

Introduction

Glaucoma

- Primary open-angle glaucoma (POAG) is the leading cause of irreversible blindness worldwide, characterized by damage to the optic nerve.
- Although POAG progression is slow, it may not show symptoms until significant damage has occurred.
- The exact cause of glaucoma is not fully understood, leading to difficulties in early detection and diagnosis.

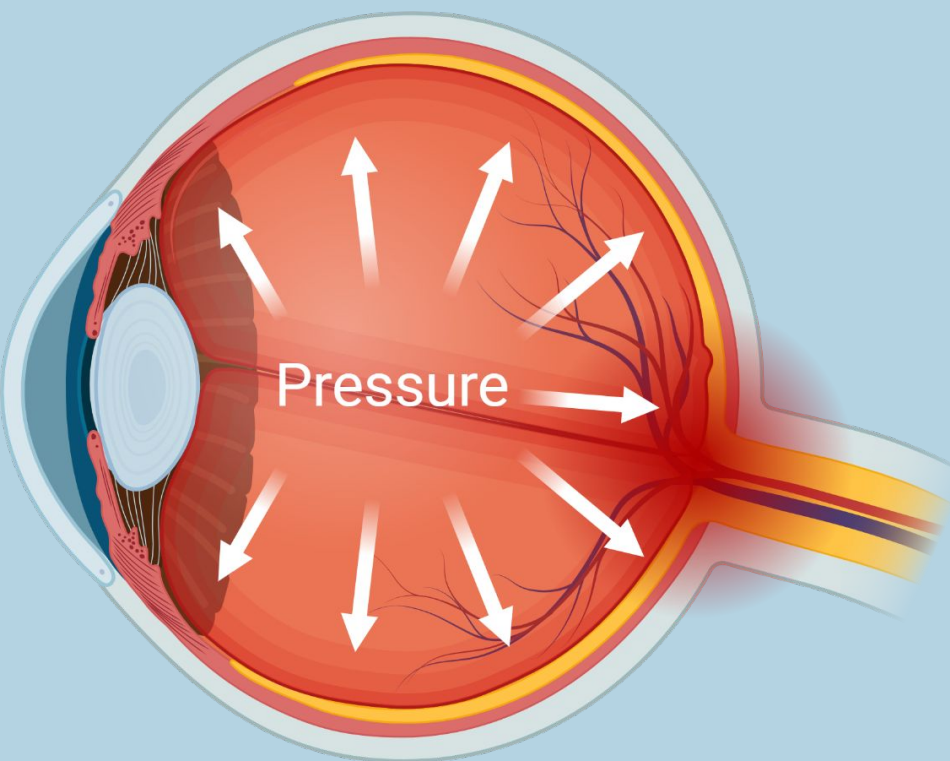


Figure 1. High intraocular pressure can lead to glaucoma.

How can data science help?

Machine learning models can recognize patterns in high-dimensional phenotypic and genotypic patient data to identify individuals at high risk for developing glaucoma.

Objective

We aimed to create a model that not only predicts glaucoma risk, but also identifies relationships between risk factors, enabling the development of personalized interventions based on the predictions of the model.

Study Cohort

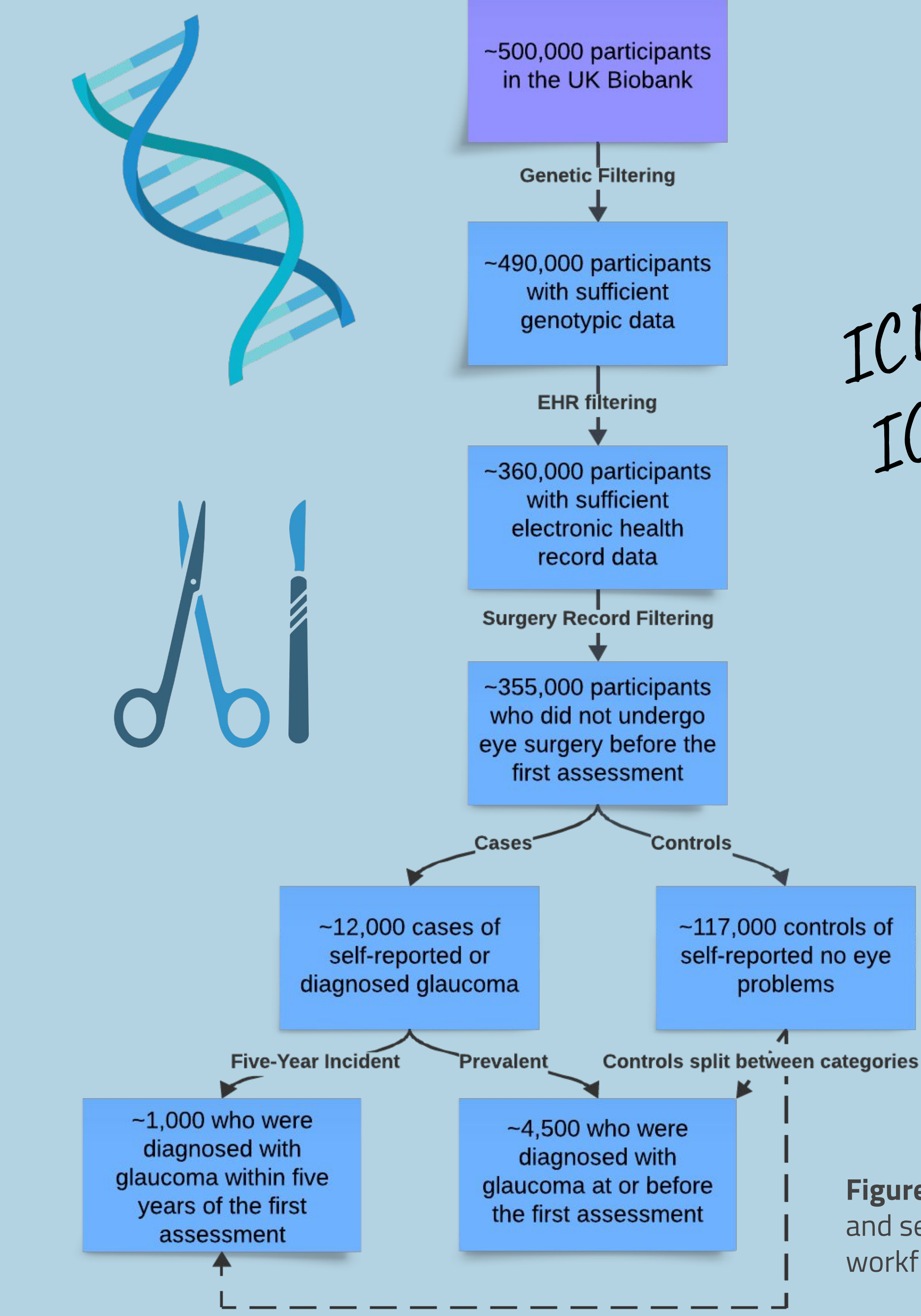
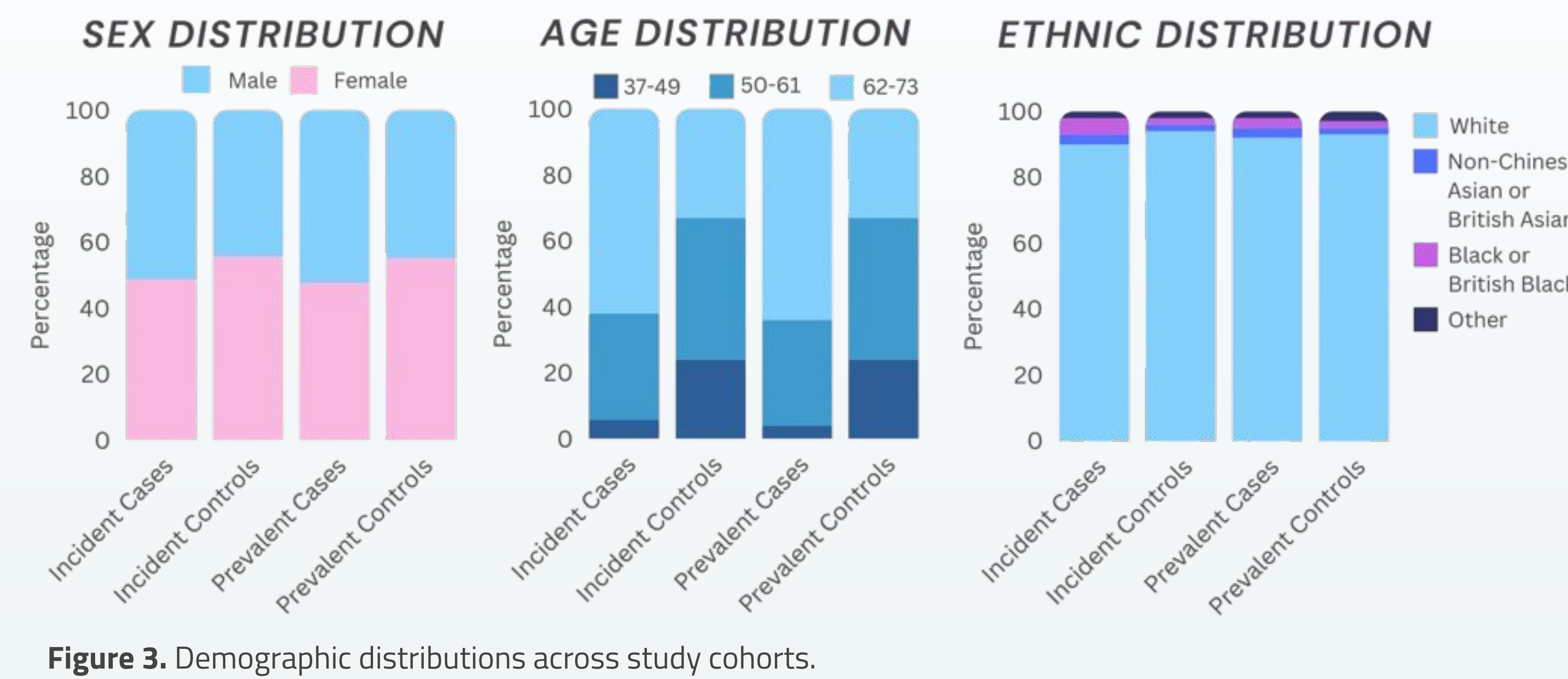


Figure 2. Data filtering and segmentation workflow.<sup>[1, 2]</sup>



Data Preprocessing

Initial Features

- Phenotypes: 253 phenotypic features were either directly taken from the UK Biobank or synthetic constructed from preexisting UK Biobank features based on expert opinion.<sup>[1]</sup>
- Genotypes: 1,097 polygenic risk scores were calculated from UK Biobank genetic data.

Imputation

- To deal with the high missingness of our data, multiple imputation by chained equations (MICE) was employed using the miceRanger package.
  - MICE uses an iterative process where each variable with missing data is imputed by regressing over the other variables in the dataset until convergence.
- Manual imputation was required for intraocular pressure and smoking variables.

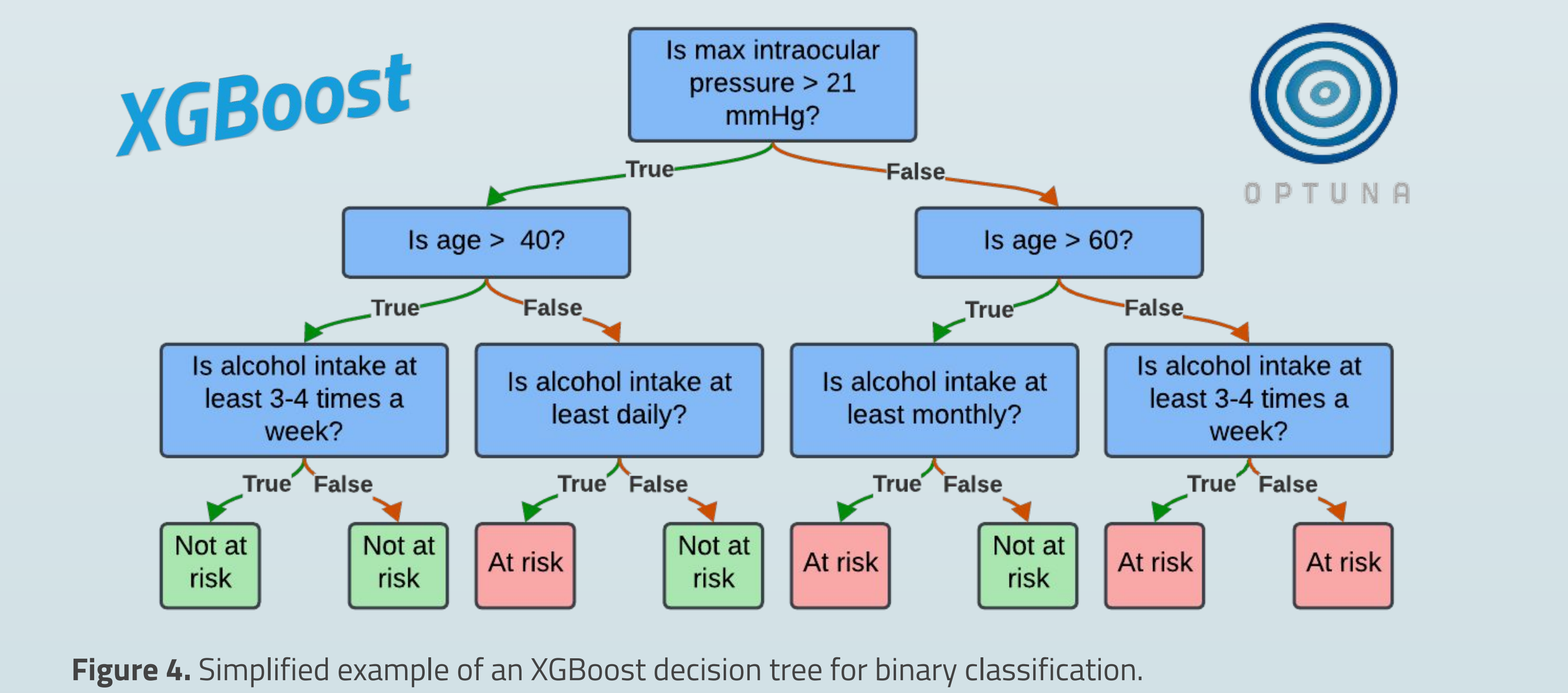
Modeling

Machine Learning with XGBoost

- XGBoost sequentially builds an ensemble of decision trees, where each tree attempts to correct the residual errors of the previous trees, thereby minimizing bias.
- Showed superior performance for predicting coronary artery disease (CAD) incident risk on similar data.
- Uses binary classification to predict whether individuals have five-year incident risk or not.
- Optuna was used for optimized hyperparameter tuning.

Incorporating Meta-Prediction

- Enables us to utilize our data on prevalent cases of glaucoma.



Conclusion

- Our XGBoost models, with and without meta-prediction, narrowly out-perform the baseline logistic regression models.
- Data Limitations**
  - Participants come from a predominantly White background.
  - The lack of five-year incident glaucoma cases makes it difficult for the models to learn the difference between a case and a control.
  - Many features had high missingness, and imputation can cause bias.
- Securing additional data may be required before pursuing risk stratification.

Acknowledgements

I would like to thank Dr. Loguercio, Dr. Torkamani, Dr. Chen, Dr. Nicholson, Dr. Jaiswal, Dr. Khawaja, Dr. Seo, Theresa, Ahmed, Kaushik, Austin, and Shreya for their guidance during my time at the SRTI.

References

[1] Torkamani, A., Chen, S.-F., et al. *Meta- Prediction of Coronary Artery Disease Risk*, 20 Dec. 2023, PREPRINT (Version 1) available at Research Square.

[2] Craig, J.E., et al. *Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression*. Nat Genet 52, 160–166 (2020).