# Vironix Problem Statement

2025 - Math Problems in Industry Workshop

Machine learning models for generation of synthetic electronic health records are getting increased attention as methods to supplement real data for training models (sometimes referred to as downstream models) for predicting chronic disease progression. We must ensure that the synthetic data meets the quality required to ensure the fidelity of downstream models. Traditionally two approaches have been used to judge the quality of the synthetic data:

1. Statistical fidelity measures - For example deception rate, α-Precision, β-Recall, and Authenticity (see [1])
2. Blinded Clinician Evaluation

In a blinded clinician evaluation, a dataset is created by randomly picking samples from both the training (real) data and the synthetic output from the generative model. The data is with extra information such as patient age, gender, and vital statistics and presented to the clinician in the form of patient profiles. Each clinician is then asked to answer a set of 1-3 questions about each patient profile. Example questions could include:

- Does this patient follow a realistic CKD trajectory? (Y/N)
- On a scale from 1 (least) to 5 (most), how likely do you think it is that this is a real patient?
- On a scale from 1 (least common) to 5 (most common), how common would you say this patient trajectory is among those being measured for chronic kidney disease (CKD)?

A typical evaluation includes 4-6 clinicians independently providing evaluations. Responses are compiled into summary statistics. While these statistics provide some insight, clinician feedback on any given patient has large variance. We would like to improve and standardize the clinician evaluation process. Specifically, identifying better measures of clinical performance of models, and better measures to quantify and interpret physician subjective responses.

Our ultimate goal is to :

a) Improve the quality of synthetic data that generalizes better
b) Incorporate cases that have large deviations from the training and validation data set but can be realistic from a clinical perspective.
c) Understand how much synthetic data can be used while training models for predicting disease progression.
d) Better understand the variability in clinician thinking and the utility in using clinician data as validation.

e) Propose and compare other models for predicting disease progression that could be used to validate/scrutinize our existing ML-predictions.

References:


[1] How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models, Ahmed M. Alaa, Boris van Breugel, Evgeny Saveliev, Mihaela van der Schaarhttps://arxiv.org/abs/2102.08921

[2] Zachary Dana, Ahmed Ammar Naseer, Botros Toro, Sumanth Swaminathan. Integrated Machine Learning and Survival Analysis Modeling for Enhanced Chronic Kidney Disease Risk Stratification, arXiv:2411.10754, 2024.

[3] Benjamin Ballyk. Privacy-Preserving Generative Modelling of Longitudinal Electronic Health Records. MSc Thesis, MMSC, University of Oxford, 2024.

[4] Yuan, Ye, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16010--16021, 2023.