A brief overview of "A semi-supervised adaptive Markov Gaussian embedding process (SAMGEP) for prediction of phenotype event times using electronic health records"

YuriAhuja, Jun Wen, Chuan Hong, Zongqi Xia, Sicong Huang & Tianxi Cai

By

Ismail Benchekroun
Quynh Vu

Department of Statistical Sciences
University of Toronto

December 6, 2022

Supervised by Professor Jessica Gronsbell

Contents

0.1 Introduction and Motivation

The growing availability of Electronic Health Records (EHRs) data collected during care delivery in recent years facilitates the full potential for epidemiological study into the course of diseases based on real-time, patient-centred information at different points in time. Prospective clinical research builds on the systematic analysis of the observable and latent clinical symptoms or characteristics of a patient's health conditions, called phenotyping. However, the majority of existing phenotyping methods focus on identifying binary phenotypes using surrogate features that are often insufficiently represented by the generic diagnostic codes such as International Classification of Diseases (ICD) or Current Procedural Terminology (CPT) codes, which, aside from data sensitivity and ethical issues, limits the scope of the implementation of EHRs to clinical research. A second and more deterrent constraint involves the lack of phenotypic gold standard labels of EHRs that possibly hamper the development of reproducible epidemiological risk prediction models due to the inefficient human resource allocation in manual annotation and structural differentials of datasets. Another common critical pitfall of EHRs data is its sparse nature and frequent evolution of lifelong medical records, preventing accurate identification of multiple relapsing phenotype events.

Recent efforts to get a handle on the instinct drawbacks of EHRs data present a trade-off in the cost-effectiveness between manually annotated standardized labels and large sample sizes. Pioneering works include unsupervised and semi-supervised phenotyping algorithms aimed at predicting breast, lung and colorectal cancer recurrence based on the earliest disease emergence by Chubak et al. or using the average of coded surrogate peak times from well-structured clinical data by Hassett et al. and an elaboration on Hassett's technique using points of the extreme increase in coded surrogates by Uno et al. Other unsupervised computational models mimic the Hidden Markov Models (HMMs), modelling disease statuses as latent states. Disease progression models by Jackson et al., Sukkar et al., and Wang et al. employed the HMM approach to an eclectic range of medical conditions (i.e. aneurysms, Alzheimer's and chronic obstructive pulmonary disease). However, the latent states in these HMM-based models are not always clinically relevant. While those unsupervised and semi-supervised phenotyping methods do not stress the manually standardized label requirement, they rely upon either designated sets of predictive codes or pre-defined latent states, resulting in sensitivity to sparseness. Traditional supervised learning methods can cope with such technical incapacity yet fail to tackle the inter-temporality of surrogate features. On the other hand, Recurrent Neural Networks (RNN) based models, particularly the Reverse Time Attention Models (RETAIN), despite the desired degree of accuracy in predicting high dimensional features of time series data, require massive standardized labels to perform solidly.

Principally, empirical experiments present mounting evidence that even highly accurate predictions of phenotype states at discrete time points often do not turn up in accurate phenotype progress predictions, causing the previous computational phenotyping methods to founder when the target object is the timing of phenotypic events. The report will revolve around the newly introduced semi-supervised adaptive Markov Gaussian embedding process (SAMGEP), which has proven its capability to handle the sparsity and noise of EHRs data and to make precise predictions associated with disease exacerbation by requiring only a minimal number of standardized labels in both simulation and real-world experiments. Even though the report centres on the projection of multiple sclerosis (MS) relapse time and heart failure (HF) diagnosis in terms of age, the prospect of generalizing to other medical conditions is unquestionable thanks to the flexible weighting scheme for labelled and unlabeled sparse EHRs features designed to optimize prediction accuracy and a limited number of observed gold standard event labels required for stable and consistent performance. In particular, the leveraging mechanism of SAMGEP confines the costly data collection to clinically relevant phenotypes event time labels using prior knowledge and allows researchers to include available unlabeled EHR data in the model with its adaptive weighting protocol. This way, researchers can maximize the information gained from the gold standard labels and assess the cumulative risks of labelled EHR features and unlabeled disease phenotypic ones not accounted for in prior knowledge. This property makes SAMGEP-based models robust to model generalization and highly accurate in estimating phenotype states compared to previous methods by various statistical classification metrics. Precise and efficient prediction of relapse time enables on-time allocation of interventions and treatments for the sake of patients' lower risk of disease recurrence or recurrence rates.

0.2 Methods

a Description

SAMGEP combines various statistical tools to ultimately derive a patient-timepoint probability of a phenotype state event for each patient i at timepoint t. Before describing the method further, let us introduce some notation.

- i, j and t to index patient, raw features and discrete time periods
- $C_{i,t}$ denotes p-dimensional raw feature vector observed for all patients
- $Y_{i,t}$ denotes phenotype states observed for a limited set of patients n $\ll$ N
- H_i is log mean healthcare encounter count per month + 1, measuring healthcare utilization

In the paper, the method consists of four parts:

1. Assembling predictive features,

   This first step involves extracting time-dependent features from the raw EHR data. For the MS relapse identification data, a patient's sense of pain, blurred vision, sensation loss, and many more variables at time t were included as predictors. We do not need to worry about over-fitting the model with too many variables because the patient-timepoint embeddings undergo a weighting process such that less informative features will be weighed down to 0.

2. Producing patient-timepoint embeddings,

   The authors reduce the high-dimensional feature vector Cit into dense, low-dimensional feature patient-timepoint embeddings $X_{i,t}$. The computation for $X_{i,t}$ is as below:

   $$X_{i,t} = C_{i,t}WV$$

   for i = 1, ..., N where

   - V is a pre-trained m × p feature embedding matrix
   - W is the square weighting matrix W obtained using L1-regularized linear discriminant analysis maximizing the following equation with gradient ascent

$$D(W) = (\mu_1 - \mu_0)\Sigma_X^{-1}(\mu_1 - \mu_0)^T - \lambda\|W\|_1^1$$

- W_1 is the L_1 norm of W
- λ is a hyperparameter
- $\mu_y = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T_i} X_{i,t} I(Y_{i,t}=y)}{\sum_{i=1}^{N}\sum_{t=1}^{T_i} I(Y_{i,t}=y)}$
- $\Sigma_X = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T_i}(X_{i,t}-\mu_{Y\,i,t})(X_{i,t}-\mu_{Y\,i,t})^T}{\sum_{t=1}^{N} T_i}$
- $y \in \{0,1\}$ is phenotype state.

λ is optimized using five-fold cross-validation, choosing the D(W)-maximizing λ. It should be noted that the L1-penalized LDA is fitted only on the labeled set. The pre-trained feature embeddings were obtained using Singular Value Decomposition (SVD). Once the abstracted, low-dimensional patient embeddings X_i are computed, we can model them as a Gaussian process - conditional on Y_i.

3. Fitting the Markov-Gaussian Process (MGP) *Patient embeddings as a Gaussian Process:* A patient i's embeddings, conditioned on Y_i and H_i, follow a Gaussian process:

$$X_i \mid Y_i \sim GP(\mu_i(t), \Sigma_i(t))$$

with mean dependent on Y_i, H_i and t as below:

$$\mu_i(t) = E[X_{i,t}] = \mu_0(1 - Y_{1,t}) + \mu_1 Y_{1,t} + \mu_H H_i + \mu_Y H H_i Y_{1,t} + \mu_2 t + \mu_3 \log t + \mu_4 Y_{1,t} t + \mu_5 Y_{1,t} \log t$$

The covariance between embedding component k and l is modeled by component-specific standard deviation parameters multiplied by their correlation and scaled by H_i with component-specific scaling coefficients α_k and α_l:

$$Cov(X_{i,t,k}, X_{i,t,l}) = \rho_{kl}\sigma_k\sigma_l \exp((\alpha_l + \alpha_k)H_i)$$

The variance follows the same logic and is equal to:

$$Var[X_{i,t,k}] = \sigma_k^2 \exp(2\alpha_k H_i)$$

Note that the variance and covariance are constant over time. From time t to time t + 1, for each patient-embedding component, the residual is modeled to follow an AR(1) process:

$$E[\epsilon_{i,t,k}|\epsilon_{i,t-1,k}] = r_k\epsilon_{i,t-1,k}$$

The coefficient mapping the previous residual to the current is broken down into an embedding component-specific effect τ_k and a regularization hyperparameter that is constant for all components.

Phenotype State as a Markov Process: Phenotype state Y_i follows a Markov process, in which the probability of the phenotype being present for patient i at time t depends on that patient's phenotype state at the previous time t - 1, as well as on the healthcare utilization of that patient, through the following logistic sigmoid equation:

$$P(Y_{i,t}|Y_{i,(t-1)}) = y_{t-1}, H_i) \equiv \pi_t(y_{t-1}, H_i) \text{ for t > 1 and } P(Y_{i,t} \mid H_i) = \pi_{i_n i_t}(H_i) \text{ for t = 1}$$

where

- $\pi_{init}(H_i) = expit(\lambda_{init}, +\lambda_{H0} H_i)$
- $\pi_t(y_{t-1} \mid H_i) = expit(\lambda_0(1 - y_{t-1}) + \lambda_1 y_{t-1} + \lambda_2 t + \lambda_3 \log t + \lambda_H H_i)$
- $expit(x) = \frac{1}{1+e^{-x}}$

We thus have an expression for Y given previous Y, and we have an expression for X given Y. The critical expression we're concerned with, however, is for Y given X. The authors offer to approximate $P(Y_{i,t} = 1)$:

$$\hat{p}_{i,t} = E[Y_{i,t} | X_{i,(t-1)}, X_{i,t}, X_{i,(t+1)}]$$

$$= \frac{\sum_{u=0}^{1} \sum_{w=0}^{1} P(Y_{i,(t-1)}{=}u) P(Y_{i,t}{=}1 | Y_{i,t-1}{=}u) P(Y_{i,t+1}{=}w | Y_{i,t}{=}1) f_{X|Y}}{\sum_{u=0}^{1} \sum_{v=0}^{1} \sum_{w=0}^{1} P(Y_{i,(t-1)}{=}u) P(Y_{i,t}{=}v | Y_{i,t-1}{=}u) P(Y_{i,t+1}{=}w | Y_{i,t}{=}v) f_{X|Y}}$$

where $f_{X|Y} = f(X_{i,t-1}, X_{i,t}, X_{i,t+1} | Y_{i,t-1}, Y_{i,t}, Y_{i,t+1})$.

The first three terms in the numerator are taken from the Markov process model of $Y_{i,t}$ given $Y_{i,t-1}$, and the f term is the density of the Gaussian feature embeddings conditioned on Y. Within this framework, there are many parameters. The Markov model and Gaussian model parameters are found using MLE on the labeled set. This is the equivalent to running a logistic regression for the Markov model and a generalized least squares for the Gaussian model with first-degree autoregressive residuals. Other parameters like are found using five-fold cross validation maximizing AUROC of $Y_{i,t}$ predictions. Once all parameters are solved for, they serve as the initialized parameters in an Expectation-Maximization (EM) algorithm aimed at estimating Y_{it} for the unlabeled set.

The second step of EM is to impute phenotype status for all patient-timpoints in the unlabeled set using the recently initialized parameters. The imputed phenotype probabilities are stored as $\mathbf{\hat{p}_{sup}}$. Next, we re-compute the MLE of the parameters, this time using the entire set and its imputed $\mathbf{\hat{p}_{sup}}$ as the dependent variable. From this new model, we make predictions on the unlabeled patient-timepoint phenotype stati, which we will refer to as $\mathbf{\hat{p}_{semisup}}$.

4. Combining semi-supervised and supervised probabilities.

   Finally, a weighted sum of the supervised p hat and semi supervised p hat is made to give our ultimate phenotype state probabilities $\mathbf{\hat{p}_{final}} = \alpha \mathbf{\hat{p}_{sup}} + (\mathbf{1} - \alpha) \mathbf{\hat{p}_{semisup}}$. The weighting parameter α is determined using five-fold cross-validation maximizing the AUROC of of $Y_{i,t}$ predictions.

b Key assumptions of the methods

The key assumptions of SAMGEP are those regarding i) the $Y_{i,t}$'s discrete time Markov process and ii) the patient embeddings X_i's Gaussian process.

Discrete time Markov Process assumptions:
Equation

$$P(Y_{i,t} = y \mid Y_{i,1}, ..., , Y_{i,t-1}, H_i) = P(Y_{i,t} = y \mid Y_{i,t-1}, H_i)$$

implies that the probability of phenotype event depends on the previous time's phenotype event, as well as healthcare utilization. As the authors explain, the t and log(t) terms are there to capture temporal effects without over-fitting.

Gaussian patient embeddings assumptions:
Expressions $X_i \mid Y_i \sim GP\big(\mu_i(t), \Sigma_i(t)\big)$ and

$$\mu_i(t) = E(X_{i,t}) = \mu_0(1 - Y_{i,t}) + \mu_1 Y_{i,t} + \mu_H H_i + \mu_{YH} H_i Y_{i,t} + \mu_2 t + \mu_3 log(t) + \mu_4 Y_{i,t} t + \mu_5 Y_{i,t} log(t)$$

summarize the assumptions relating to the Gaussian nature of the patient embeddings. To re-emphasize them concisely, the embeddings of a patient follow a normal distribution whose mean changes with time. This mean is also assumed to depend on phenotype state at time t, healthcare utilization, and time - with interactive effects thereof included. Healthcare utilization is assumed to act as a scaling effect for covariance between embeddings. Finally, the residuals of the embeddings at time t are assumed to be a linear mapping of their value in the previous time t-1. A lag of 1 was decided for the autoregression for computational ease but also to, again, avoid overfitting.

The authors note the normality assumption of the patient embeddings conditioned on phenotype state is justified because the feature embeddings were engineered to follow a multivariate normal distribution. We also assume phenotype state can be known for a small number of patients using laborious chart review, and that the feature vector is known for all patients. An assumption that is not mentioned but made implicit from the method are that time periods t for a given patient i must be evenly spread apart. Another unmentioned assumption is that the labeled set is assumed to be a proper representation of the general sample, meaning that the model covariate parameters in the labeled and unlabeled sets should be the same.

c Comparison and Diagnostics

Unlike LASSO-penalized logistic regression and random forest, SAMGEP models leverage the time sequence nature of the features and outcome measures. Meanwhile, LSTM and RETAIN integrate the time series data into a Neural Network-based architecture, making for a distribution-free but more sophisticated method than SAMGEP. Both SAMGEP and HMM can model a Markov process and esitmate its parameters using an EM algorithm. The difference with SAMGEP is that it initializes the parameters using already-consistent estimates learned from standard supervised learning techniques on the labeled set. Thanks to the leveraged gold-standard labels, the EM algorithm only needs one iteration to reach convergence under SAMGEP methodology.

SAMGEP is ideal for predicting time-specific phenotype states using EHR data. It leverages the vast amounts of features available from EHR data, and the L1 norm regularization for W filters out the uninformative ones (of which there are many in EHR). The method's MGP nature is good because it preserves the time sequence of the embeddings and phenotype state in a structure that is simple relative to neural network-based methods. Such simplicity may be desired since the labelled dataset is small in practice, and thus SAMGEP can learn from the gold-standard labels while being less likely to overfit to them when predicting phenotype state for the unlabeled set.

0.3 Results

The researchers initiated the study with a large-scale programmatic simulation experiment to assess the robustness of SAMGEP to violations of model assumptions at different thresholds of labelled phenotype events. After verifying SAMGEP assumptions and discerning the number of labelled phenotype events optimizing SAMGEP performance (about 50 to 100 labels), the research team then utilized real-world EHRs data sets of size 50, 100, 200 and 500 for MS relapse and size 50, 100, and 200 for HF onset to compare the predictive accuracies between SAMGEP and existing phenotyping methods (i.e. PheCode, LSTM, RETAIN, LDA Embed, and HMM). The key performance metrics used were the scale-invariant AUC (a relative measure of predictions), F1-score (a measure of the model's accuracy on the dataset), and ABC classification (a measure of predictive accuracies in independent, correct, and complex settings). This approach allowed researchers not only to gain an insight into the robustness and correctness of SAMGEP predictions at several levels of abstraction but also to analyze the possibilities for generalization and reproducibility to modelling of diseases other than multiple sclerosis and heart failure without the need for many more real-world data sets, diminishing the cost of conducting research. More importantly, it is not always possible to have data sets that match up to the intended purpose of every analysis, making simulation a satisfactory alternative to real-world ERHs in preliminary analysis.

In the simulation experiment, researchers simulated 150 count features, including ICD codes, RxNorm drug codes, CPT codes, etc., along with the average number of monthly health appointments, denoted by $\mathbf{H}$, for 1000, 5000, and 20000 unlabelled patients as well as 100 labelled patients, each with an expected period 25-time points. The researchers ran simulations with the number of standardized labels varying from 5 to 100. SAMGEP was proven to outshine deep learning methods (i.e. RETAIN and LSTM) if $Y_{i,t} \mid Y_{i,t-1}, H_i$ follows a Markov process and $\mathbf{X}|\mathbf{Y}$ follows a Gaussian one, where $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{T}$ denote, respectively, binary phenotype states, weighted sum over feature embeddings and the patient's case history length. That said, SAMGEP still performed relatively better than deep learning methods if $\mathbf{Y}$ was independent of $\mathbf{T}$ (i.e. $\mathbf{Y} \perp \mathbf{T}$) or if $\mathbf{Y}(\mathbf{T})$ was over-parametrized, though lower predictive accuracy is inevitable if simulation data diverge from SAMGEP assumptions, as evidenced by the AUC, F score, ABC_{cdf} gain, and ABC_{count} gain metrics in Panels A and B (Figure 1 on page x). Moreover, the L_1 regularized weighting mechanism fixed in SAMGEP reduced the high dimensionality due to the sparsity in ERHs data, allowing it to outperform unsupervised learning methods (i.e. HMMs and LDA_{Embed}) when using as little as five trained labels and deep learning methods when using the medium number of 20 train labels as indicated in Panel C (Figure 1 on page x). In the real-world experiment, SAMGEP modelled MS relapse of 4706 patients and HF onset of 59395 patients, each of whom had at least one ICD code. The predictions for MS relapse and HF onset either outperformed or worked relatively as well as previous methods across key metrics, especially for a small number of labelled phenotype features. It is noteworthy that SAMGEP achieved the highest accuracy in predicting MS relapse timing even in the large data setting (n > 100), a favourable environment for deep learning methods, implying SAMGEP successfully modelled phenotype events as a process rather than as discrete time points (Figure 2 on page xi).

The simulation setup is justifiable since it accurately imitates real-world situations and represents an actual EHR data collection process with real parameters (e.g. number of feature counts, diagnostic codes, number of patients, number of standardized labels, etc.) while allowing the flexibility to replicate this process with varying parameter values. Moreover, the SAMGEP assumptions stress the importance of $Y_{i,t} \mid -Y_{i,t-1}, H_i$

being a stochastic Markov process, meaning that the event a patient suffers an MS relapse or an HF onset, given their last hospitalization and number of visits (i.e. the seriousness of patient's physical health condition) randomly occurs. Here, the real EHRs data align with SAMGEP assumption since $\mathbf{Y}|\mathbf{H}$ for MS relapse was defined as the time points new or recurrent MS symptoms randomly occurred, and HF onset dates since the last visit were collected from randomly chosen patients. On the other hand, one can verify the assumption that $\mathbf{X}|\mathbf{T}$ follows a Gaussian process by using tests for normality such as Quantile-Quantile plot or Shapiro-Wilk test on a finite collection of patients. Above all, though more works are ongoing to minimize the number of labels required, the SAMGEP phenotyping algorithm has closed the "loophole" of previous computational phenotyping algorithms by lowering the cost to obtain gold-standard labels, reducing the dependence on sets of designated codes or pre-defined latent states, and, most importantly, eradicating the sensitivity to the inherent sparseness in ERHs data of unsupervised and semi-supervised phenotyping methods.

0.4 Conclusion

Plenty of methods have been developed to determine overall phenotype status or at a certain time, but few have focused on phenotype status over multiple timepoints. Out of the approaches that do model a time sequence, some supervised learning methods have high accuracy but require reliable labels throughout the EHR data. The unsupervised learning methods are free of this constraint, but are not guaranteed to be estimating the right parameter of interest - which is, the probability of phenotype being present over time. SAMGEP, on the other hand, offers a practical solution to this, using supervised learning on the labeled set of data to then impute labels onto the remaining data. The authors' approach achieves this through two key distributional assumptions on $Y_{i,t} \mid Y_{t-1}, H_i$ and on $X_i \mid Y_i$. The model's performance was compared to that of previous methods, under various settings. In the cases of correct model specification and smaller datasets, SAMGEP outperformed all other methods. LSTM and RETAIN become better options as the labeled set increases in sample size, and also as the true model distribution gets farther from the MGP.

Even though only a small proportion of the data may have gold-standard labels, incorporating these few labels improves model performance against complex deep learning tools. SAMGEP is most successful in leveraging limited gold-standard labels when in a longitudinal context. Few gold-standard labels and longitudinal data are the exact settings in which the SAMGEP was intended to add value as those are when neural network and standard supervised learning methods fall short. To wit, we can conclude that SAMGEP is successful at improving upon previous work methods in these specific settings.

The results were thorough in showing SAMGEP's performance relative to alternatives using i) various metrics and ii) various data settings. We see that it is not an overwhelmingly superior method to RETAIN or HMM, but that it stays competitive with them even in less-than-ideal settings like when the Markov model is mis-specified. In the settings for which the model was said to be helpful, ie few gold-standard labels and few informative features and longitudinal data, SAMGEP outperformed all other methods across multiple metrics. To wit, we can conclude that SAMGEP delivers its promise to more efficiently use EHR data than any of the previous work. This at least warrants further work to be done on SAMGEP.

This paper's contribution is very practical. No individual part of the method is groundbreaking; it is simply a combination of statistical and machine learning tools that each address shortcomings of the EHR phenotyping process (e.g. sparsely important features, expensive gold-standard labels, longitudinal data, etc.) Its most substantial contribution is the method's ability to make efficient use of few gold-standard labels. No semi-supervised method on longitudinal data has done this, and the sophisticated supervised learning methods require more than a few such labels to achieve good performance. This is particularly important because manually annotating these labels is an expensive task. While the authors point out current limitations of SAMGEP such as non-robustness to covariate shift between labeled and unlabeled sets, they offer potential solutions that could be worth implementing in the future. For now, the method is promising in that it makes use of the vast information offered by EHR data to determine phenotype status over time with relatively good accuracy.

0.5 References

1. Tayefi, M, Ngo, P, Chomutare, T, et al. Challenges and opportunities beyond structured data in analysis of electronic health records. WIREs Comput Stat. 2021; 13:e1549. https://doi.org/10.1002/wics.1549

2. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, Goldman S, Janmohamed S, Kreuzer J, Leenay M, Michel A, Ong S, Pell JP, Southworth MR, Stough WG, Thoenes M, Zannad F, Zalewski A. Electronic health records to facilitate clinical research. Clin Res Cardiol. 2017 Jan;106(1):1-9. doi: 10.1007/s00392-016-1025-6. Epub 2016 Aug 24. PMID: 27557678; PMCID: PMC5226988.

3. Ajami S, Arab-Chadegani R. Barriers to implement Electronic Health Records (EHRs). Mater Sociomed. 2013;25(3):213-5. doi: 10.5455/msm.2013.25.213-215. PMID: 24167440; PMCID: PMC3804410.

4. Morris, TP, White, IR, Crowther, MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019; 38: 2074– 2102. https://doi.org/10.1002/sim.8086

5. Batra, S., Sachdeva, S. (2021). Pre-Processing Highly Sparse and Frequently Evolving Standardized Electronic Health Records for Mining. In G. Rani, P. Tiwari (Ed.), Handbook of Research on Disease Prediction Through Data Analytics and Machine Learning (pp. 8-21). IGI Global. https://doi.org/10.4018/978-1-7998-2742-9.ch002

0.6 Figures

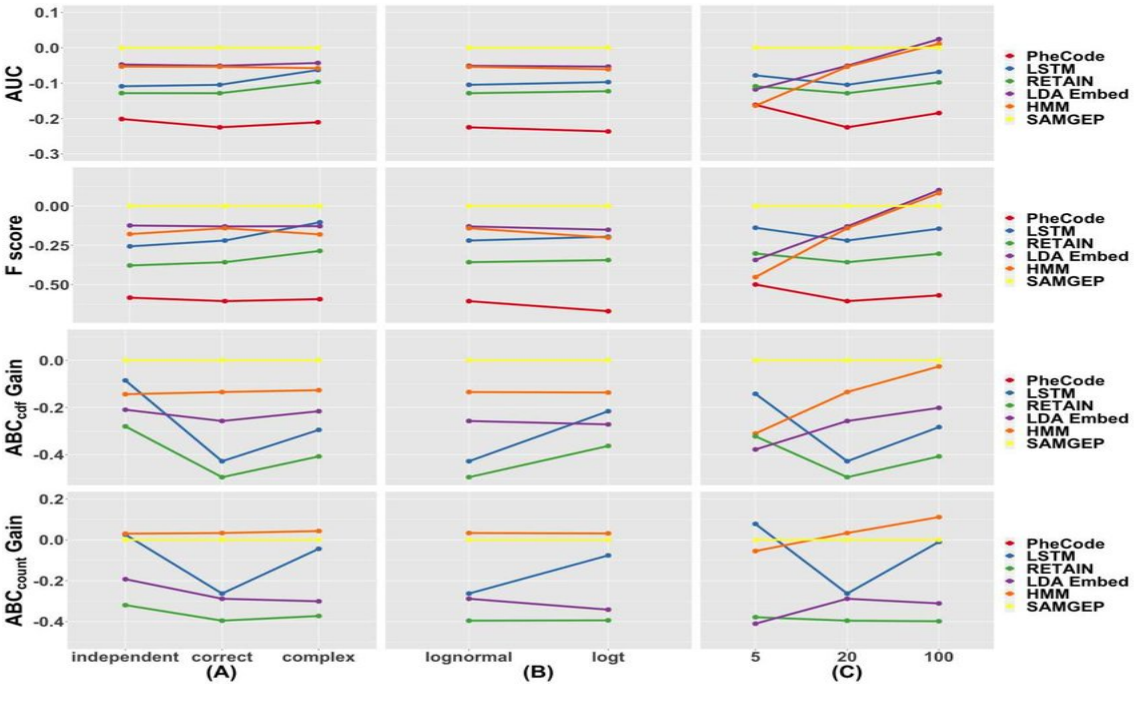
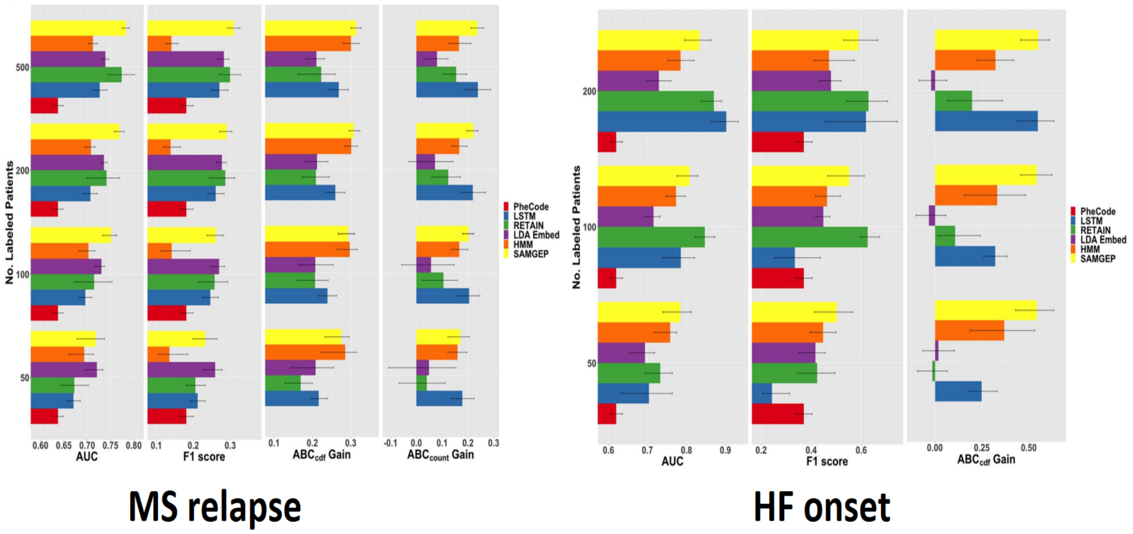


Figure 1: Robustness of SAMGEP and comparator methods' AUCs, F scores, ABC_{cdf} gains, and ABC_{count} gains to various generative parameters



MS relapse

HF onset

Figure 2: Predictive accuracies of SAMGEP and various comparator methods using real-world EHR data

0.7 Contributions

- Introduction and Motivation - Quynh Vu
- Methods - Ismail Benchekroun
- Result - Quynh Vu
- Conclusion - Ismail Benchekroun