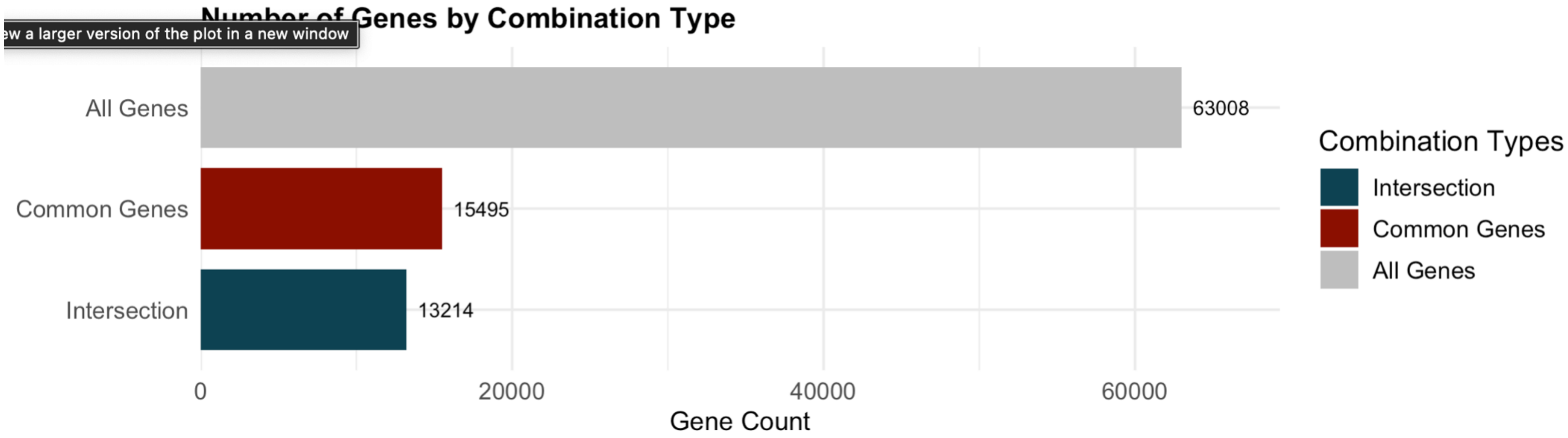
# Gene Data - Challenges

- Varying availability of gene data for all different cohorts:
  - 63008 unique genes over all cohorts
  - 16810 genes in smallest and 58178 in biggest expression data set
- Correlation > 0.75 for 202 genes (based on genes present across all cohorts of Group A in Group A patients)
- Resulting challenges:
  1) Effective use of available data
  2) Multicollinearity, potentially affecting model accuracy
  3) Risk of COD and computational issues due to high dimensionality



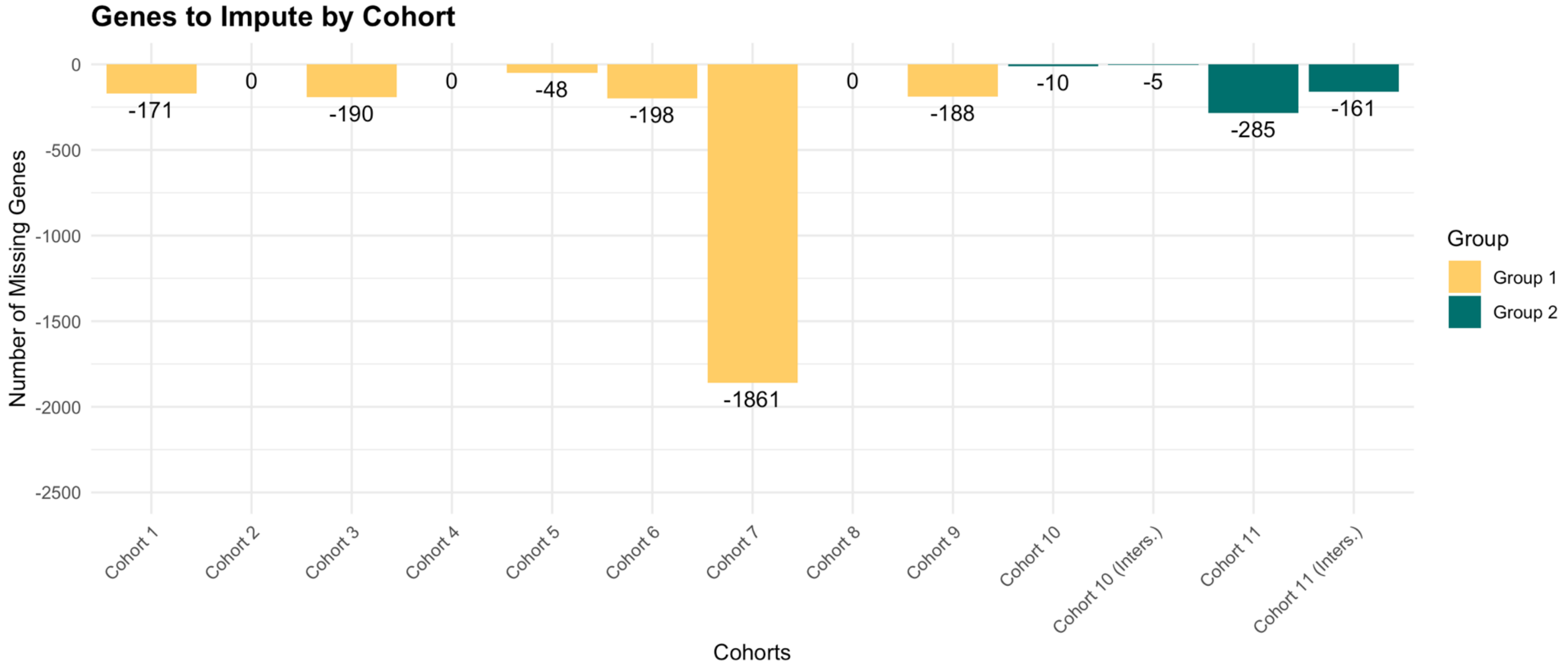
# Gene Data - Challenge I

1) Effective use of available data:
   - Use of models that can handle (block-wise) missing data
   - Use the intersection of genes
     - Pros: Easy to implement, not induced bias due to imputation
     - Cons: Potentially high loss of information
   - Use common genes (available for ≥ 80% of patients of Group A) [17]
     - Pros: Increased use of information
     - Cons: Quality heavily depends on imputation method



# Gene Data - Challenge I

- Imputation of missing expressions for each cohort
  - For group A only relevant for the Common Genes data set
  - For group B relevant for Common and Intersection Genes data sets
  - Imputation method: k-nearest neighbor with k=35



# Gene Data – Challenge II & III

2) Multicollinearity
    - Use of methods that can handle multicollinearity
    - Dimensionality reduction
3) Risk of COD and computational issues due to high dimensionality
    - Dimensionality reduction

Dimensionality reduction Pros:

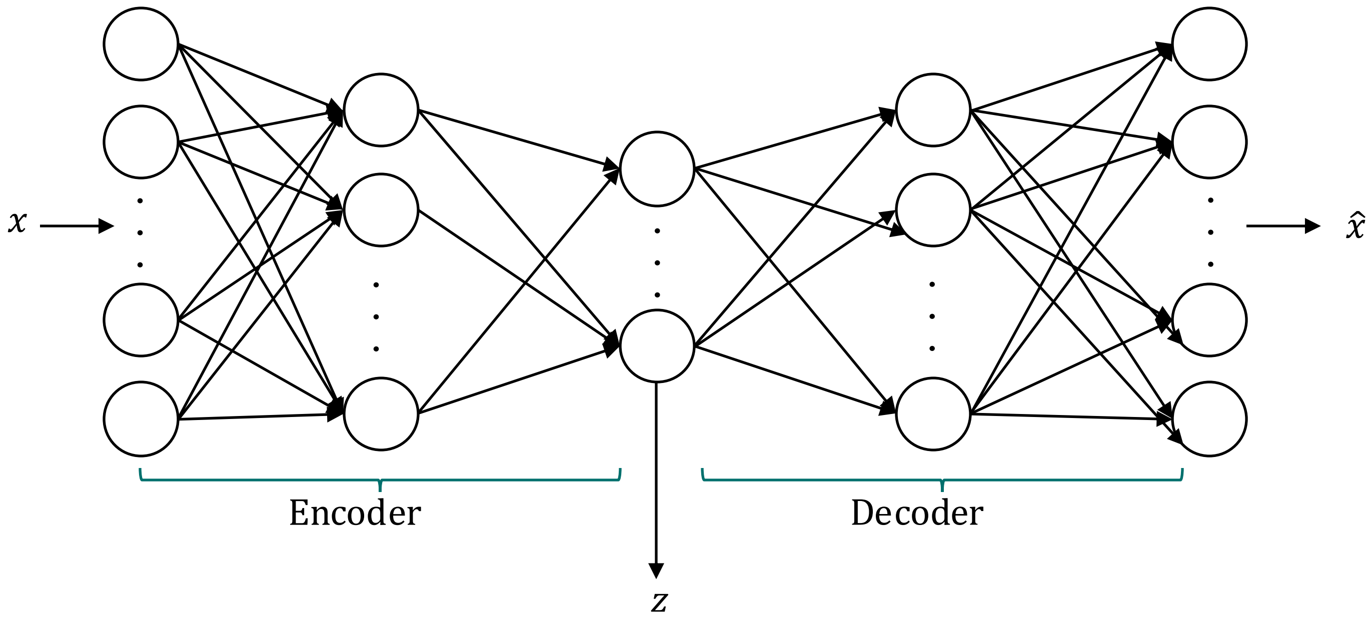
- Can reduce the impact of highly correlated feature
- Reduction of computational costs

Dimensionality reduction Cons:

- Potential loss of information
- Quality heavily depends on reduction method

# Gene Data – Challenge II & III

- Autoencoder: Unsupervised neural network to obtain latent representation z of input x
- Architecture [4]:
  - Input: Intersection Genes
  - Encoder: Compresses x into z
  - Decoder: Reconstructs x from z
    Mirrors architecture of encoder
- Loss function: MSE
- Reason for usage:
  - Captures non-linear relationships in data as opposed to e.g. PCA



# Resulting Data Sets

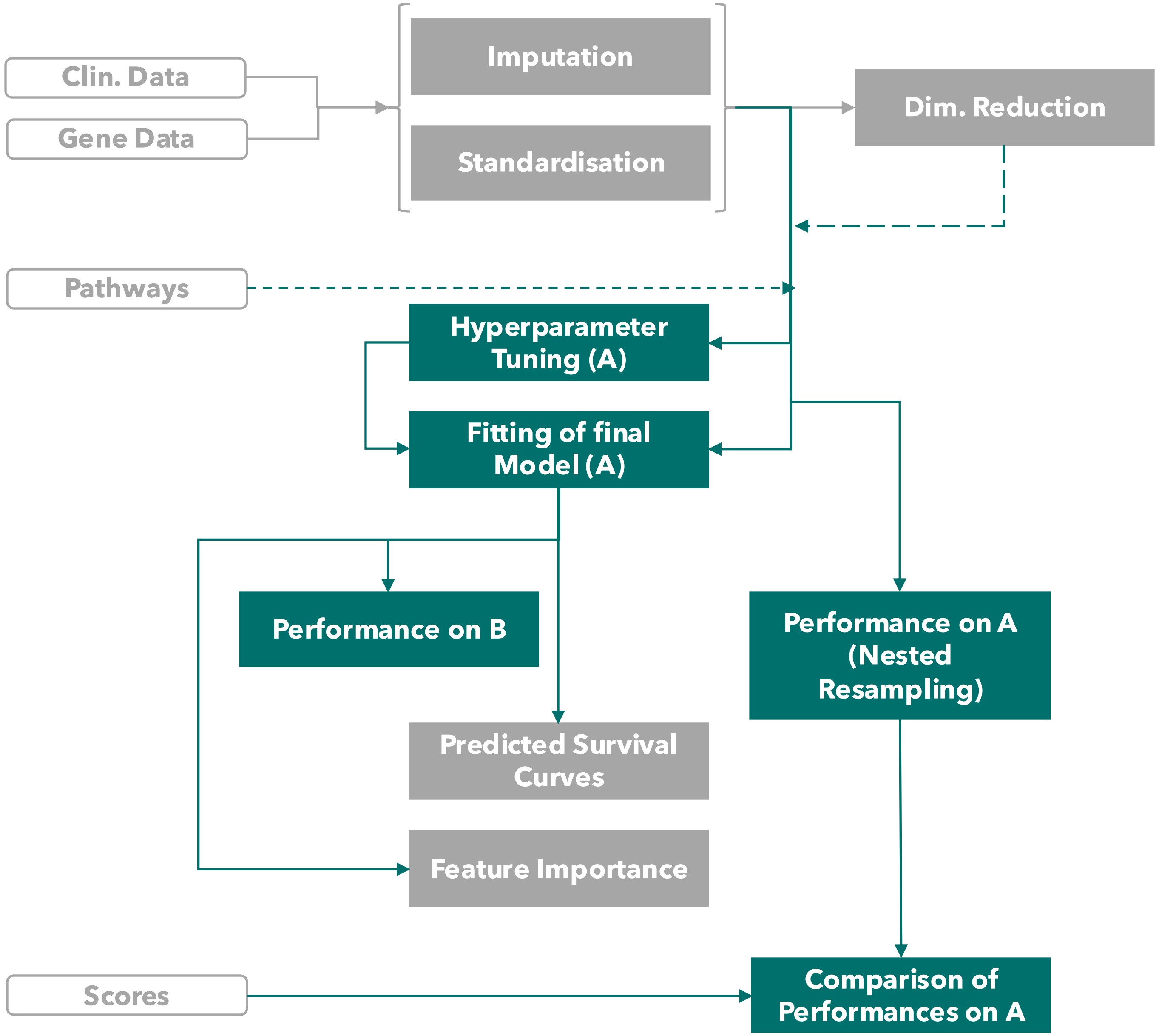
Resulting data sets:

- Intersect. Genes + Clinical Data
- Autoencoder + Clinical Data
- Common Genes + Clinical Data
- Pathways
- Only Intersect. Genes
- Only Clinical Data
- Only Autoencoder
- Only Common Genes
- Block Data

Preprocessing steps across all data sets:

- Correcting data types
- Standardization of numerical features (based on respective cohort)
  - Expressions
  - Gleason score
  - Preoperative PSA
  - Age

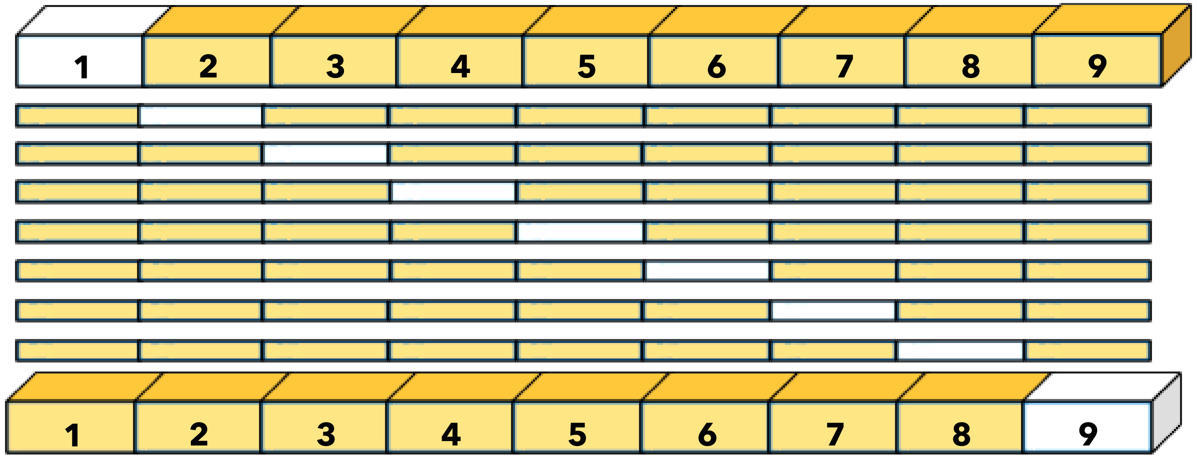
- Data
- Modelling Process
  - Resampling Strategy
  - Performance Evaluation
- Models
- Results
- Summary and Critique


Clin. Data
Gene Data
Imputation
Standardisation
Dim. Reduction
Pathways
Hyperparameter Tuning (A)
Fitting of final Model (A)
Performance on B
Predicted Survival Curves
Feature Importance
Performance on A (Nested Resampling)
Scores
Comparison of Performances on A

# Resampling Strategy

Leave-one-cohort-out Cross-Validation as resampling method:

- Iterates through cohorts with each cohort serving once as test data
- Reasons for usage:
  - Mimics the real use case of a new patient having to be modeled
  - Disregarding the cohort structure leads to over-optimistic performance estimation [12]



# Resampling strategy

GridSearch as Search Strategy:

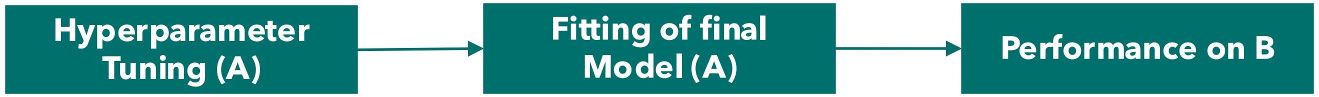
- Iterates all possible combinations of the hyperparameter values
- Reason for usage: Relatively small hyperparameter grids

Concordance Index (C-Index) as performance measure:

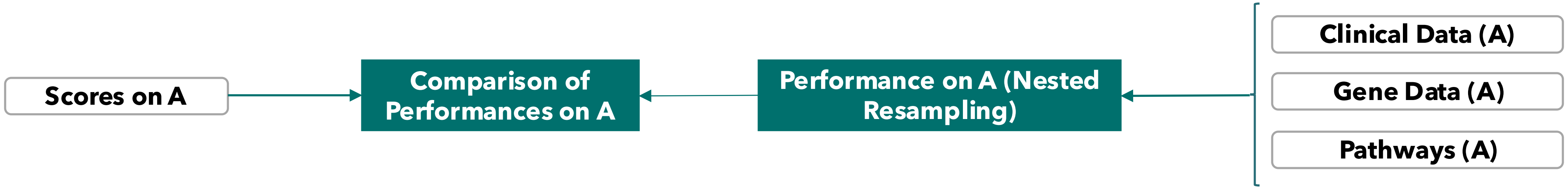
- Proportion of correctly ordered pairs among all comparable pairs of observation
  → 0.5: Random guessing; 1: Perfect order predicted
- Reason for usage: Availability for all applied methods

# Performance Evaluation - Overview

1. Performance of the models on B using the C-Index of final models on the B-cohorts

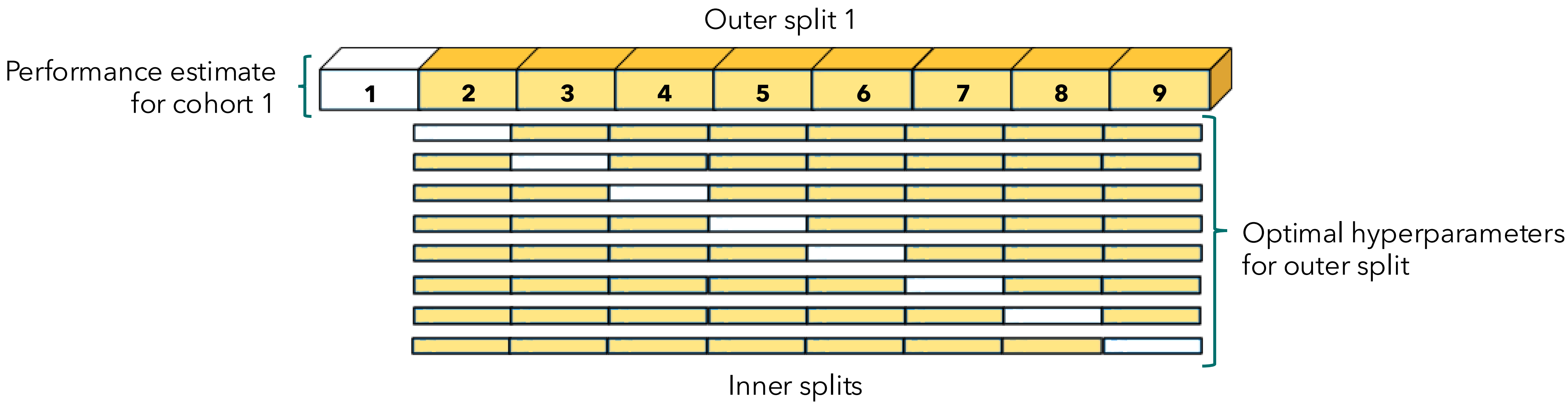


2. Performance comparison on A between the models and the ProstaTrend-ffpe:
    - C-Index calculation of ProstaTrend-ffpe scores on A
    - Estimated performance of the models on A obtained during nested resampling

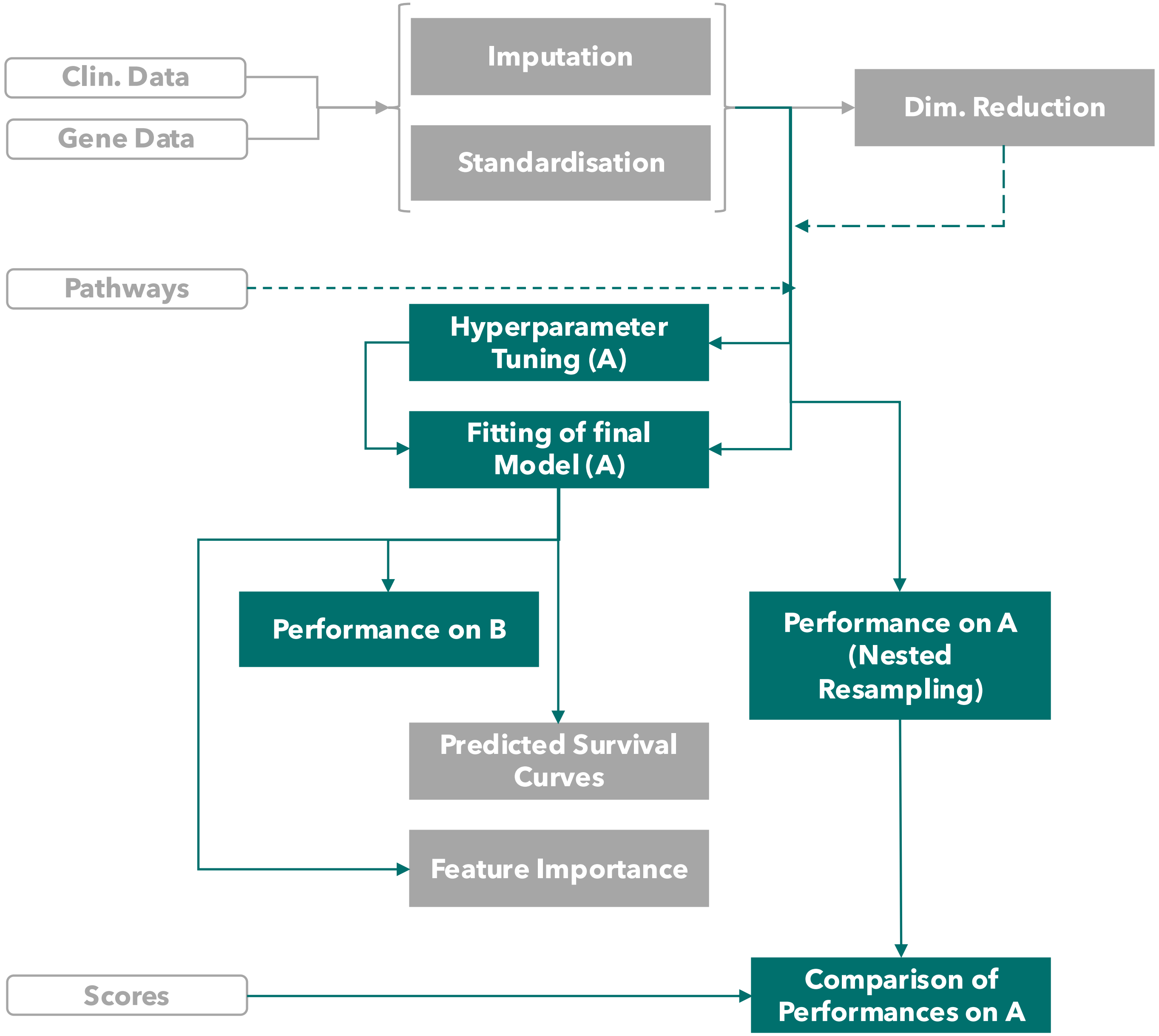

Scores on A
Comparison of Performances on A
Performance on A (Nested Resampling)
Clinical Data (A)
Gene Data (A)
Pathways (A)

# Performance Evaluation – Nested Resampling

- Estimated performance of the models on A obtained during nested resampling
  - Outer splits: Performance eval. on respective cohort
  - Inner splits: Hyperparameter tuning for the respective outer split
  - Reason for usage: Prevents information leakage and thus provides more reliable performance estimates [11]


Outer split 1
Performance estimate for cohort 1
1
2
3
4
5
6
7
8
9
Optimal hyperparameters for outer split
Inner splits

- **Data**
- **Modelling Process**
- **Models**
  - Standard Models
  - DeepSurv
  - Cox-PASNet
  - Priority Lasso
- **Results**
- **Summary and Critique**


Clin. Data
Gene Data
Imputation
Standardisation
Dim. Reduction
Pathways
Hyperparameter Tuning (A)
Fitting of final Model (A)
Performance on B
Predicted Survival Curves
Feature Importance
Performance on A (Nested Resampling)
Scores
Comparison of Performances on A

# Models - Standard Models

Penalized Cox Proportional-Hazards model:

- Assumes proportional hazards and a linear predictor; Uses L1 Penality; Estimates the hazard function [9]
- Reasons for usage: Scales well with high-dimensional data [10], interpretability of coefficients [9]

Random Survival Forest:

- Tree based bagging technique; Estimates the survival function [6]
- Reasons for usage: Low risk for overfitting; Can capture non-linearities and complex interactions; Can be robust against feature correlation[7]; Measures for feature importance readily available

Gradient Boosting:

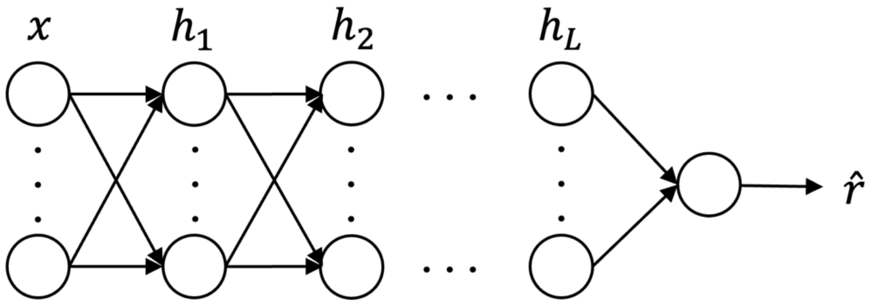
- Tree based boosting technique; Hazard rates as final model output [8]
- Reasons for usage: Can capture non-linearities and complex interactions; Can deal with unbalanced data Measures for feature importance readily available

# Models - DeepSurv

- Neural network using an objective similar to Cox PH regression models [18]
- Architecture:
  - Fully connected hidden layers $h_1,..,h_L$
  - Output layer: Risk Score $\hat{r}$
- Objective function:
  - Neg. log partial likelihood:

$$l(\beta) = -\frac{1}{N}\sum_{i=1}^{N} \delta_i\left( \hat{r}_i - \log\left(\sum_{j \geq i} \exp(\hat{r}_i)\right)\right)$$

→ Linear predictor from Cox PH regression is replaced by network output

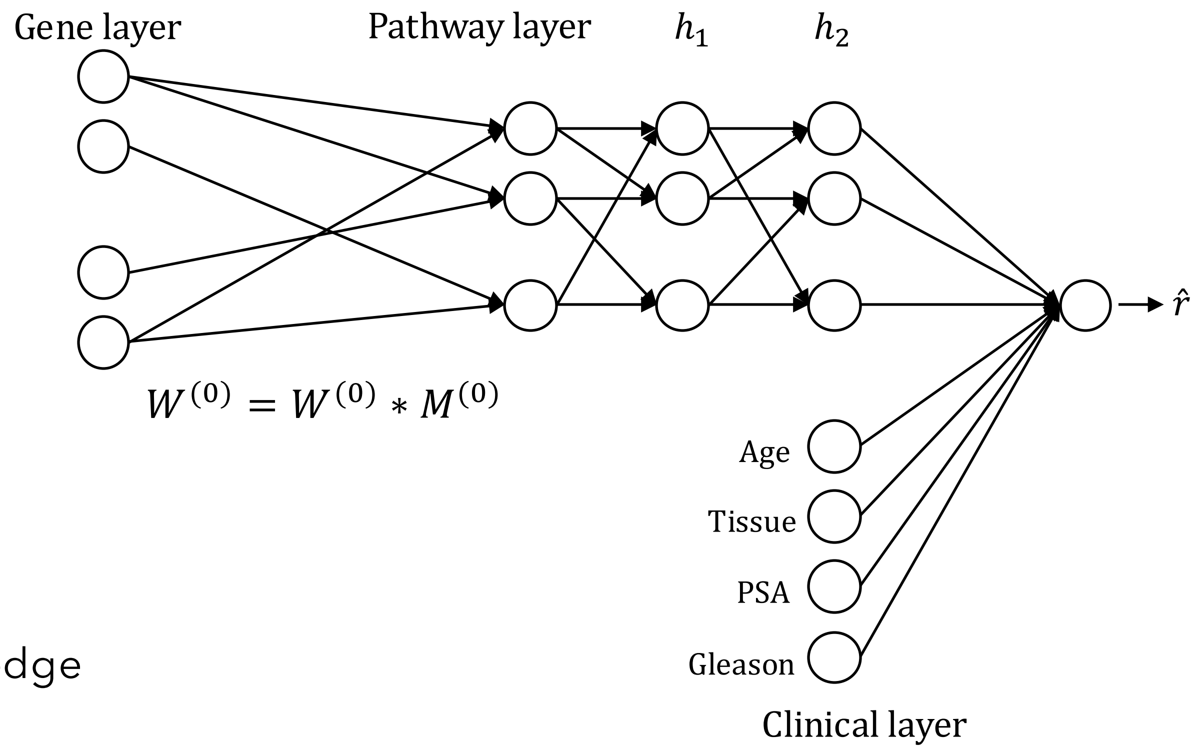


- Reason for usage:
  - Models non-linear relationships and feature interactions
  - Comparison to potentially more sophisticated neural network-based approaches

# Models – Cox-PASNet

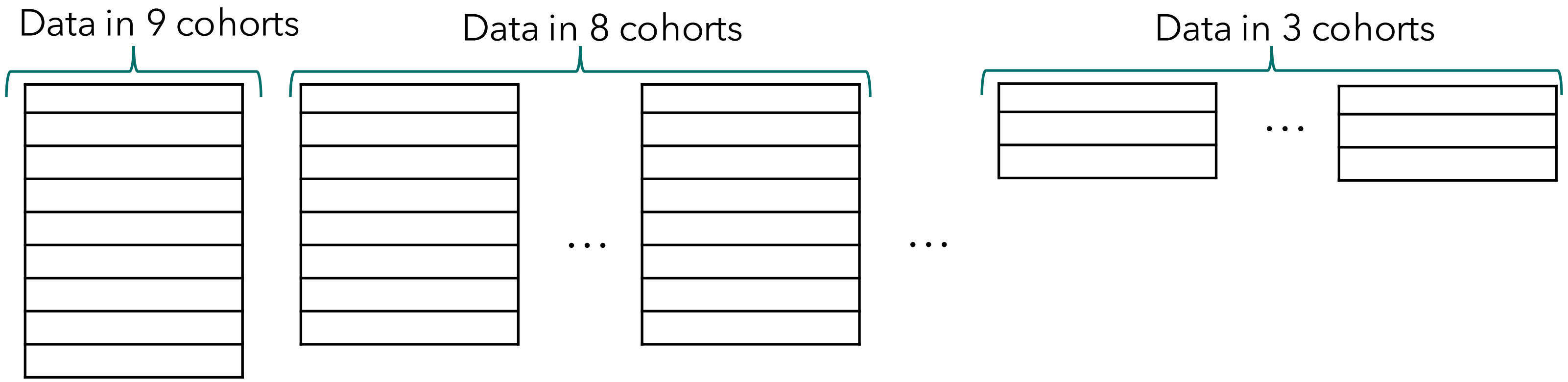
- Neural Network based approach which combines pathway data with clinical and gene data [5]
- Architecture:
  - Input layers: Gene and clinical layer
  - Pathway layer:
    - Constructed using a mask $M^{(0)}$ resulting in sparse connections $W^{(0)}$
    - Represents biological pathways
  - Output layer: Risk score $\hat{r}$
  - Obj. function:
    - Neg. log partial likelihood

    → Analogous formulation to DeepSurv
- Reason for usage: Leverages prior biological knowledge



# Models – Priority Lasso

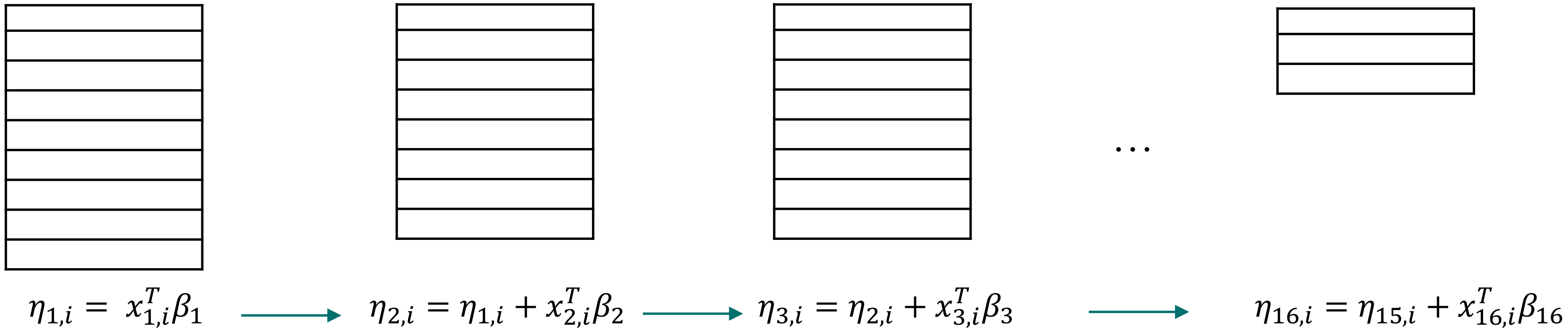
- Performs successive pen. Cox PH regressions on ordered blocks of data [3]:
  1. Define Covariate blocks and induced priority between them
     Here: Block structure based on available covariates
     - E.g.: Block 1: Data available in all cohorts; Blocks 2-10 : Data available in all but cohort 1,…,9



     - Further refinement: Min. # of cohorts = 3 and min. # of genes = 100
       → 16 Blocks, 42995 genes

# Models – Priority Lasso

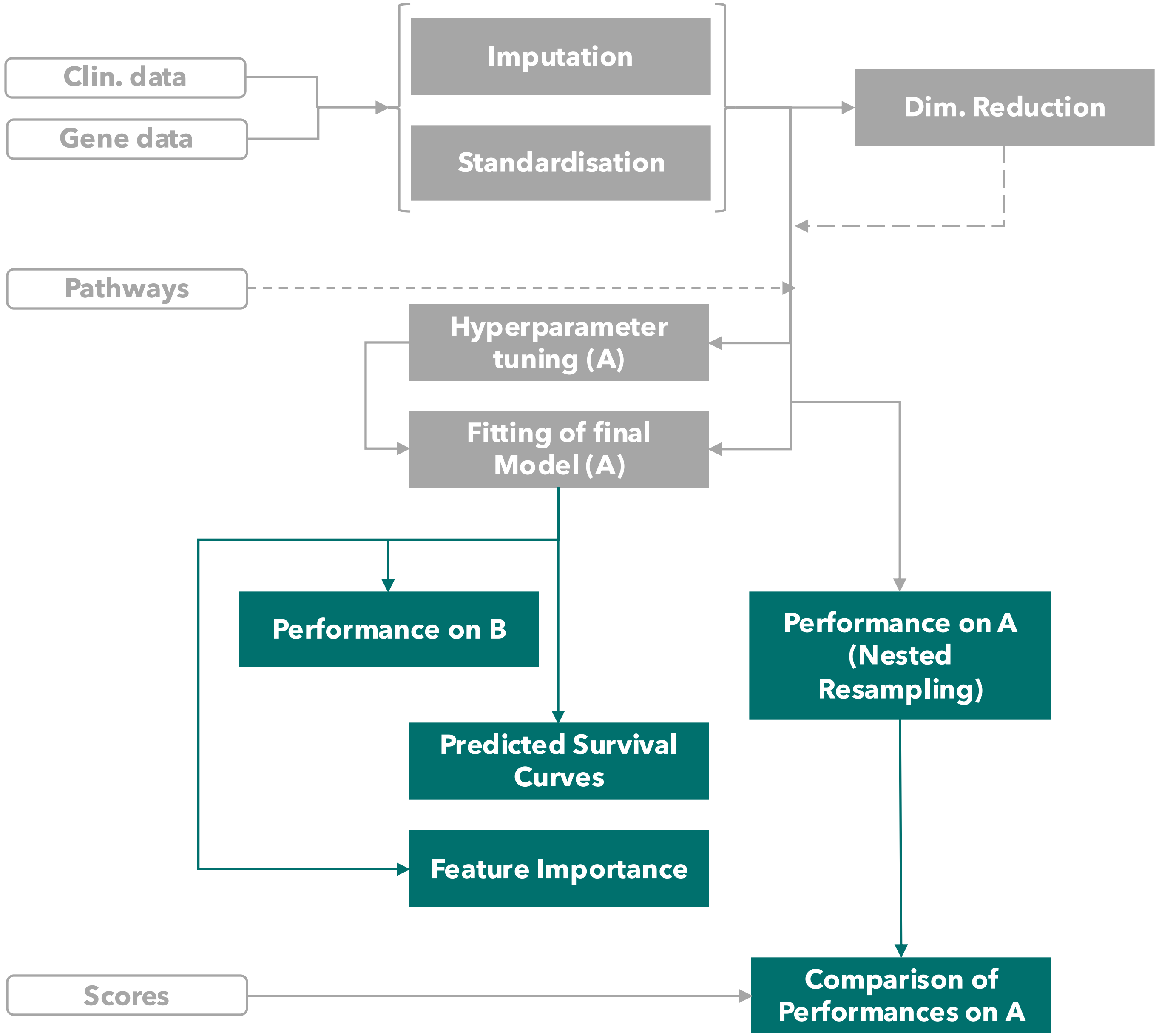
2. Based on block hierarchy, stepwise fitting of pen. Cox PH regressions
3. Hierarchical consideration of information by including predictions from higher-priority blocks as offsets for subsequent Lasso fits



$$\eta_{1,i} = x_{1,i}^T\beta_1 \longrightarrow \eta_{2,i} = \eta_{1,i} + x_{2,i}^T\beta_2 \longrightarrow \eta_{3,i} = \eta_{2,i} + x_{3,i}^T\beta_3 \quad \dots \quad \longrightarrow \eta_{16,i} = \eta_{15,i} + x_{16,i}^T\beta_{16}$$

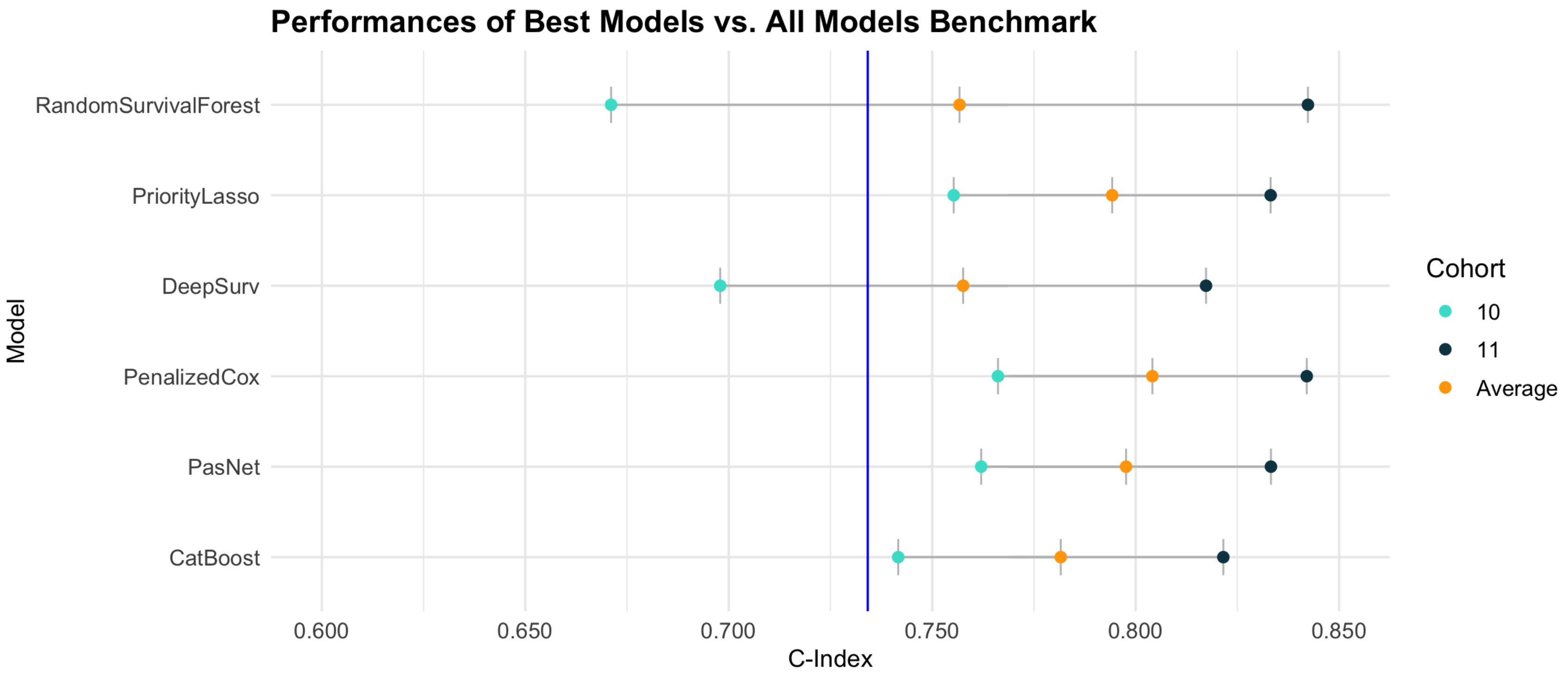
- Reason for usage: Manages blockwise-missing data during training and prediction via ignoring observations with missing values for the respective block

- **Data**
- **Modelling Process**
- **Models**
- **Results**
  - Model Performance on Group B
  - Models vs. Score on Group A
  - Survival Curves
  - Feature Importance
- **Summary and Critique**



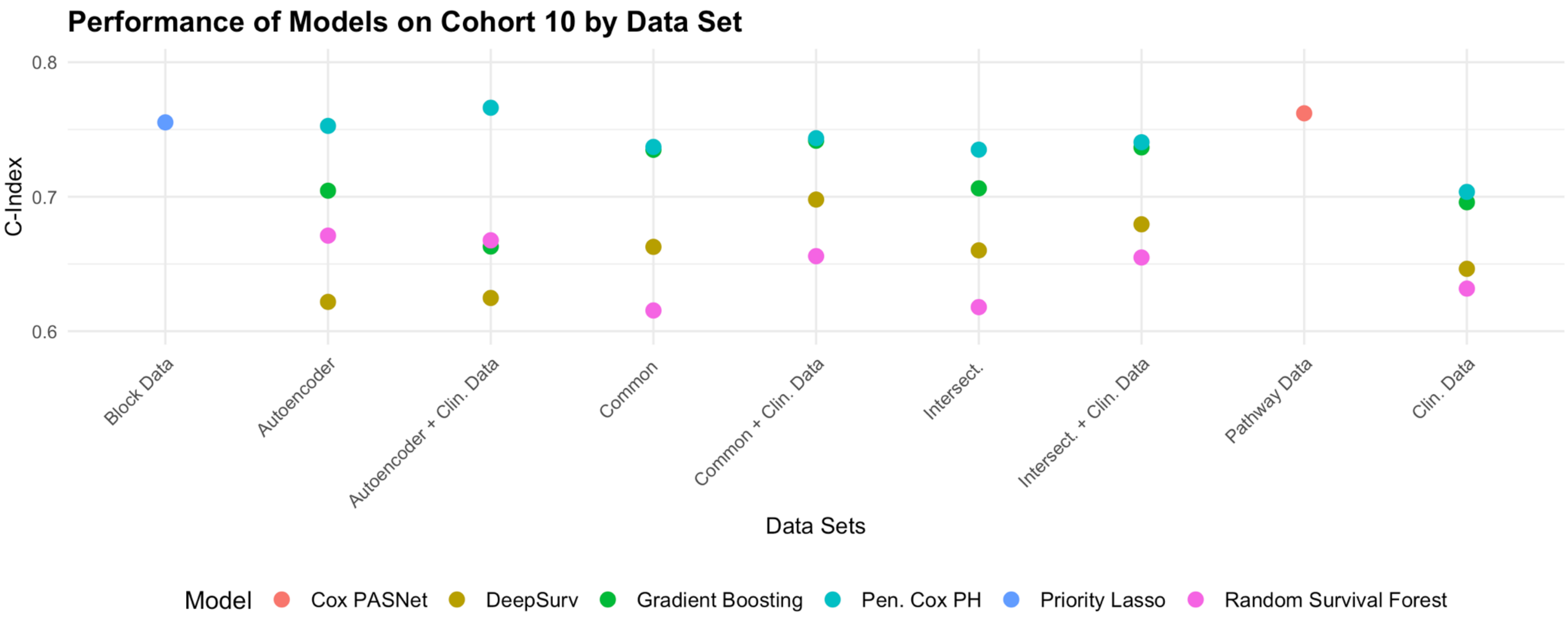
Clin. data
Gene data
Imputation
Standardisation
Dim. Reduction
Pathways
Hyperparameter tuning (A)
Fitting of final Model (A)
Performance on B
Predicted Survival Curves
Feature Importance
Performance on A (Nested Resampling)
Scores
Comparison of Performances on A

# Model Performance on Group B



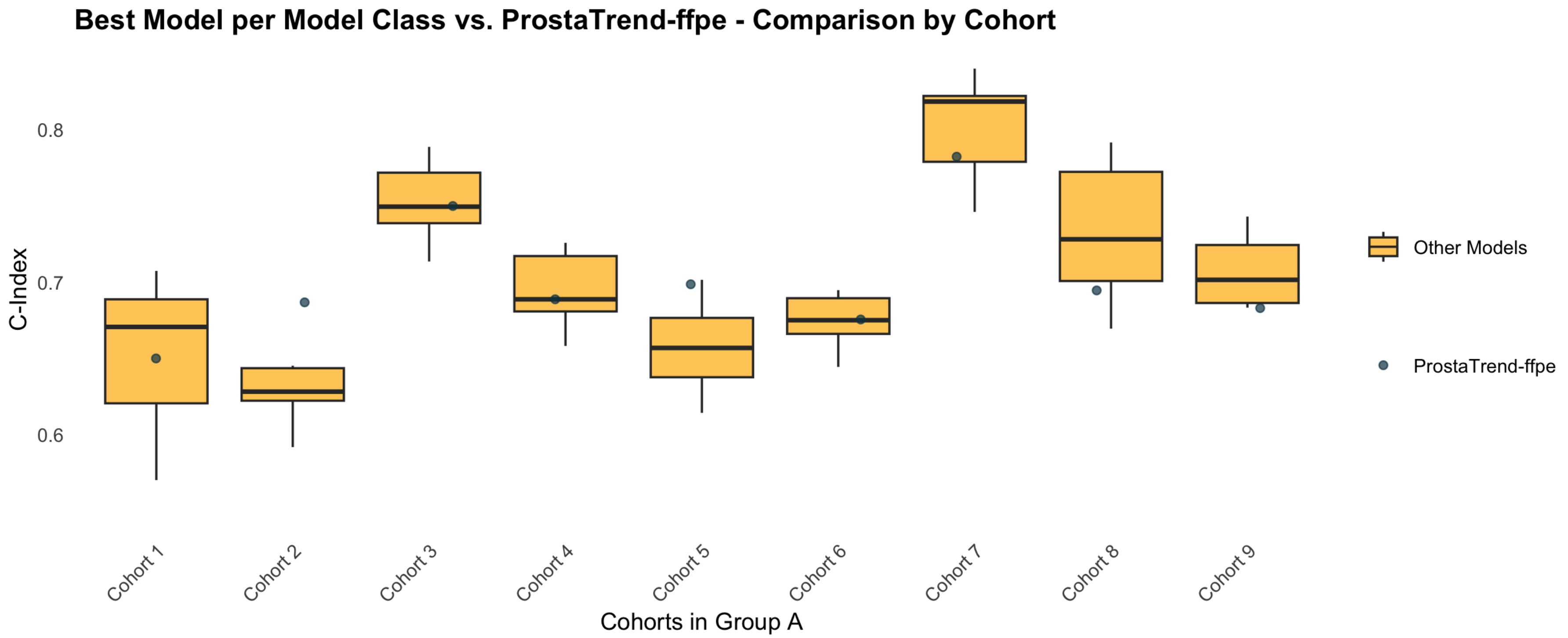
- Clear separation of performance between Cohorts 10 and 11
- Varying performance differences on cohorts across model classes
- Best model class on cohort 10: Cox PH; Best model class on cohort 11: RSF
- Worst model class on cohort 10: RSF; Worst model class on cohort 11: DeepSurv

# Model Performance on Group B



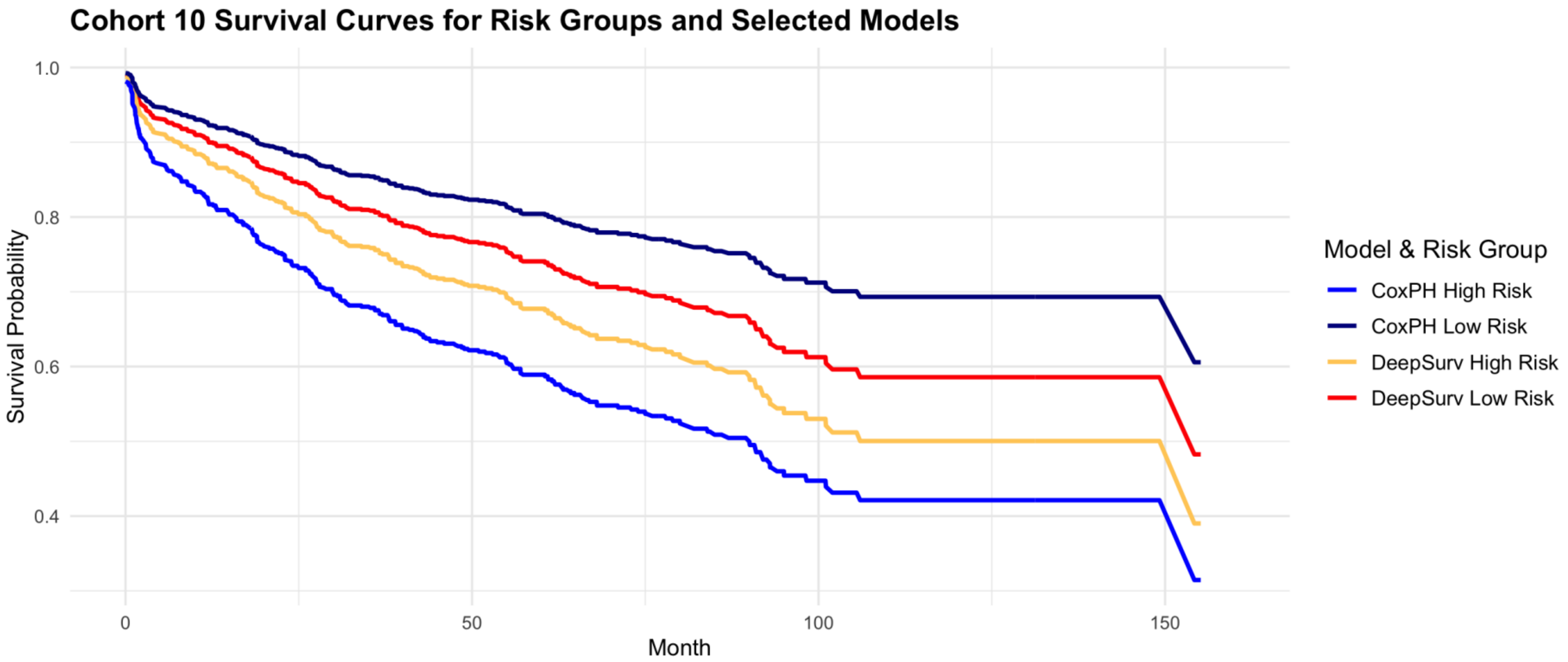
- Clear performance dominance of some models across all data sets
- Inclusion of clinical data mostly increases consistency of model performance per data set
- Combining gene data with clinical data (avg. CI = 0.71) seems to have prognostic value compared to “stand-alone” data sets (avg. CI = 0.68)

# Models vs. ProstaTrend-ffpe on Group A



- Varying performance across different cohorts
- ProstaTrend-ffpe:
  - Worse than median for 4 cohorts
  - Roughly equal to median performance for 3 cohorts
  - Better than median for cohort 2 and 5

# Survival Curves

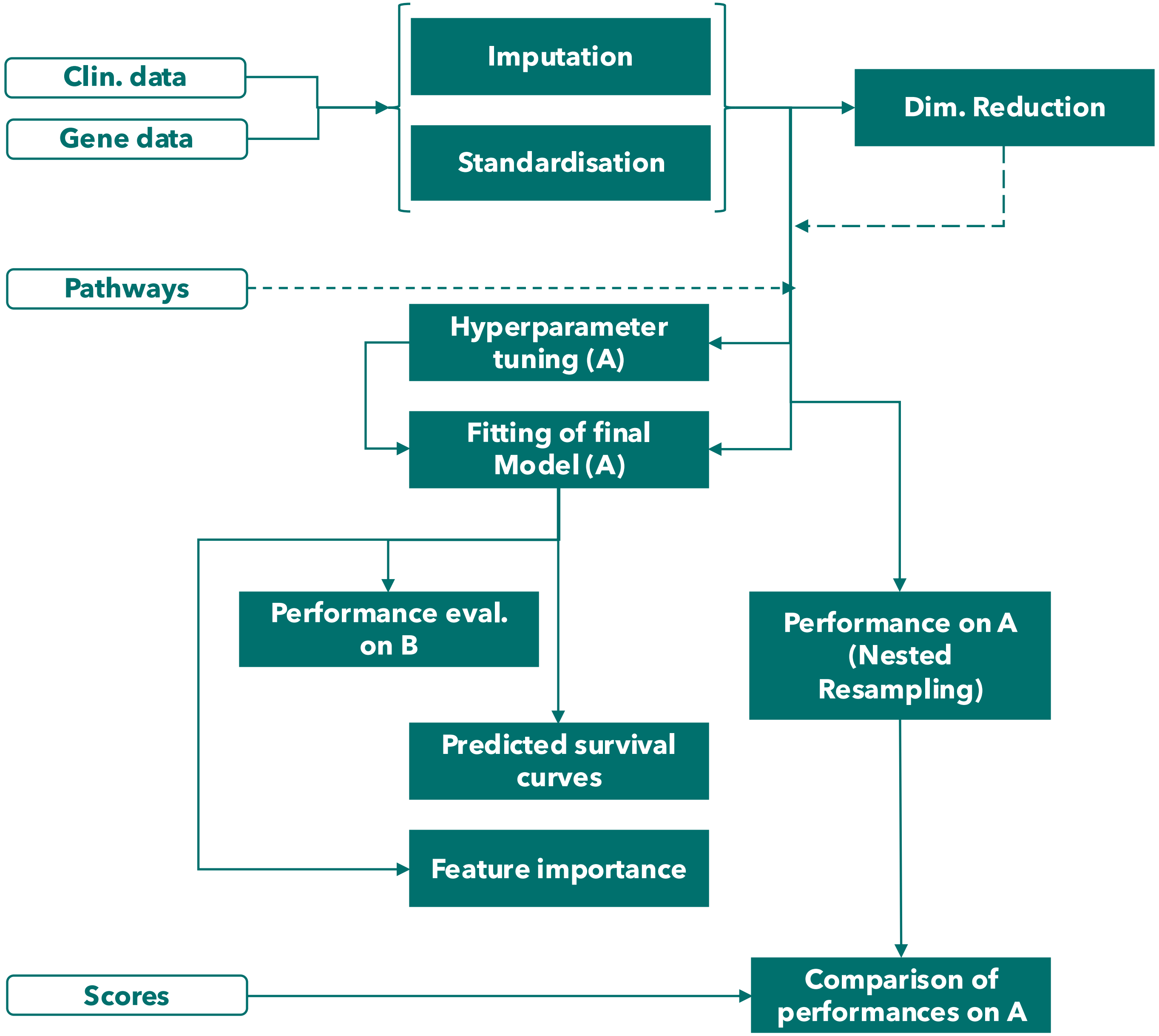


- High risk group: ProstaTrend-ffpe > 0
- Low risk group: ProstaTrend-ffpe < 0
- Clear separation of risk groups → coherence between applied models and ProstaTrend-ffpe
- Higher discriminative power for Cox PH regression model

# Feature Importance

- 4 Models with direct feature importance measure: RSF, Grad. Boosting, Pen. Cox PH (and Priority Lasso)
- Feature on Intersection Genes combined with clinical data:
  - Gleason Score selected in 2 Models:
    1) Penalized Cox PH: Rank 1 (Estimated coefficient: 0.17)
    2) Gradient Boosting: Rank 2
  - Preoperative PSA selected by Cox PH and
  - Age, Tissue not selected at all
- Comparison of ProstaTrend-ffpe with selected genes (Intersection data sets)
  1) 1078 Genes selected across all models → 23 of them in ProstaTrend-ffpe
  2) 17 Genes selected in at least 2 models → 2 of them in ProstaTrend-ffpe
  - General remark: Only 149 of ProstaTrend-ffpe's 204 genes available in the intersection data
    → Limits the comparability of the results

- Data
- Modelling Process
- Models
- Results
- Summary and Critique
  - Summary
  - Challenges and Critique


Clin. data
Gene data
Imputation
Standardisation
Dim. Reduction
Pathways
Hyperparameter tuning (A)
Fitting of final Model (A)
Performance eval. on B
Predicted survival curves
Feature importance
Performance on A (Nested Resampling)
Scores
Comparison of performances on A

# Summary

Performance evaluation:

- Performance across applied models:
  - Pen. Cox PH as best model class on cohort 10
  - RSF as best model class on cohort 11
- No coherent picture w.r.t. to the performance of the models and ProstaTrend-ffpe

Other:

- Performance for both ProstaTrend-ffpe and models varies across cohorts
- Increased performance for the combination of gene and clinical data compared to individual data sets
- Selected features across the different models:
  - Gleason score as important clinical variable
  - Small overlap of selected features with ProstaTrend-ffpe

# Challenges and Critique

General challenges :

- Different data availably

  → Intersection, imputation and Priority Lasso
- Computational issues due to high dimensional data

  → Dimensionality reduction
- Limited comparability of models and ProstaTrend-ffpe

  → Performance estimates of applied models obtained via nested resampling

(Resulting) critique and possible solutions:

- Improvement of both imputation and dimensionality reduction techniques

  → e.g. Pathway-based Autoencoder [4]
- Increasing the hyperparameter grids for tuning to address model specific weaknesses

# Bibliography

[1] Kreuz, M., Otto, D. J., Fuessel, S., Blumert, C., Bertram, C., Bartsch, S., Loeffler, D., Puppel, S. H., Rade, M., Buschmann, T., Christ, S., Erdmann, K., Friedrich, M., Froehner, M., Muders, M. H., Schreiber, S., Specht, M., Toma, M. I., Benigni, F., Freschi, M., … Horn, F. (2020). ProstaTrend-A Multivariable Prognostic RNA Expression Score for Aggressive Prostate Cancer. *European urology*, *78*(3), 452–459. https://doi.org/10.1016/j.eururo.2020.06.001

[2] Rade, M., Kreuz, M., Borkowetz, A. *et al.* A reliable transcriptomic risk-score applicable to formalin-fixed paraffin-embedded biopsies improves outcome prediction in localized prostate cancer. *Mol Med* **30**, 19 (2024). https://doi.org/10.1186/s10020-024-00789-9

[3] Klau, S., Jurinovic, V., Hornung, R., Herold, T., & Boulesteix, A.-L. (2018). Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. BMC Bioinformatics, 19(1), 322. https://doi.org/10.1186/s12859-018-2344-6

[4] da Costa Avelar, P. H., Wu, M., & Tsoka, S. (2023). Incorporating Prior Knowledge in Deep Learning Models via Pathway Activity Autoencoders. https://arxiv.org/abs/2306.05813

[5] Hao, J., Kim, Y., Mallavarapu, T. *et al.* Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. *BMC Med Genomics* **12** (Suppl 10), 189 (2019). https://doi.org/10.1186/s12920-019-0624-2

[6] H. Ishwaran, M. S. Lauer, E. H. Blackstone, M. Lu, and U. B. Kogalur. 2021. "randomForestSRC: random survival forests vignette." http://randomforestsrc.org/articles/survival.html.

[7] Robin Genuer, Jean-Michel Poggi, & Christine Tuleau-Malot (2010). Variable selection using random forests. *Pattern Recognition Letters, 31(14), 2225-2236.*

[8] *Regression: Objectives and metrics*. CatBoost. (n.d.). https://catboost.ai/docs/en/concepts/loss-functions-regression#Cox

[9] Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival analysis part II: multivariate data analysis--an introduction to concepts and methods. *British journal of cancer*, *89*(3), 431–436. https://doi.org/10.1038/sj.bjc.6601119

[10] *Penalized cox models*. scikit-survival 0.23.1. (n.d.). https://scikit-survival.readthedocs.io/en/stable/user_guide/coxnet.html

# Bibliography

[11] Bischl, B., Mersmann, O., Trautmann, H., & Weihs, C. (2012). Resampling Methods for Meta-Model Validation with Recommendations for Evolutionary Computation. Evolutionary Computation, 20, 249–275. https://doi.org/10.1162/EVCO_a_00069

[12] Hornung, R., Nalenz, M., Schneider, L., Bender, A., Bothmann, L., Bischl, B., Augustin, T., & Boulesteix, A.-L. (2023). Evaluating machine learning models in non-standard settings: An overview and new findings. https://arxiv.org/abs/2310.15108

[13] Knipper, S., Pecoraro, A., Palumbo, C., Rosiello, G., Luzzago, S., Deuker, M., Tian, Z., Shariat, S. F., Saad, F., Tilki, D., Graefen, M., & Karakiewicz, P. I. (2020). The effect of age on cancer-specific mortality in patients with prostate cancer: a population-based study across all stages. *Cancer Causes & Control, 31*(3), 283–290. https://doi.org/10.1007/s10552-020-01273-5

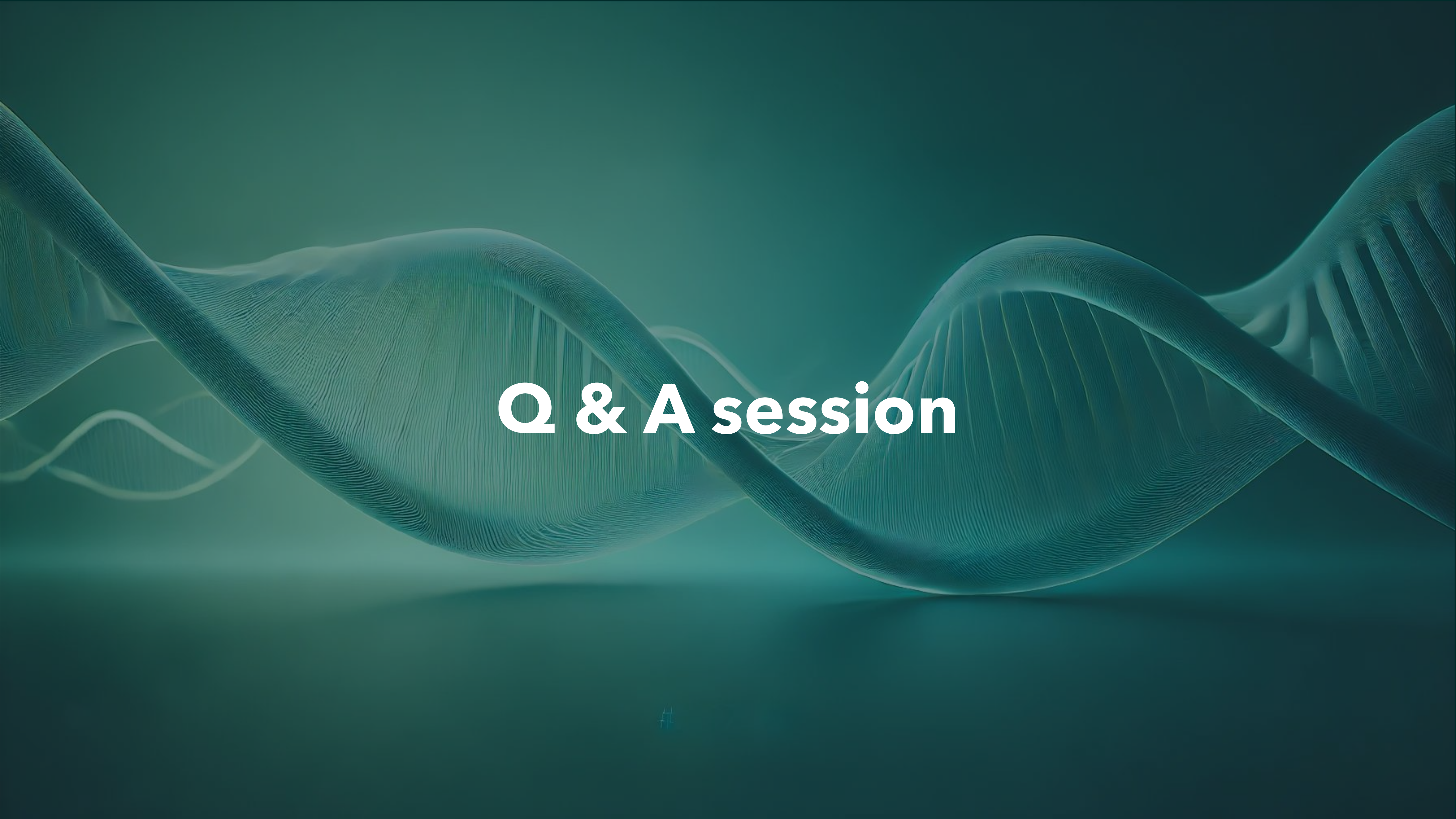
[14] Esteve-Codina, A., Arpi, O., Martinez-García, M., Pineda, E., Mallo, M., Gut, M., Carrato, C., Rovira, A., Lopez, R., Tortosa, A., Dabad, M., Del Barco, S., Heath, S., Bagué, S., Ribalta, T., Alameda, F., de la Iglesia, N., & Balaña, C. (2017). A Comparison of RNA-Seq Results from Paired Formalin-Fixed Paraffin-Embedded and Fresh-Frozen Glioblastoma Tissue Samples. *PLoS One, 12*(1), e0170632. https://doi.org/10.1371/journal.pone.0170632

[15] Egevad, L., Micoli, C., Samaratunga, H., Delahunt, B., Garmo, H., Stattin, P., & Eklund, M. (2024). Prognosis of Gleason Score 9–10 Prostatic Adenocarcinoma in Needle Biopsies: A Nationwide Population-based Study. *European Urology Oncology, 7*(2), 213–221. https://doi.org/10.1016/j.euo.2023.11.002

[16] D'Amico, A. V., Chen, M.-H., Roehl, K. A., & Catalona, W. J. (2004). Preoperative PSA Velocity and the Risk of Death from Prostate Cancer after Radical Prostatectomy. *The New England Journal of Medicine, 351*(2), 125–135. https://doi.org/10.1056/NEJMoa032975

[17] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17*(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

[18] Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network. *arXiv preprint*. https://arxiv.org/abs/1606.00931



# Q & A session

# Appendix

# Miscellaneous

| Model | Clin. | Common | AE | Inter. | AE + Clin. | Inter. + Clin | Common + Clin. | Block + Clin. | Pathways |
|---|---|---|---|---|---|---|---|---|---|
| Prio. Lasso | | | | | | | | X | |
| DeepSurv | X | X | X | X | X | X | X | | |
| Cox PH | X | X | X | X | X | X | X | | |
| PASNet | | | | | | | | | X |
| RSF | X | X | X | X | X | X | X | | |
| Grad. Boost. | X | X | X | X | X | X | X | | |

Total Number of models: 30

Patterns in model performances:

- Negative correlation (-0.474) between model performance and percentage of BCR events

# Autoencoder

- Architecture details:
  - 2 Hidden layers, with 128 and 62 nodes and ReLU activation
  - Based on already existing architecture used in [4]
- Loss function: $\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2$
- Possible critique:
  - Due to computational reasons the architecture wasn't systematically tuned but in the end an architecture analogously to [4] was used

# RSF

- Used splitting criterion: Log-rank splitting

$$L(X,c) = \frac{\sum_{j=1}^{m}\left(d_{j,L} - Y_{j,L}\,\frac{d_j}{Y_j}\right)}{\sqrt{\sum_{j=1}^{m}\frac{Y_{j,L}}{Y_j}\left(1-\frac{Y_{j,L}}{Y_j}\right)\left(\frac{Y_j-d_j}{Y_j-1}\right)d_j}}.$$

  - Interpretation: |*L*(*X*,*c*)| is a measure of node separation. The larger the value, the greater the survival difference between *L* and *R*, and the better the split is. The best split is determined by finding the feature *X*∗ and split-value *c*∗ such that |*L*(*X*∗,*c*∗)|≥|*L*(*X*,*c*)| for all *X* and *c* [6]

- Implementation details:
  - Used package: randomForestSRC
  - (Nested) resampling via custom function

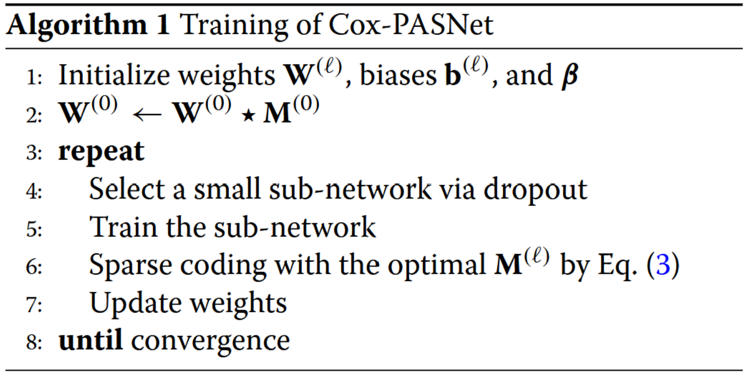
- Feature importance:
  - Used metric: Anti (Assigns cases to the opposite split)
  - VIMP calculated for the entire forest by comparing the perturbed OOB forest ensemble to the unperturbed OOB forest ensemble
    → yields ensemble VIMP, which does not measure the tree average effect of a variable, but rather its overall forest effect [6]

# Grad. Boosting

- Implementation details:
  - Used package: catBoost (Python implementation)
  - (Nested) resampling via sklearn.model_selection import GridSearchCV
- Feature importance:
  - Used metric: Prediction Values Change (shows how much on average the prediction changes if the feature value changes)

# Cox-PASNet I

- Implementation details:
  - Custom implementation based on authors' official repo
  - (Nested) resampling via sklearn.model_selection import GridSearchC
- Objective function:
  - $l(\Theta) = -\frac{1}{N}\sum_{i=1}^{N}\delta_i\big(\ \alpha_i - \log\big(\sum_{j\geq i}\exp(\alpha_i)\big)\big) + \lambda(||\Theta||_2)$
  - with $\Theta = \{\beta, W\}$ and $\alpha_i = h_{2,i}^T\beta$
- Sparse coding for interpretable networks:
  - Sparse connections between pathway layer and h1, h1 and h2 respectively via subnetwork training and local/subnetwork wise optimization of sparsity
- Possible critique:
  - Information loss of available genes due to restriction to genes only associated with pathways



**Algorithm 1** Training of Cox-PASNet

```
Initialize weights $\mathbf{W}^{(\ell)}$, biases $\mathbf{b}^{(\ell)}$, and $\boldsymbol{\beta}$
$\mathbf{W}^{(0)} \leftarrow \mathbf{W}^{(0)} \star \mathbf{M}^{(0)}$
repeat
    Select a small sub-network via dropout
    Train the sub-network
    Sparse coding with the optimal $\mathbf{M}^{(\ell)}$ by Eq. (3)
    Update weights
until convergence
```

# Priority Lasso

- Implementation details:
  - Used package: prioritylasso (R)
  - (Nested) resampling via custom function
- Objective for second block:

$$\sum_{i=1}^{n}\left(y_i-\hat{\eta}_{1,i}(\boldsymbol{\pi})_{\mathrm{CV}}-\sum_{j=1}^{p_{\pi_2}} x_{ij}^{(\pi_2)}\beta_j^{(\pi_2)}\right)^2+\lambda^{(\pi_2)}\sum_{j=1}^{p_{\pi_2}}\left|\beta_j^{(\pi_2)}\right|.$$

- Possible critique:
  - Propagating the off-set for missing features might lead to overfitting on the present training data per split
    → Off-Set calculation via Cross Validation as solution, but currently not implemented for family "Cox"
  - $$\hat{\eta}_{1,i}(\boldsymbol{\pi})_{\mathrm{CV}}=\hat{\beta}^{(\pi_1)}_{S\setminus S_k,1}x_{i1}^{(\pi_1)}+\ldots+\hat{\beta}^{(\pi_1)}_{S\setminus S_k,p_{\pi_1}}x_{ip_{\pi_1}}^{(\pi_1)}.$$

    → Imputation methods for missing values ("impute offset") but not computationally demanding
  - Definition of block heavily influence the modelling results
    → modelling splits should be treated as HPs

## Deep Surv

- Implementation details:
  - Based on git repo https://github.com/czifan/DeepSurv.pytorch
  - Final implementation uses
    - 2 Hidden Layers [256, 128]
    - Drop-Out of 0.2
    - Adam Optimization
    - Early Stopping with patience of 12, max 500 epochs
    - Learning Rate: 0.00001
    - Batch Size: 64
    - ReLU
- Possible critique:
  - Black Box → no information on relevant features /effect directions
  - Risk of overfitting
  - Many different possible HPs and grids make finding the optimal specifications difficult

# Feat. Imp. Stats

| Model | Value | data set |
|---|---|---|
| Gleason Score | 70.590153 | cBoost |
| Tissue | 18.635959 | cBoost |
| Preoperative PSA | 8.195472 | cBoost |
| Age | 2.578415 | cBoost |
| Gleason Score | 0.402625 | pen_cox |
| Preoperative PSA | 0.139578 | pen_cox |
| Tissue.FFPE | -0.057436 | pen_cox |
| Tissue.Snap frozen | 0.128038 | pen_cox |
| Tissue | 0.015558 | rsf |
| Gleason Score | 0.122606 | rsf |
| Preoperative PSA | 0.024001 | rsf |

| Model | data set | |
|---|---|---|
| rsf | pData | 3 |
| cBoost | pData | 4 |
| pen_cox | pData | 4 |
| prioLasso | block_data | 22 |
| pen_cox | Intersection | 37 |
| pen_cox | pData_Intersection | 39 |
| pen_cox | Imputed | 43 |
| pen_cox | pData_Imputed | 44 |
| cBoost | Intersection | 80 |
| cBoost | Imputed | 162 |
| cBoost | pData_Imputed | 183 |
| cBoost | pData_Intersection | 210 |
| rsf | Intersection | 987 |
| rsf | pData_Intersection | 1086 |

| Gene | Count | Max. rank | Min. rank | Coeff. Pen. CoxPH | Combinded logHR in Meta analy. of [2] |
|---|---|---|---|---|---|
| TPX2 | 2 | 72 | 26 | 0.06 | 0.48 |
| FKBP10 | 3 | 912 | 3 | 0.01 | 0.32 |