Impact of random oversampling and random undersampling on the development and validation of prediction models using observational health data

PRESENTER: **Cynthia Yang**

INTRO:

- Many datasets used for clinical prediction modelling are imbalanced.
- In the machine learning literature, it has been suggested that developing models using resampled data may improve prediction performance.
- The aim of this study is to empirically investigate the impact of random oversampling and random undersampling on the development and validation of prediction models using observational health data.

METHODS:

1. We used the PatientLevelPrediction (PLP) framework and a sample of 100,000 patients from each database: CCAE, MDCR, MDCD and IQVIA Germany. We investigated 21 binary outcomes within a target population of people suffering from depression.
2. The imbalance ratio (IR) is defined as IR = (# patients who **do not** experience the outcome) / (# patients who **do** experience the outcome). For these tasks, the original IR ($IR_{original}$) ranged from 9.6 to 246.3 with a median of 80.9.
3. We investigated XGBoost and lasso logistic regression. 75% of the data was used for training (including 3-fold cross-validation (CV) for hyperparameter tuning) and the remaining 25% was used for testing. Random sampling was only applied to the training folds, where we varied IR = min($IR_{original}$, x) with x ∈ {20,10,2,1}. We evaluated the area under the receiver operating characteristic curve (AUROC).

Our results suggest that random sampling strategies on average do not improve the prediction performance in terms of AUROC.

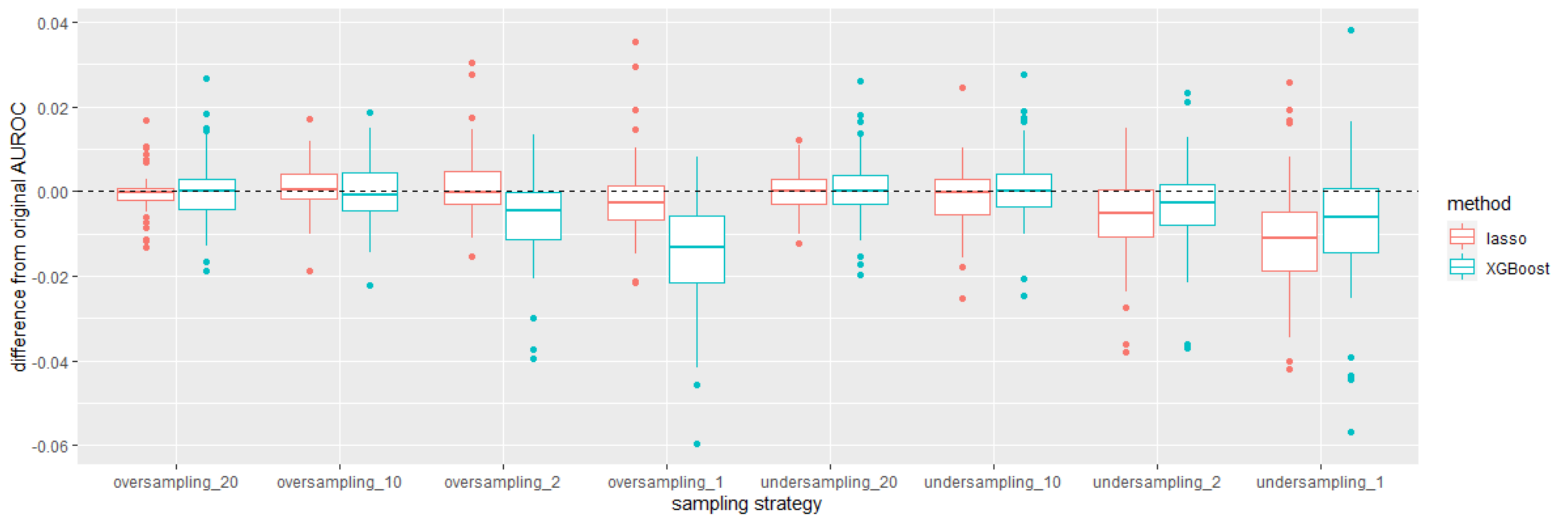

Figure 1. Difference from original AUROC across all prediction problems and all databases per sampling strategy.

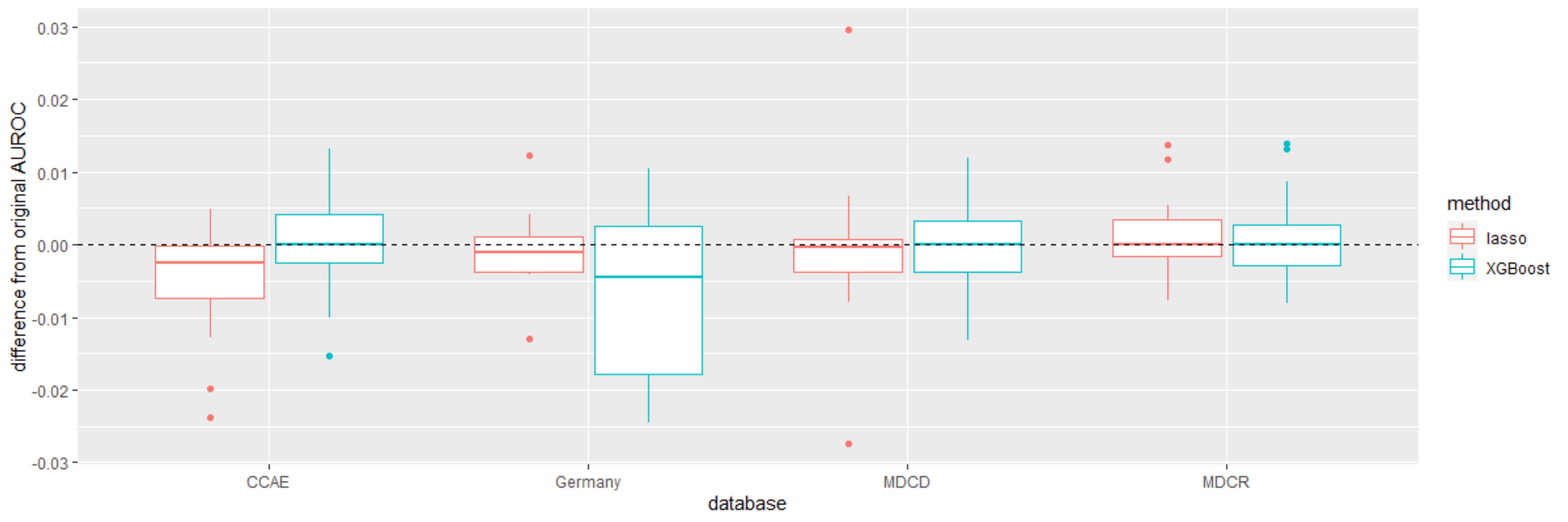

Figure 2. Difference from original AUROC when choosing the sampling strategy based on highest AUROC during CV for each prediction problem per database.

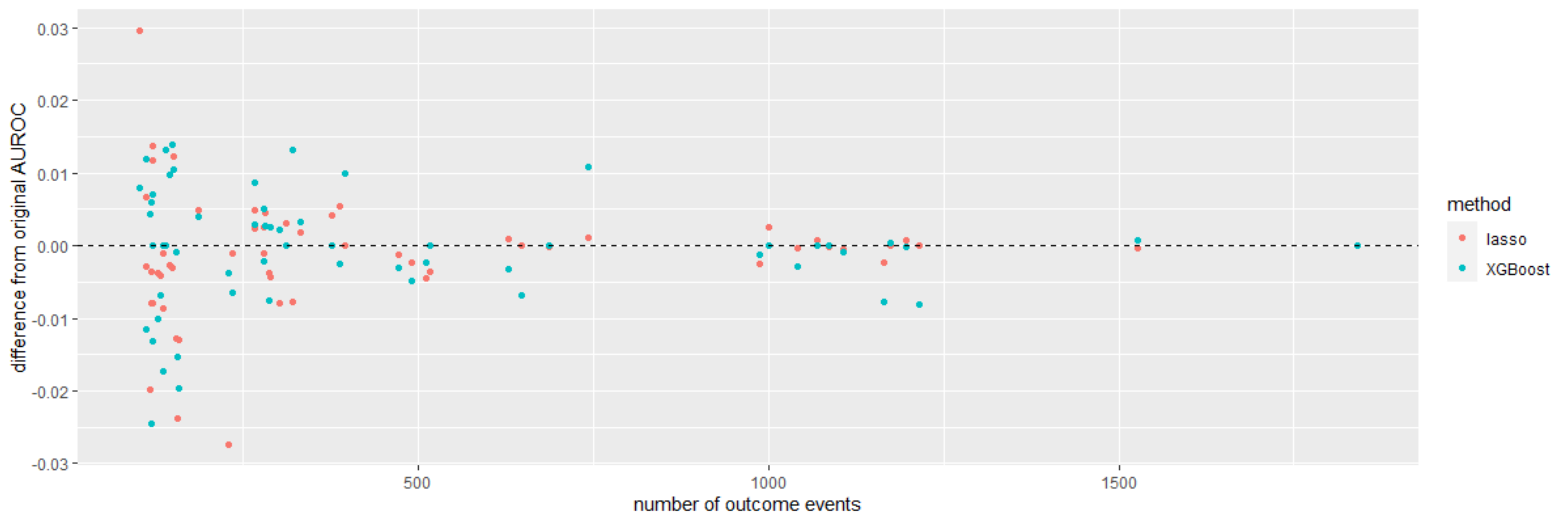

Figure 3. Difference from original AUROC when choosing the sampling strategy based on highest AUROC during CV for each prediction problem by number of outcome events.

RESULTS:

1. Figure 1 shows the difference from the original AUROC (test AUROC with sampling – test AUROC without sampling) across all prediction problems and all databases per sampling strategy. We can see that although there are some cases where a small improvement in test AUROC is found, on average random oversampling and random undersampling do not improve the test AUROC compared to the original setting without sampling. The changes in test AUROC are generally very small, with a maximum absolute difference below 0.06.
2. Figure 2 shows the difference from the original AUROC when choosing the sampling strategy based on highest AUROC during CV for each prediction problem per database. This suggests that if the sampling strategy is considered a hyperparameter during CV, the test AUROC would on average not improve.
3. Figure 3 shows the difference from the original AUROC when choosing the sampling strategy based on highest AUROC during CV for each prediction problem by number of outcome events. Overall, the impact of random sampling on the AUROC shows more variation when the number of outcome events is lower.

Cynthia Yang, MSc,
Egill A. Fridgeirsson, MSc,
Jenna M. Reps, PhD,
Jan A. Kors, PhD,
Peter R. Rijnbeek, PhD

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.