SYNTHETIC DATA HACKATHON 2020

Team TEMB
Bioinformatics track

WHO ARE WE?

Team TEMB

- **T**homas Koefoed (Danish, BSc. Mol. Biomedicine)
- **E**lla Hedeboe (Danish, Mol. BSc. Biomedicine)
- **M**aria Matorra (Spanish, BSc. Genetics)
- **B**rede Sørland (Norwegian, BSc. Bioinformatics)

4 students from University of Copenhagen,
studying a Master's degree in Bioinformatics ☺

OUR APPROACH

Data exploration

Hypotheses

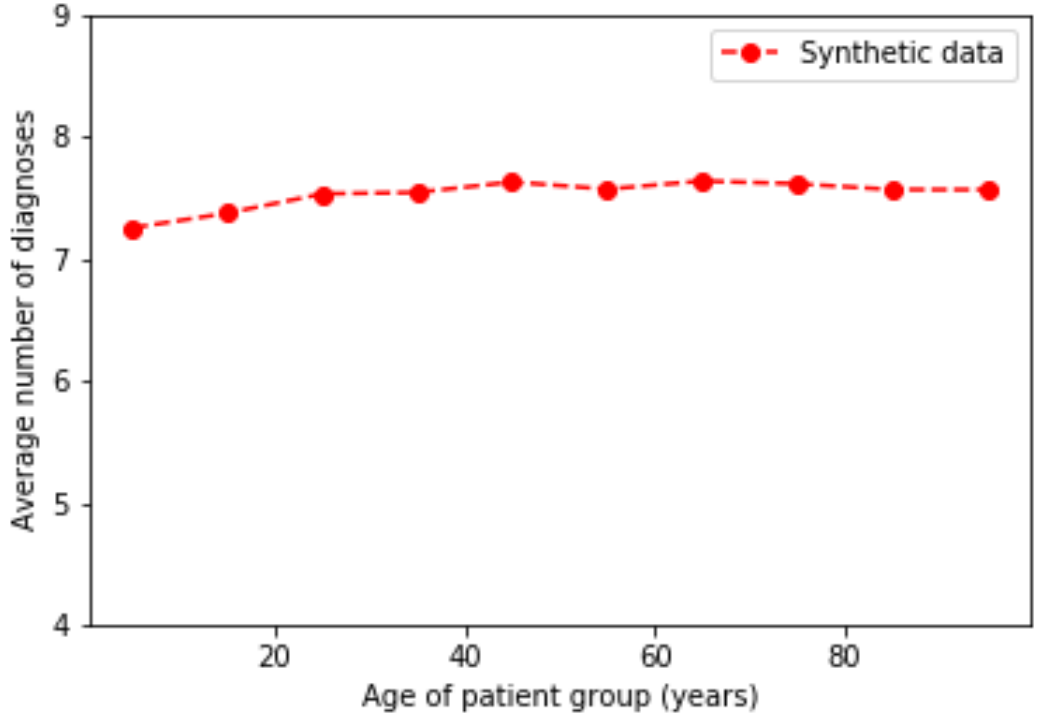
DATA EXPLORATION
INVESTIGATION SHOWS MISSING EXPECTED CORRELATIONS

Investigation
Average number of diagnoses per patient in each age group

Expectation
Strong positive correlation

Observation
No correlation

What about the real, un-synthesized?



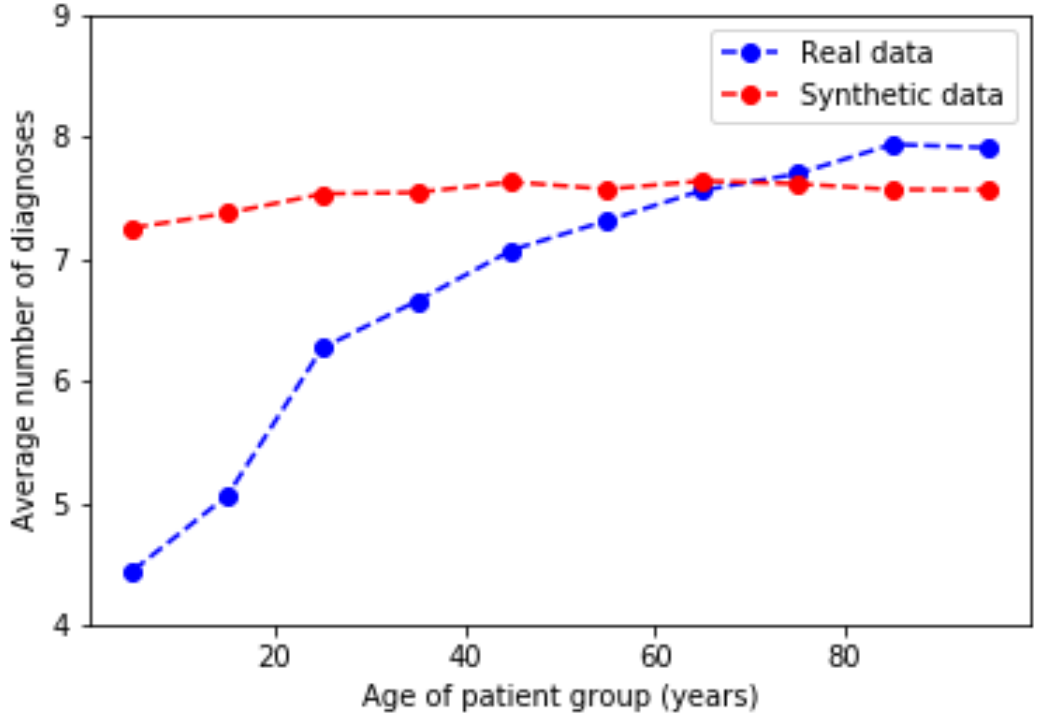
DATA EXPLORATION
INVESTIGATION SHOWS MISSING EXPECTED CORRELATIONS

Investigation
Average number of diagnoses per patient in each age group

Expectation
Strong positive correlation

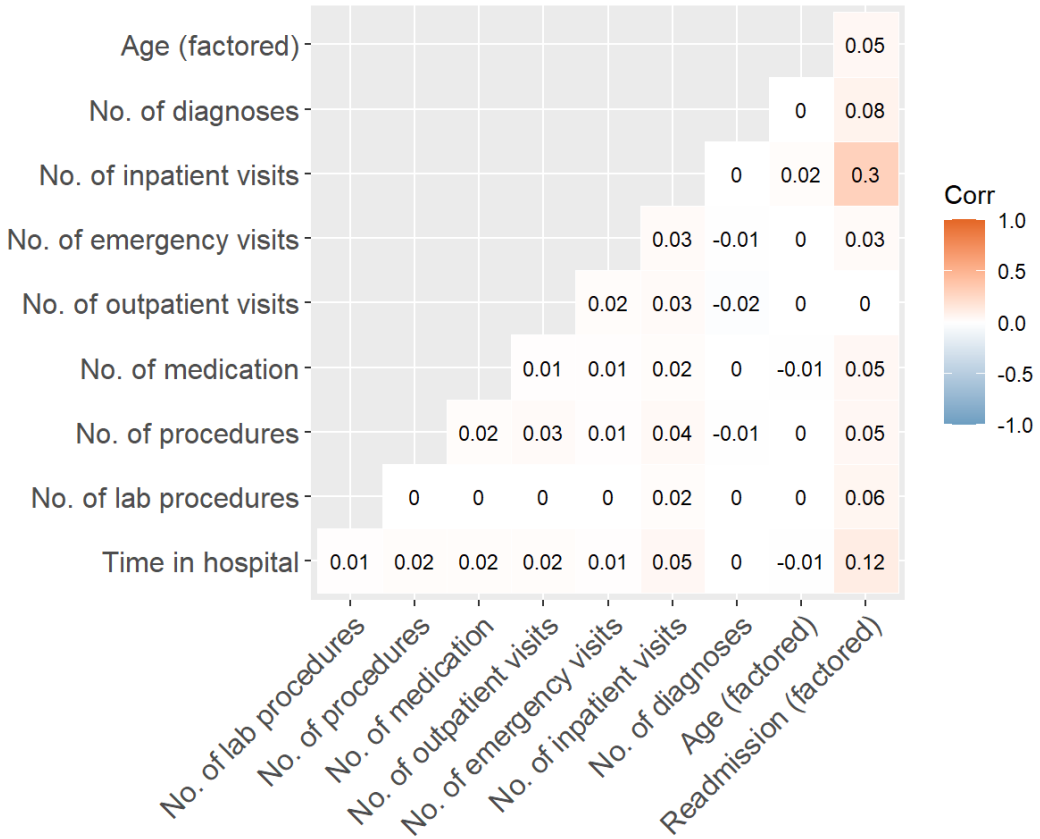
Observation
No correlation

What about the real, un-synthesized?

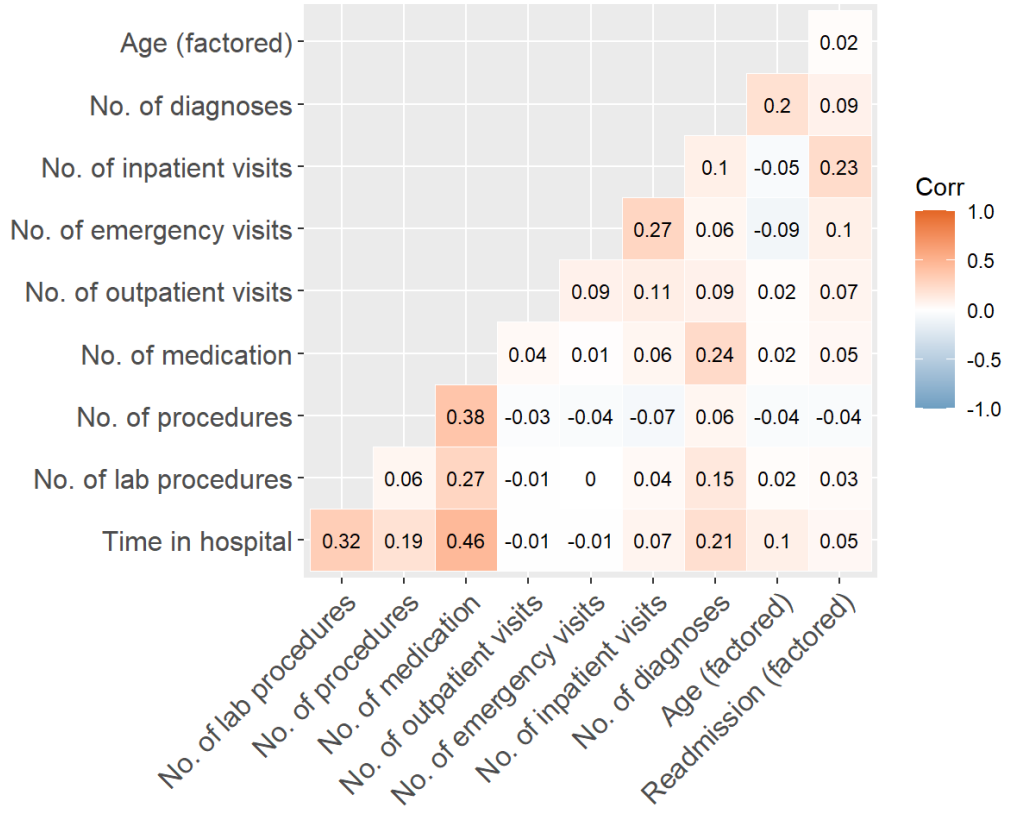


SYNTHETIC VS. REAL DATA SET
CORRELATION MATRIX OF VARIABLES

Synthetic

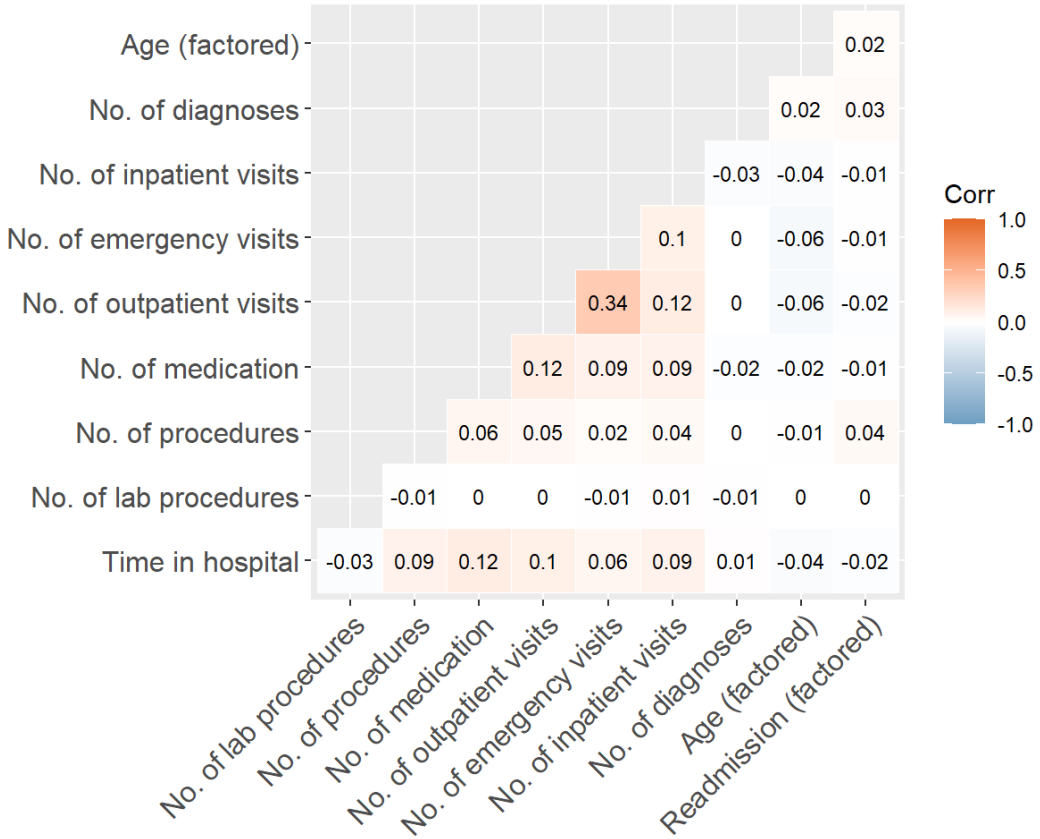


Real

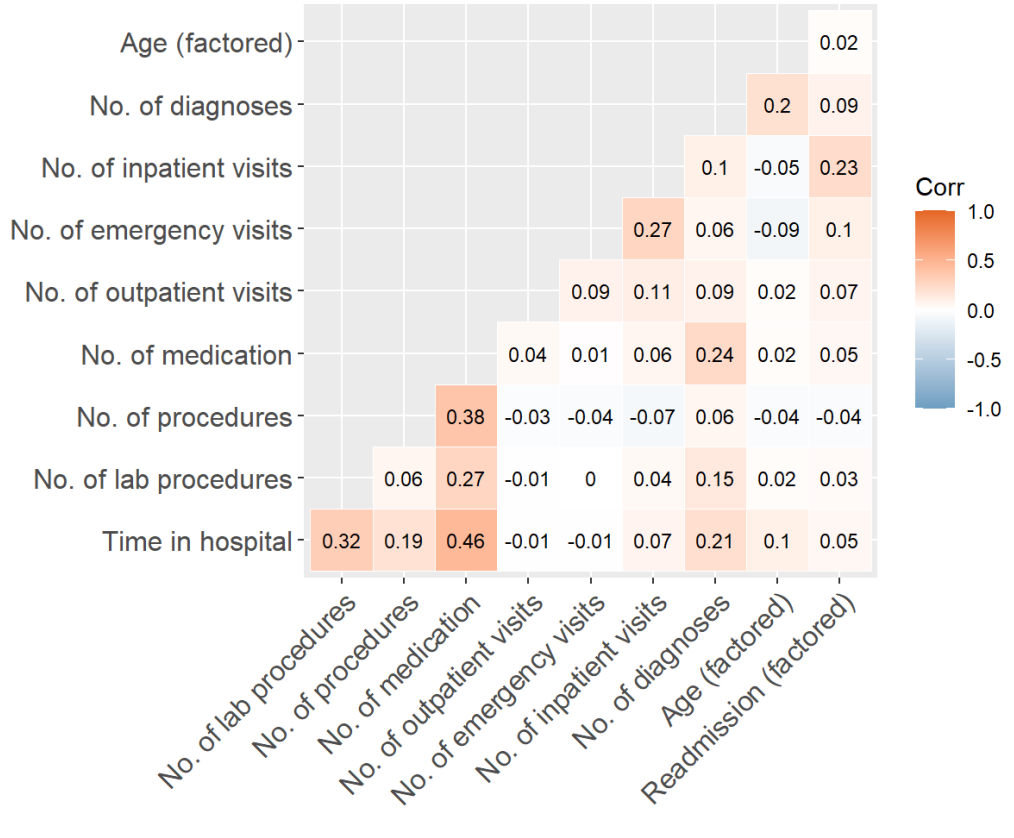


SYNTHETIC VS. REAL DATA SET
CORRELATION MATRIX OF VARIABLES

Synthetic 2.0



Real



SYNTHETIC VS. REAL DATA SET

PREDICTING TIME IN HOSPITAL USING LINEAR REGRESSION

Synthetic, $R^2 = 0.003$

Standardized variables	Weight
Age	-0.02
# lab procedures	0.02
# procedures	0.07
# medications	0.05
# out-patient visits	0.04
# emergency visits	0.04
# in-patient visits	0.13
# diagnoses	0.006
Intercept	4.03

Real, $R^2 = 0.27$

Standardized variables	Weight
Age	0.22
# lab procedures	0.60
# procedures	0.10
# medications	1.12
# out-patient visits	-0.11
# emergency visits	-0.06
# in-patient visits	0.15
# diagnoses	0.22
Intercept	4.42

CONCLUSION

- Synthetic data is a very useful approach, but important signals may be lost
- Quality must be investigated thoroughly before building predictive models

IDEAS FOR FUTURE WORK

- Use linear regression with more, non-numerical variables relevant for diabetes
- Use logistic regression instead of linear regression to predict future re-admission

THANKS TO THE ORGANIZERS FOR HOSTING THIS EVENT!