Improving AI/ML readiness of data generated under the R01: Protein signatures of APOE2 and cognitive aging

Paola Sebastiani, Tufts Medical Center
Ofer Mendelevitch, Syntegra
Eric Reed, Tufts University

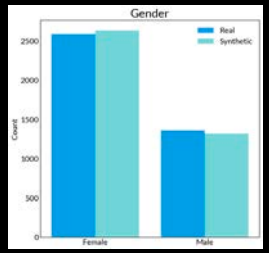
Goal Of The Project

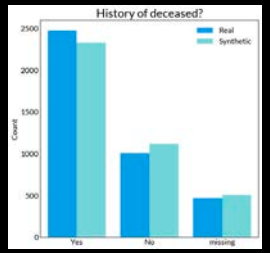
- We work with data from centenarians (ages > 100 years).
- Sharing data with restrictions is not a problem.
- Sharing data with no restriction is not possible: data include HIPAA identifiers, particularly age >89.
- Unrestricted sharing of data would be an attractive option for AI/ML investigators.

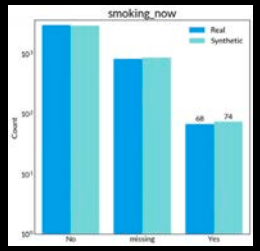
The goal of our project was to use advanced machine learning techniques to generate high-fidelity, privacy-preserving, synthetic versions of the data to be shared without restriction

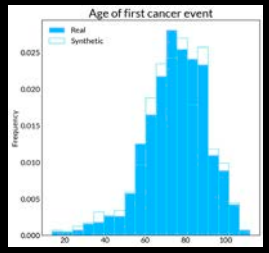
NECS Dataset - Comparing Real to Synthetic

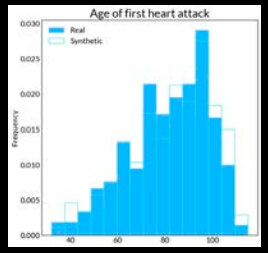
Univariate Distributions and Pairwise Correlations

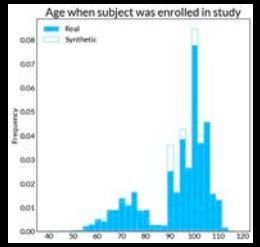


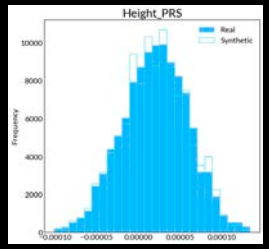


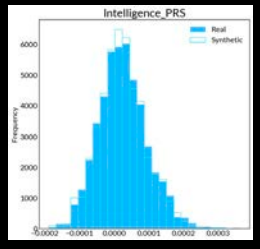


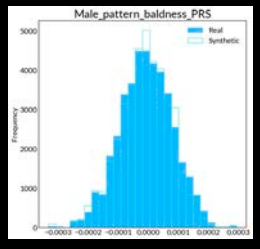












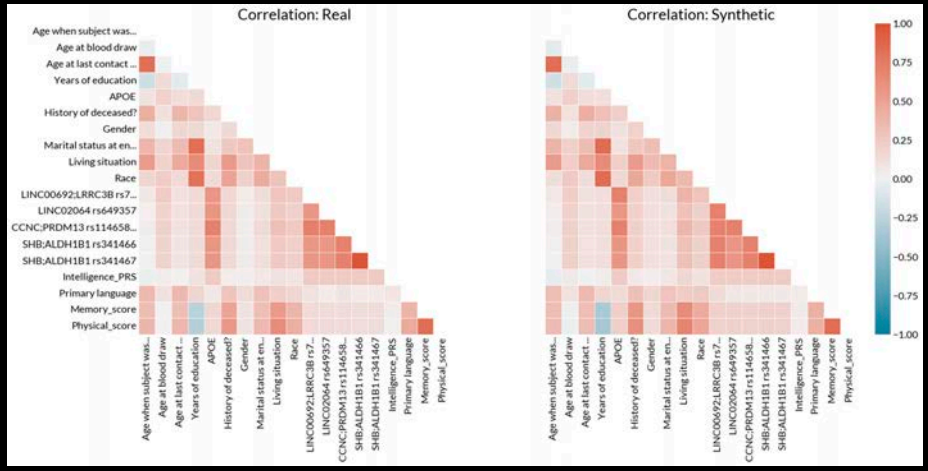
The Dataset consists of 4203 patient records with 230 variables

- Demographics like gender, race, age, years of education, etc
- Clinical outcomes like heart attack, cancer, angina, cataracts, etc along with age of first diagnosis
- Other clinical information like height, weight, BMI, blood pressure
- Genetic information: APOE as well as 34 SNPs and 54 PRS
- Various proteomics variables

Only ~200 patients have proteomics data

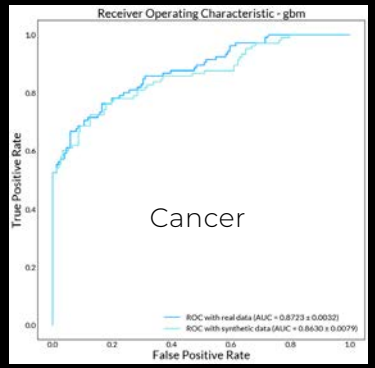
Only ~2000 patients have genomics data

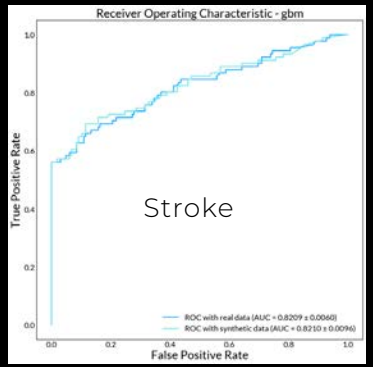
Synthetic

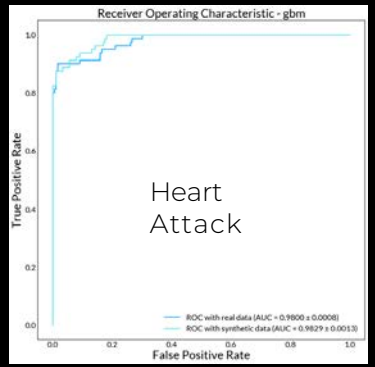


NECS Dataset - Comparing Real to Synthetic

Predictive Modeling and Survival Analysis

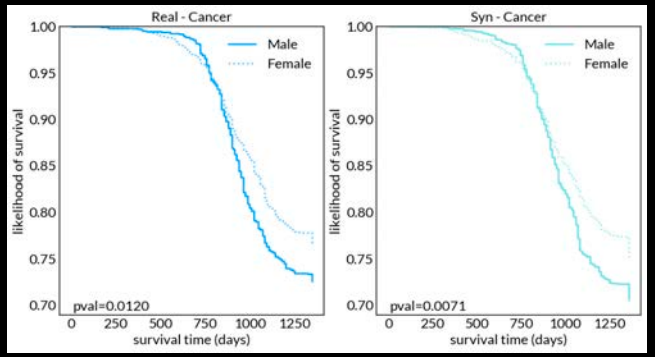




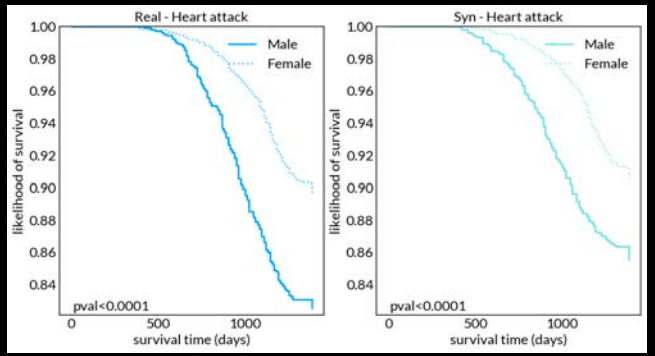


Predictive Models

Synthetic



Survival Analysis



NECS Dataset - Privacy Metrics

Disclosure risk is low

Disclosure risk from attribute inference attack simulation (select variables)

Variable	Depression	Cancer	APOE	rs78043944	rs114658003	rs341466
Disclosure risk	1.3%	0.4%	0.9%	0.46%	0.45%	1%

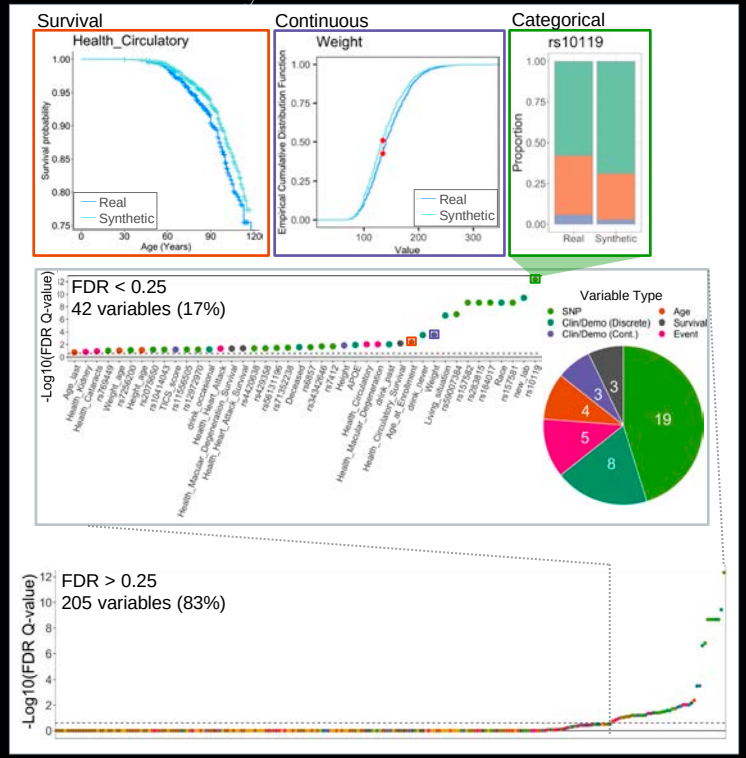
Maximum disclosure risk for dataset = 1.58% (age_last)
Total/Aggregate disclosure risk: 1.62%

In healthcare it is common to consider disclosure risk < 5% is “very low risk”

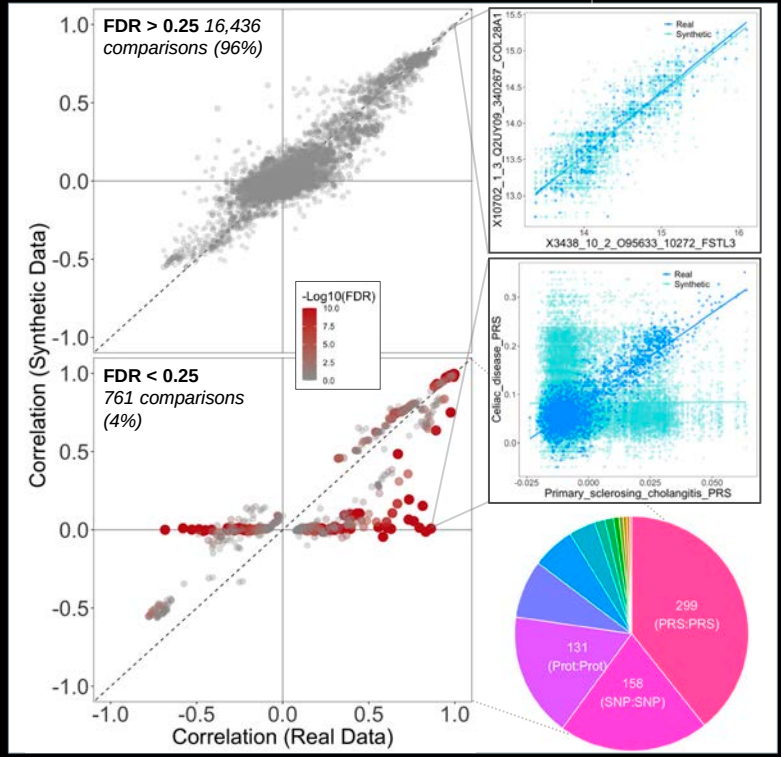
NECS Dataset - Comparing Real to Synthetic

Independent Validation

Probability Distribution



Pairwise Correlation Comparisons



Discussion

- Generally, univariate distributions and variable dependences were conserved
  - Especially for phenotypic variables
  - Exceptions include subset of genetic and proteomics variables

- Analytical challenges include
  - Exhaustiveness
  - Decision Making (Pass/Fail)

Ongoing and future work

- Syntegra is generating a new version of the data
  - Correcting some initial data coding issues/errors

- Some students in our lab are using the data for methods development and evaluation

- We plan to share the validated data through the ELITE program of the AMP-AD knowledge portal