Our Project

Synthetic Health Data Hackathon 2020

REGION Rigshospitalet

Digital Hub Denmark

Erhvervsministeriet

biolib

Our Amazing Team

Stefano Pellegrini

Mahdi Robbani

Jean-Baptiste Van Den Broucke

Challenge 1 and 2 - Diabetes

Methods

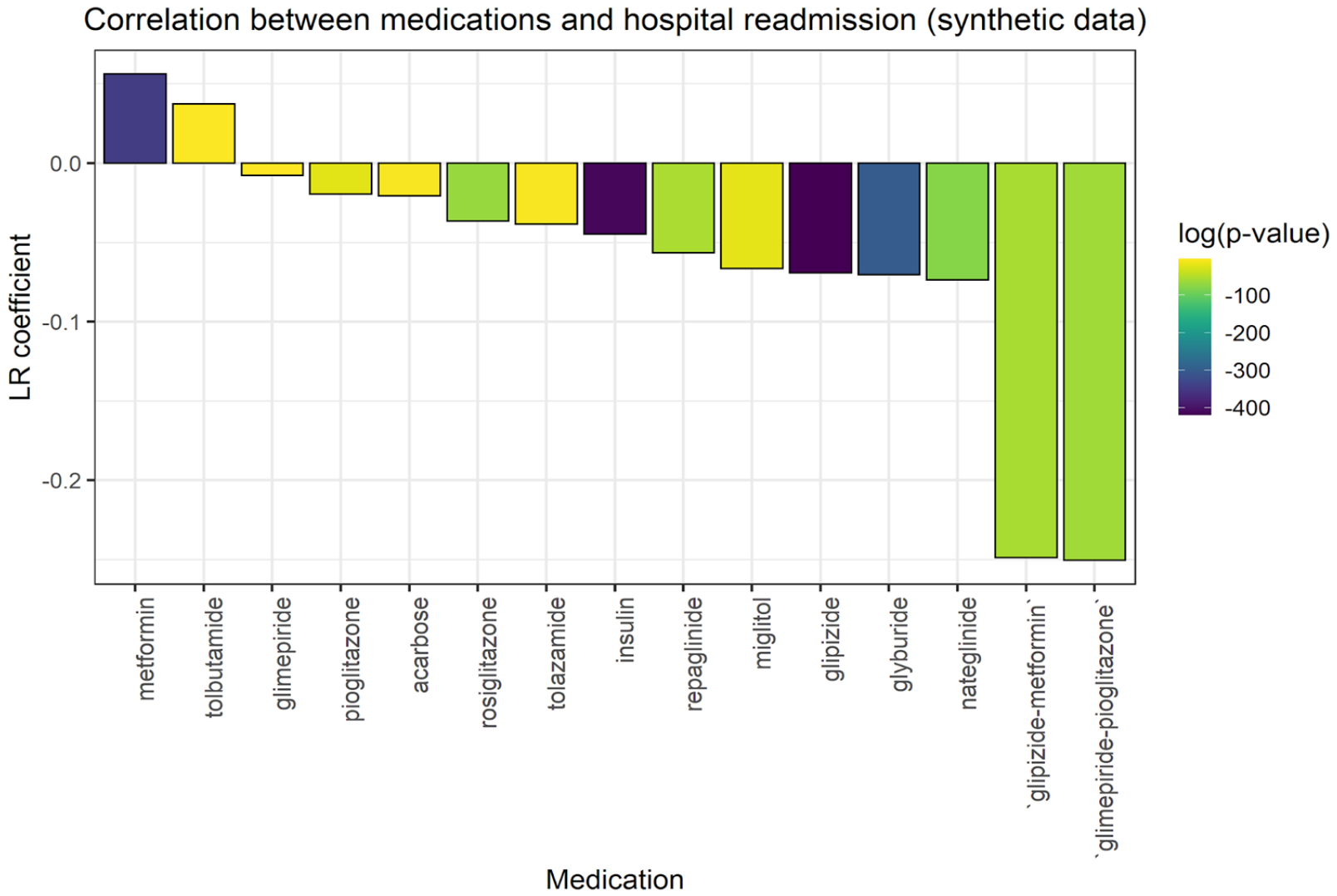
- Estimate the correlation between medication and hospital readmission (**linear regression**)
- Evaluate features importance in both synthetic and real data (**random forest**)
- Compare performance between models trained on real data versus synthetic data (**lightGBM**)

Results

- It is possible to extract biological insights from synthetic data
- Using the synthetic data to predict real data leads to underperformance

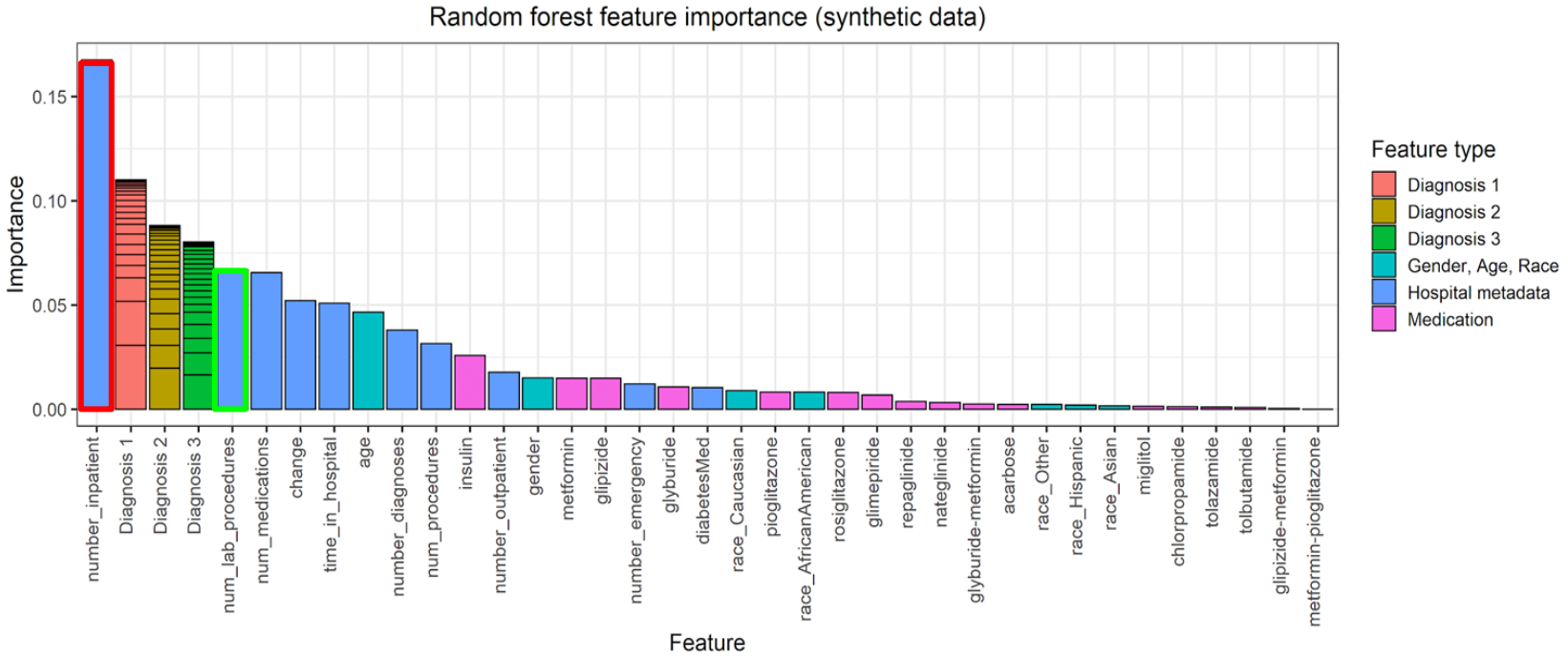
Correlation between medications and hospital readmission in synthetic data (LR)

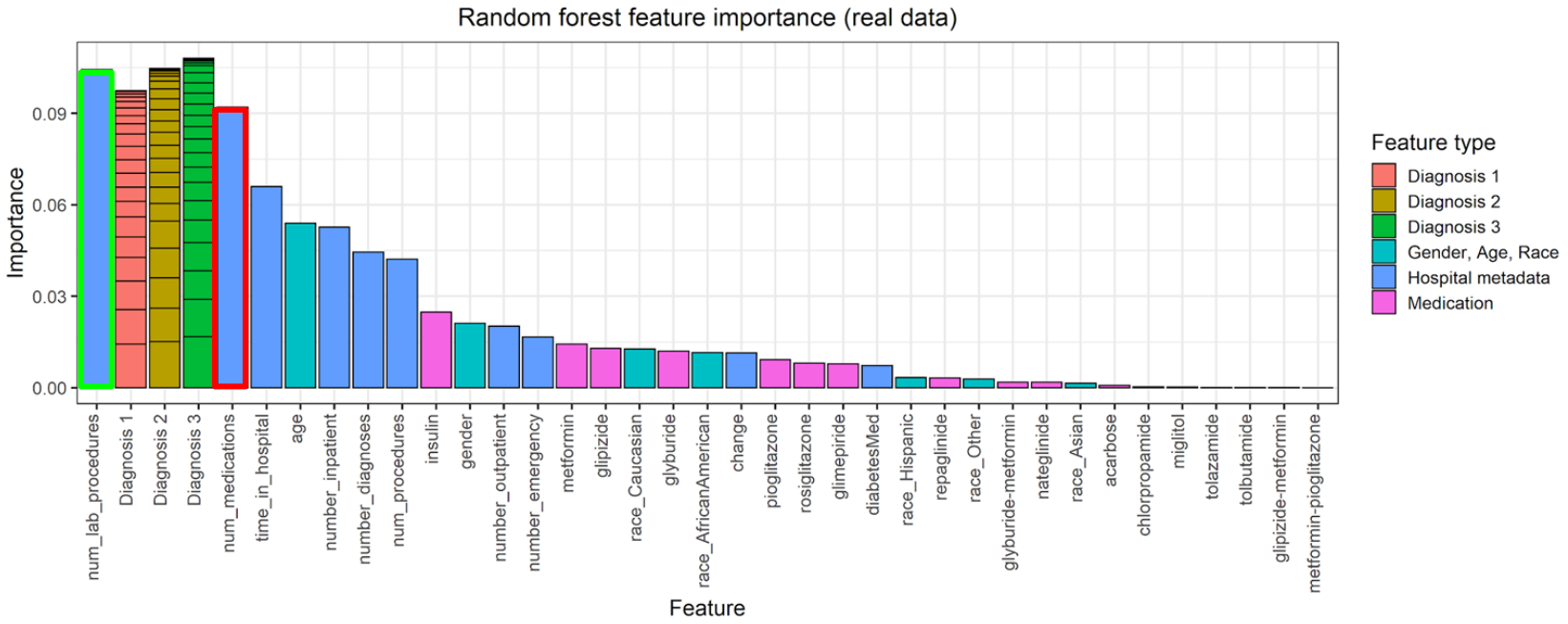
- Metformin shows the highest positive correlation with the target variable
- Insulin and glipizide show the most significant linear relationship



Factors importance for hospital readmission in real and synthetic datasets (RF)

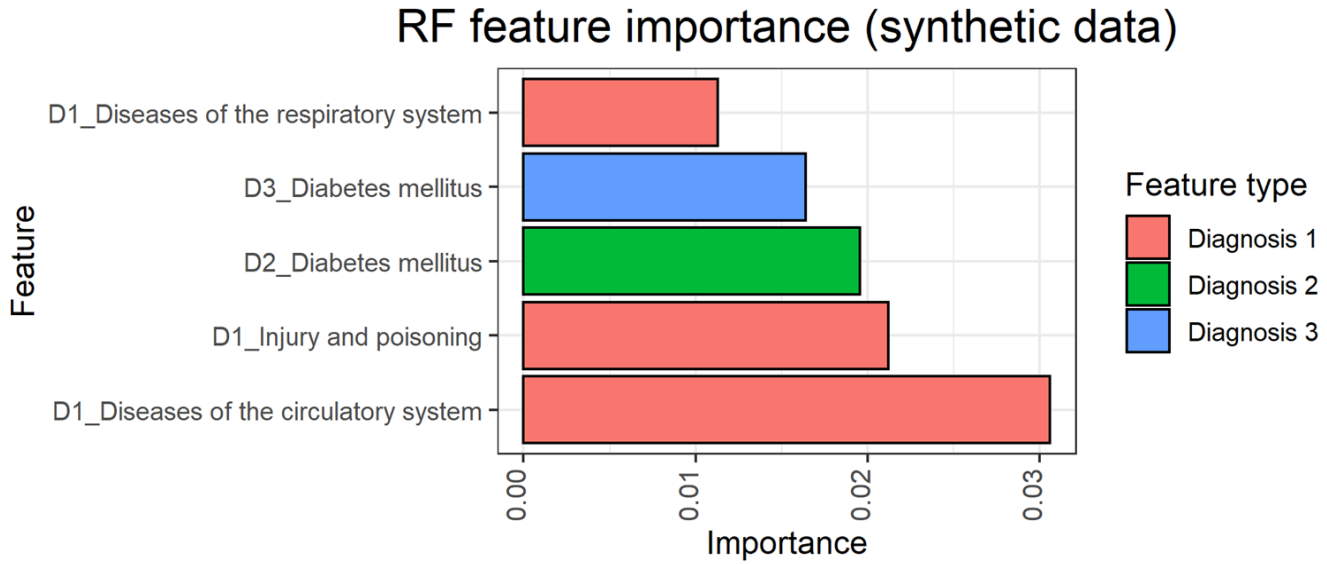
- Overall similarity
- Most important factor for synthetic data is *number_inpatient*
- Most important factor for realdata is *num_lab_procedures*
- *Insulin* is the most important drug

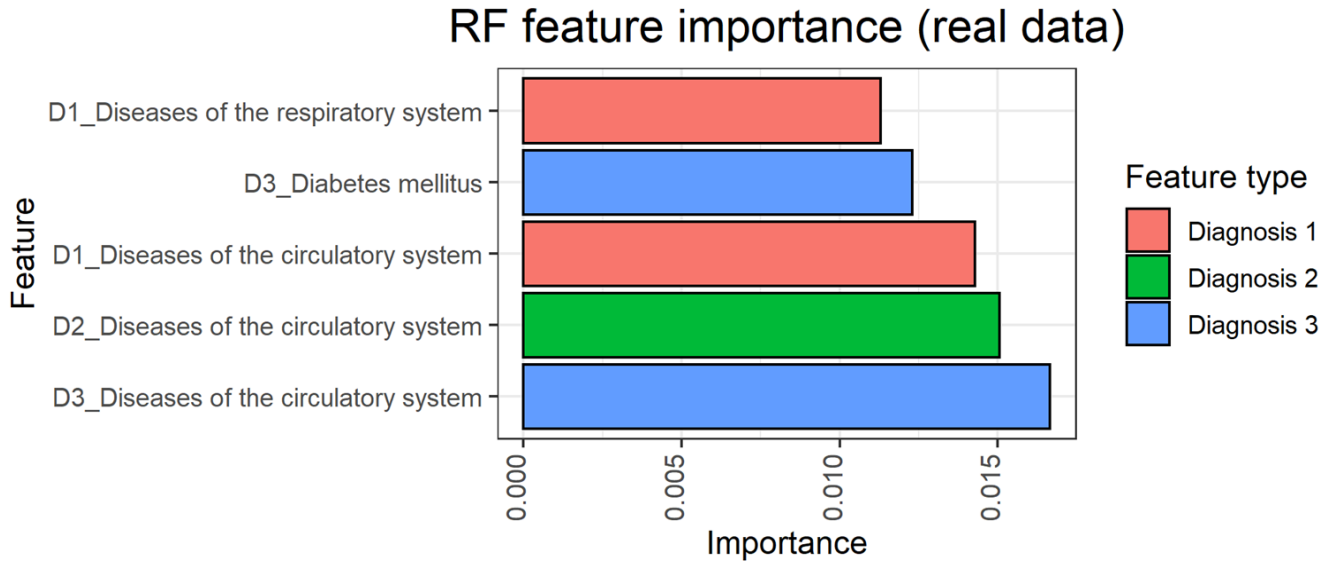




Top 5 important diagnosis in synthetic and real data (RF)

- Overall similarity
- Disease of the circulatory system is the most important diagnosis for hospital readmission





Comparison of model performances (LightGBM)

- Problem simplified to binary classification task
- Training and prediction on synthetic data (blue)
- Training and prediction on on real data (orange)
- Training on synthetic, prediction on real data (green)

	Synthetic	Real	Synthetic → real
AUC	0.89	0.67	0.51
Accuracy	0.81	0.62	0.56