SAMGEP: A Novel Method for Prediction of Phenotype Event Times Using the Electronic Health Record

Yuri Ahuja (yuri_ahuja@hms.harvard.edu)
Harvard T.H. Chan School of Public Health
Jun Wen
Harvard T.H. Chan School of Public Health
Chuan Hong
Harvard Medical School
Zongqi Xia
University of Pittsburgh
Sicong Huang
Harvard Medical School
Tianxi Cai ScD
Harvard T.H. Chan School of Public Health

Research Article

Keywords: phenotyping, phenotypic big data, electronic health records, precision medicine, time series prediction

Posted Date: December 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1119858/v1

SAMGEP: A Novel Method for Prediction of Phenotype Event Times Using the Electronic Health Record

Yuri Ahuja PhD[1,2*], Jun Wen PhD[1*], Chuan Hong PhD[2], Zongqi Xia MD PhD[3], Sicong Huang MD[2,4,5], Tianxi Cai ScD[1,2,5]

[1] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[2] Harvard Medical School, Boston, MA, USA
[3] Department of Neurology, University of Pittsburgh, Pittsburgh, PA, USA
[4] Division of Rheumatology, Inflammation, and Immunity, Brigham and Women's Hospital, Boston, MA, USA
[5] VA Boston Healthcare System, Boston, MA, USA

*These authors contributed equally

Corresponding Author: Yuri Ahuja, 677 Huntington Ave, Boston, MA 02115; yuri_ahuja@hms.harvard.edu; (917)-297-2421

Keywords: phenotyping, phenotypic big data, electronic health records, precision medicine, time series prediction

Word Count: 4025

ABSTRACT

While there exist numerous methods to identify binary phenotypes (i.e. COPD) using electronic health record (EHR) data, few exist to ascertain the timings of phenotype events (i.e. COPD onset or exacerbations). Estimating event times could enable more powerful use of EHR data for longitudinal risk modeling, including survival analysis. Here we introduce Semi-supervised Adaptive Markov Gaussian Embedding Process (SAMGEP), a semi-supervised machine learning algorithm to estimate phenotype event times using EHR data with limited observed labels, which require resource-intensive chart review to obtain. SAMGEP models latent phenotype states as a binary Markov process, and it employs an adaptive weighting strategy to map timestamped EHR features to an embedding function that it models as a state-dependent Gaussian process. SAMGEP's feature weighting achieves meaningful feature selection, and its predictions significantly improve AUCs and F1 scores over existing approaches in diverse simulations and real-world settings. It is particularly adept at predicting cumulative risk and event counting process functions, and is robust to diverse generative model parameters. Moreover, it achieves high accuracy with few (50-100) labels, efficiently leveraging unlabeled EHR data to maximize information gain from costly-to-obtain event time labels. SAMGEP can be used to estimate accurate phenotype state functions for risk modeling research.

INTRODUCTION

Electronic Health Record (EHR) data collected during the routine delivery of care have in recent years enabled countless opportunities for translational and clinical research.[1-3] Comprising freeform clinical notes, lab results, prescriptions, and codified features including International Classification of Diseases (ICD) and Current Procedural Terminology (CPT) billing codes, EHRs encode rich information for research. However, EHRs' lack of gold-standard phenotype labels limits utilization of these data to precisely estimate epidemiological parameters such as prevalence and treatment effects, or to develop and validate well-calibrated risk prediction models for clinical events. Phenotype surrogate features such as ICD diagnosis codes often exhibit dismal specificity that can bias or de-power the downstream study.[4,5] Meanwhile, manual annotation of phenotypes via chart review is laborious and unscalable. These limitations become even more pronounced when the object of interest is the *timing* of clinical events, which is important for survival analysis or evaluating disease course. Event time surrogates derived from EHR codes often exhibit systematic biases, and multiple features may be needed to accurately predict timing.[6-8]

For binary phenotypes, researchers have proposed a variety of unsupervised and semi-supervised methods requiring few-to-no manually-annotated gold-standard labels.[9-20] However, few methods exist to predict phenotype event times. Chubak et al. developed a rule-based algorithm that predicts breast cancer recurrence time based on the earliest encounter times of expert-specified codes.[8] Hassett et al. proposed a similar algorithm averaging the peak times of selected codes.[7] Uno et al. expanded on this by using points of maximal increase in lieu of peak values, and adjusting for systematic temporal biases between code timings and phenotype onset.[6] While these approaches achieve notable performance, they are limited by 1) reliance on a limited, curated set of predictive codes, and 2) sensitivity to sparsity, a common characteristic of EHR data. In addition, these algorithms cannot identify multiple event times as might be pertinent for a relapsing and remitting phenotype such as multiple sclerosis.

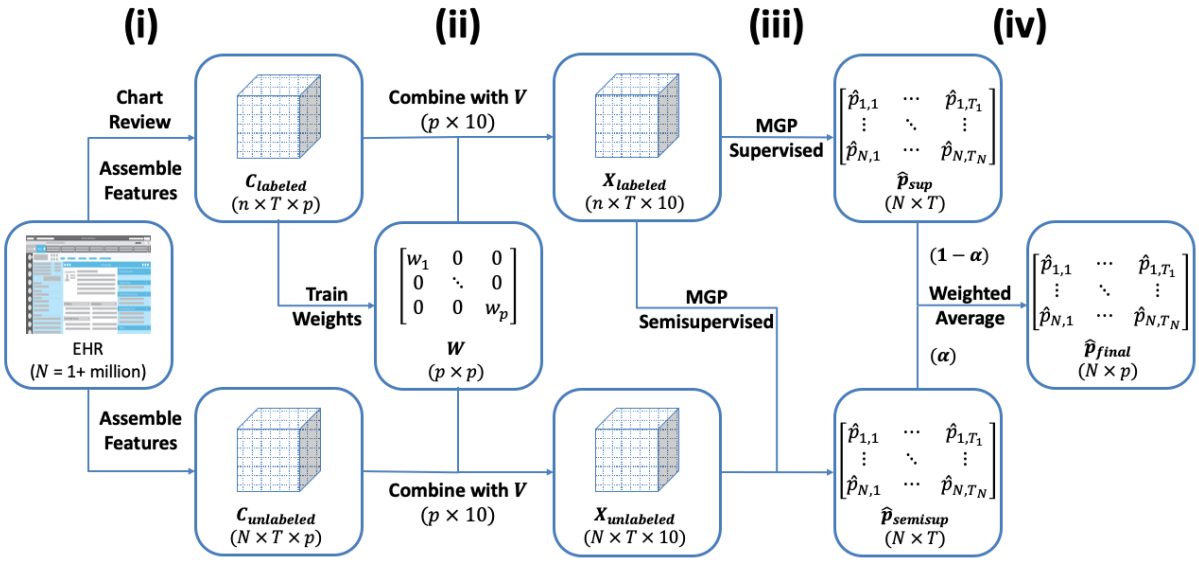
Using machine learning to predict event times can potentially address these limitations. Traditional supervised learning methods such as logistic regression, random forest, and naive Bayes are

suboptimal for modeling longitudinal processes as they cannot account for intertemporal associations in either outcomes or features. Recurrent neural networks (RNNs), designed for sequence data and well-conditioned to high feature dimensions, have enjoyed widespread application to prediction using longitudinal data.[21-25] One recent RNN-based method optimized for healthcare data, the Reverse Time Attention Model, (RETAIN), offers particularly notable accuracy to this end.[26] However, neural networks often require large numbers of training labels to achieve stable performance, which can be very expensive to attain. Consequently, existing applications of RNNs to EHR-based prediction typically use readily available outcome measures such as discharge billing codes, hindering application to outcomes without reliable codified proxies. Additionally, apart from RETAIN these models are generally not intuitively interpretable.

On the other end of the spectrum, researchers have developed unsupervised computational models of chronic disease progression that do not use any gold-standard labels.[26-31] Many of these approaches employ Hidden Markov Models (HMMs) in which latent states represent disease stages or status. For instance, Jackson et al. apply a multistage discrete HMM to aneurysm screening, Sukkar et al. apply one to Alzheimer's disease, and Wang et al. apply a continuous HMM to progression of chronic obstructive pulmonary disease.[26-29] However, the latent states learned from these unsupervised models may not be reflective of the target phenotype or even clinically relevant.

In this paper we propose Semi-supervised Adaptive Markov Gaussian Embedding Process (SAMGEP), a label-efficient, semi-supervised machine learning method to predict the presence of a well-defined binary phenotype over time using longitudinal EHR data. By leveraging large-scale unlabeled data to train the temporal prediction model, SAMGEP makes efficient use of limited (~50-100) gold-standard event labels. Unlike existing event identification algorithms, SAMGEP can leverage hundreds of sparse EHR features rather than a handful of surrogates by combining features and their embeddings into dense patient-timepoint embeddings via a novel data-driven weighting procedure. It then models the patient-timepoint embedding progression as a Gaussian Process emission of an HMM to predict the target phenotype, combining desirable aspects of existing phenotyping methods.

(a)



(b)

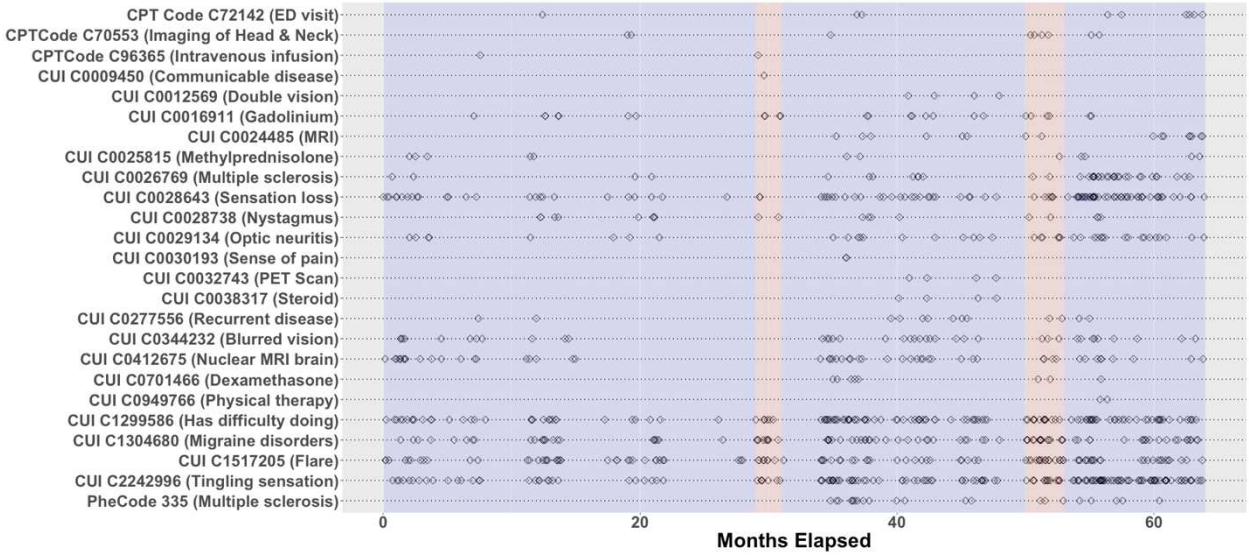


Figure 1: (a) Schematic of the overall SAMGEP algorithm. (b) Depiction of the sparsity and temporal irregularity of EHR data. In this study we aim to predict MS relapse event times (red bands) using timestamped EHR feature observations (black diamonds).

RESULTS

Model Overview

SAMGEP predicts well-defined (i.e. not new) binary phenotype processes from longitudinal EHR data with few observed gold-standard event labels. It does so in four steps: (i) assembling time-dependent candidate features, (ii) optimizing weights for combining feature embeddings into dense patient-timepoint embeddings, (iii) fitting supervised and semi-supervised Markov Gaussian Process (MGP) models to the embedding process to predict phenotype status over time, and (iv) taking a weighted average of these semi-supervised and supervised predictions with weights determined adaptively to optimize prediction performance. Figure 1A illustrates the overarching SAMGEP procedure, and Figure 1B depicts the form of raw longitudinal EHR data for input into SAMGEP. A detailed description of the SAMGEP procedure is included in the *Methods*. R source code is available at https://cran.r-project.org/web/packages/SAMGEP/index.html.

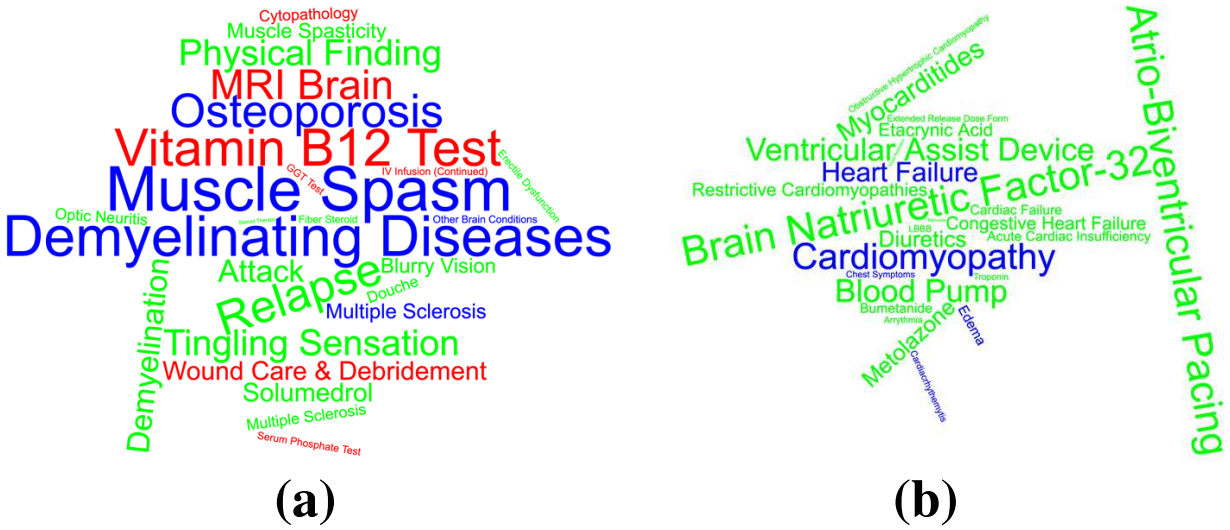


Figure 2: Feature word clouds for (a) MS relapse and (b) HF onset using the product of SAMGEP's feature weights and the empirical standard deviations of corresponding features. See the *Evaluation Metrics* subsection of the Methods for details.

Feature Selection

Figure 2 depicts feature clouds generated using SAMGEP's feature weights for identification of (A) MS relapse and (B) HF onset. SAMGEP identified PheCodes for demyelinating diseases and muscle

spasm, CPT codes for vitamin B12 testing and MRI brain, and CUIs for "relapse" and "tingling sensation" as most predictive of MS relapse. For identification of HF onset, SAMGEP selected PheCodes for heart failure and cardiomyopathy, and CUIs for "brain natriuretic factor 32" and "atrio-biventricular pacing." In both cases, SAMGEP reassuringly selected and upweighted appropriate, clincially relevant features for the respective outcomes.

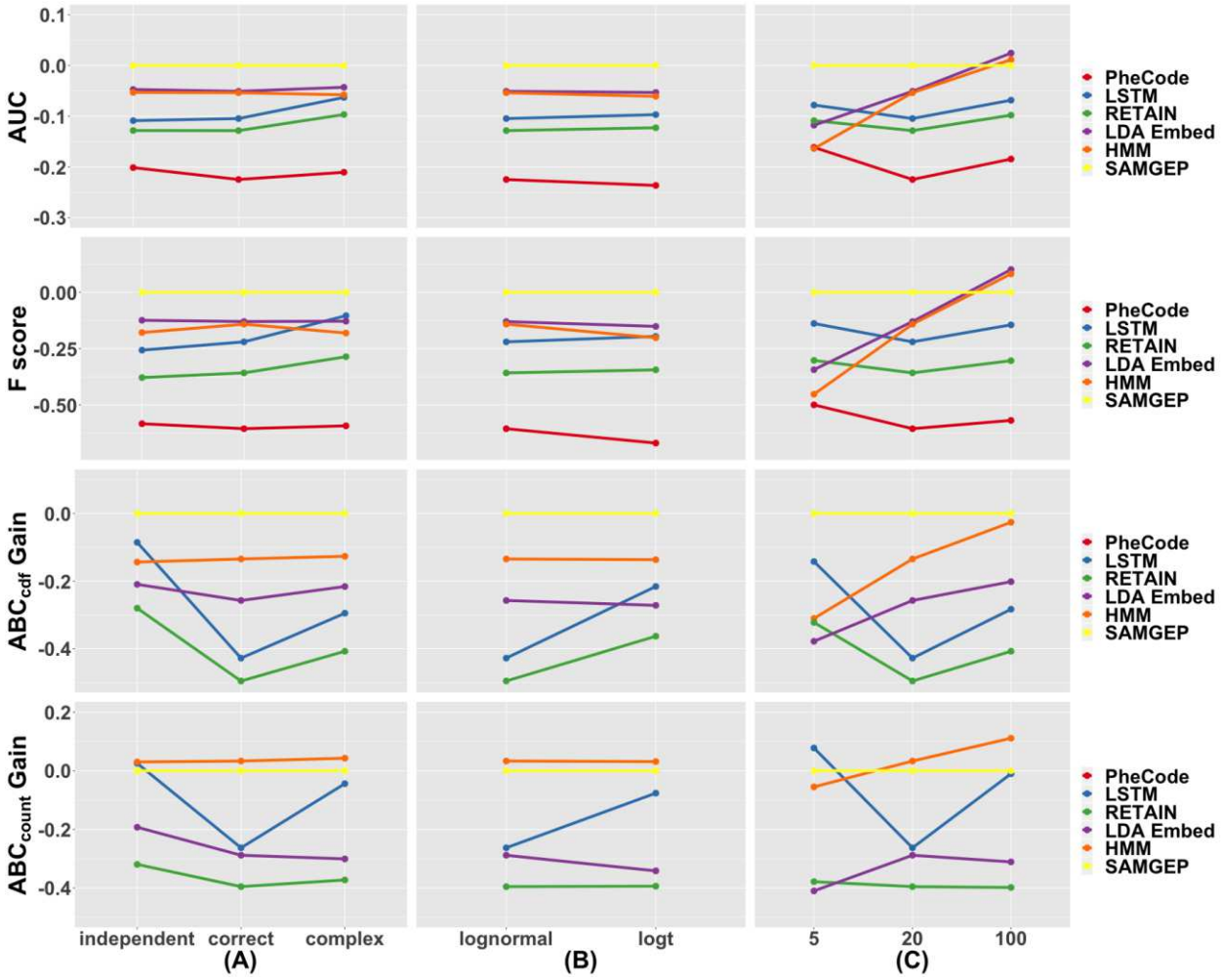


Figure 3: Robustness of SAMGEP and comparator methods' AUCs, F scores, ABC_{cdf} gains, and ABC_{count} gains to various generative parameters, including the (a) specification of $Y|T$, (b) specification of $\boldsymbol{C}|Y$, and (c) number of informative (i.e. non-sparse) features. Details of the experiments are delineated in the *Simulation Study* subsection of the Methods, and more extensive results are displayed in Supplementary Figure S5.

Robustness to Data Generative Characteristics

Figure 3 explores SAMGEP and comparators' robustness to various generative model specifications. Note that only relative performance between methods, not absolute performance, is meaningful as different generative settings may portend disparate inherent levels of information.

Panels A and B demonstrate unsurprisingly that SAMGEP outperforms RETAIN and LSTM when SAMGEP's distributional assumptions regarding (A) $\boldsymbol{Y}|\boldsymbol{T}$ and (B) $\boldsymbol{X}|\boldsymbol{Y}$ are correctly specified. That said, SAMGEP achieves strong relative performance notwithstanding misspecification of $\boldsymbol{Y}|\boldsymbol{T}$ or $\boldsymbol{X}|\boldsymbol{Y}$. Practically this indicates that SAMGEP is robust to model misspecification, though the more the true distribution diverges from SAMGEP's assumption, the less desirable SAMGEP is relative to highly flexible deep learning models.

Panel C demonstrates that SAMGEP provides more benefit over LDA_{Embed} and HMM the more sparsely information is distributed over features (i.e. 5 rather than 100 informative features out of 150), reflecting the robustness of SAMGEP's L_1-regularized weighting protocol to information sparsity. This robustness makes SAMGEP well-conditioned for the EHR, which typically contains a handful of informative features out of millions. Meanwhile, SAMGEP offers the most benefit over LSTM and RETAIN, which also utilize L_1 weighting, when the number of informative features is 20. This likely reflects the fact that we set the dimension of $\boldsymbol{X}$ to 10, resulting in bias when the true number of informative dimensions exceeds 10. Thus, SAMGEP is optimized relative to alternatives when ~10 degrees of freedom are needed to fully capture $\boldsymbol{C}|\boldsymbol{Y}$.

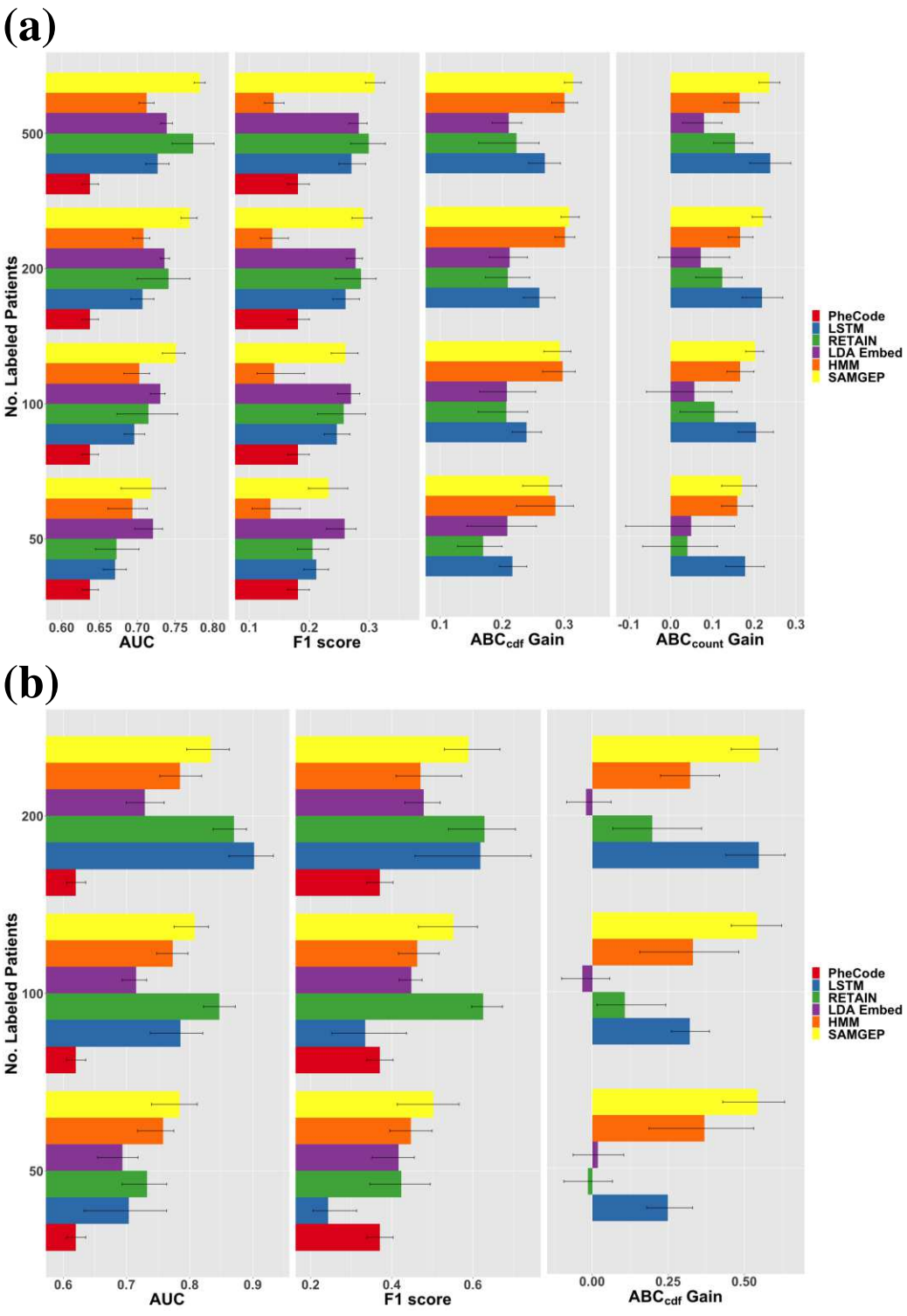


Figure 4: Predictive accuracies of SAMGEP and various comparator methods using real-world EHR data to predict (a) MS relapse with $n \in \{50,100,200,500\}$ labeled patients, and (b) HF onset with $n \in \{50,100,200\}$ labels. 95% confidence intervals were empirically estimated by bootstrapping with 100 replicates. See the *Evaluation Metrics* subsection of the Methods for details about the evaluation metrics. More extensive results are displayed in Supplementary Figure S6.

Identification of MS Relapse and HF Onset Using Real-World EHR Data

Figure 4 depicts mean AUCs, F1 scores, ABC_{cdf} gains, and ABC_{count} gains for SAMGEP and comparator methods predicting (A) MS relapse and (B) HF onset using real-world EHR data. For identification of MS relapse, SAMGEP achieved significantly higher AUCs than all other methods, though RETAIN approached SAMGEP for $n = 500$ labels. SAMGEP also achieved F1 scores equivalent to that of the top-performing method for all n. LSTM and RETAIN achieved lackluster AUCs for $n \in \{50,100,200\}$, suggesting that even 200 labels is insufficient for such complex models. SAMGEP also achieved the highest ABC_{cdf} gains, though not significantly so relative to HMM. Finally, SAMGEP achieved the highest ABC_{count} gains, though statistically equivalent to LSTM per Student's t test and only marginally superior to HMM. The fact that SAMGEP, HMM, and LSTM were the top performers by both ABC metrics, despite HMM and LSTM's unremarkable AUCs, suggests that jointly modeling $\{Y_{i,1}, \dots, Y_{i,T(i)}\}$ is singularly beneficial for longitudinal phenotype process prediction. LDA_{embed} does not even significantly improve upon the null model's ABC metrics for $n \in \{50,100,200\}$, demonstrating that accurately predicting phenotype states at individual timepoints does not necessarily translate into accurate phenotype process prediction.

For identification of HF onset, SAMGEP achieved the highest AUCs and F1 scores for $n = 50$ but was outperformed by RETAIN for $n = 100$, and both RETAIN and LSTM for $n = 200$. SAMGEP achieved the highest ABC_{cdf} gains across the board but was statistically equivalent to HMM for $n = 50$ and LSTM for $n = 200$. The fact that SAMGEP, LSTM, RETAIN, and HMM achieved high accuracies across metrics reflects the powerful benefit of leveraging the full time sequence when predicting the onset of a chronic disease, wherein $Y_t \mid (Y_{t-1} = 1) = 1$ with probability 1. Moreover, the fact that RETAIN often outperformed SAMGEP for HF onset but not MS relapse identification suggests that SAMGEP offers particular benefit over deep learning for prediction of a relapsing and remitting process, which expends more degrees of freedom than prediction of disease onset.

While SAMGEP does not always outperform all comparators, its consistency is notable. Whereas some comparators achieve high accuracy by only certain metrics or on only one of our two outcomes, SAMGEP achieves consistently strong performance, and for $n = 50$ it always achieves statistically equivalent accuracy to the top performing method. It demonstrates proficiency at predicting both phenotype states at individual timepoints and phenotype processes over time, most notably in the label-poor setting (i.e. 50-100 labels). Finally, it achieves high accuracy on two contrasting phenotypic outcomes, bolstering our claim of generalizability.

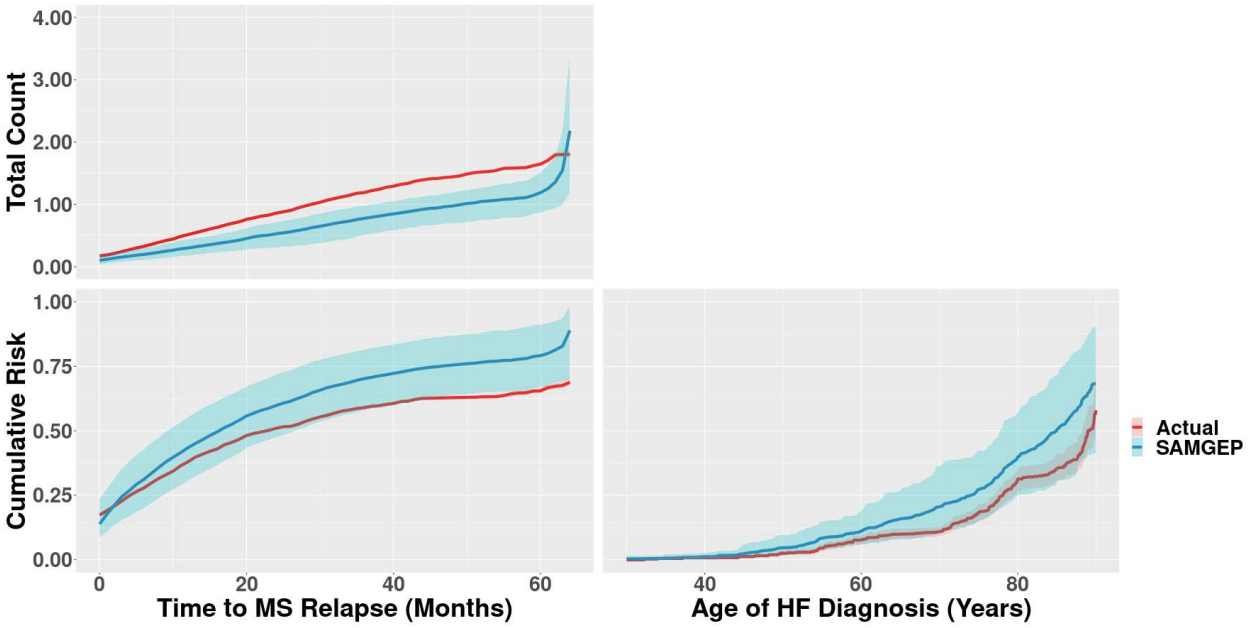


Figure 5: Estimation of population-wide cumulative probability (bottom) and counting process (top) curves for MS relapse (left) and HF development (right) using the identifications of SAMGEP with $n = 100$ labeled patients. 95% confidence intervals were empirically estimated by bootstrapping with 100 replicates. More extensive results are displayed in Supplementary Figure S7.

Estimation of Cumulative Probability and Counting Process Curves

Figure 5 depicts the estimated CDF, obtained as $\hat{F}(t) = N^{-1}\sum_{i=1}^{N}\hat{F}_{i,t}$, and counting process, obtained as $\hat{N}(t) = N^{-1}\sum_{i=1}^{N}\hat{N}_{i,t}$, based on the identifications of SAMGEP using $n = 100$ labels, along with 95% confidence intervals. Notably, CDF estimation using SAMGEP's identifications is relatively

unbiased for onset of both (A) first MS relapse and (B) HF. As Supplementary Figure S7 demonstrates, comparator methods' predictions tends to markedly overestimate the true cumulative risk, whereas SAMGEP consistently achieves the least biased estimates.

Counting process estimation using SAMGEP's MS relapse identifications appears to systematically but slightly underestimate the true counting process. SAMGEP significantly improves upon comparators later in patients' disease courses, where all other methods except HMM appear to markedly overestimate. SAMGEP's identifications again improve bias at the expense of increased variance, overall significantly improving ABC_{count}.

DISCUSSION

While identification of binary phenotypes using EHR data is well-trodden in the literature, identification of longitudinal phenotype processes, or event times, remains underdeveloped. As our results demonstrate, accurate identification of a patient's phenotype status overall – or even at a particular timepoint – does not necessarily translate into accurate phenotype process prediction. SAMGEP accurately predicts phenotype processes, constituting a meaningful step forward for computational phenotyping.

SAMGEP excels relative to existing methods because it 1) can leverage numerous EHR features, which is particularly important for phenotypes that are insufficiently represented by a handful of surrogates (i.e. ICD codes); 2) incorporates prior knowledge by utilizing low-dimensional feature embeddings trained using all available EHR data; 3) can efficiently utilize few gold-standard labels by leveraging unlabeled data in a semi-supervised manner; and 4) jointly models the time sequence of relapses and features rather than treating individual timepoints as independent observations. Figures 4A, 5, and S7 demonstrate that SAMGEP achieves particularly accurate (per ABC_{cdf} and ABC_{count}) and unbiased CDF and counting process estimates for a relapsing-and-remitting phenotype, particularly in the

setting of few (i.e. 50-100) observed labels. Manual annotation of event times is an extremely laborious process, so SAMGEP's label efficiency is a key attribute.

We envision SAMGEP's phenotype process predictions being used as outcomes in a downstream clinical or epidemiological study. For instance, researchers aiming to measure the effect of an MS treatment on the rate of relapse could 1) annotate the relapse histories of ~100 patients via chart review, 2) use SAMGEP to estimate cumulative relapse probabilities for all remaining patients, and 3) use these relapse probabilities as outcomes to measure treatment effect. Further research is warranted to assess the bias-variance tradeoff of such a workflow relative to traditional predictive modeling methods using the labeled set alone.

While SAMGEP is label-efficient, manually annotating even 50-100 event time labels is a labor-intensive process. Modifying SAMGEP to handle current status labels – indicators of phenotype status at censor time – would greatly diminish the chart review time required to utilize the algorithm for phenotype onset prediction. Further work is warranted to explore this possibility.

In summary, SAMGEP is a novel semi-supervised machine learning method that accurately predicts the course of a binary phenotype over time using EHR data with few observed event time labels. Singularly adept at estimating cumulative probability and counting process functions, SAMGEP promises to enable more powerful use of EHR data for epidemological research involving event timings, including survival analysis.

METHODS

Assembling Predictive Features

To assemble feature counts and define phenotype states, we consider consecutive non-overlapping time periods starting at the patient's first ICD code for the target phenotype. Candidate features include log-transformed counts of ICD codes, RxNorm drug codes, CPT codes, lab tests, and mentions of clinical concepts in a patient's record. Features can be selected manually or identified

automatically via label-free methods such as surrogate-assisted feature extraction.[33] Figure 1B depicts the form of raw EHR feature data for MS relapse identification, with red bands indicating relapse events. Since SAMGEP employs sparse feature weighting to select informative features, it is preferable to include features liberally rather than aiming for parsinominousness in feature assembly.

Henceforth we let $\boldsymbol{V}_{m\times p}$ denote the matrix of m-dimensional embedding vectors for p features. See the *Producing Feature* Embeddings section of the Supplementary Materials for details on how $\boldsymbol{V}_{m\times p}$ is pre-trained. We use i, j, and t to index patients, raw features, and time periods respectively. We assume there are N patients and T_i periods for patient i in our dataset. For patient i at timepoint t, let $\boldsymbol{C_{i,t}}$ denote the p-dimensional raw feature vector and $Y_{i,t} \in \{0,1\}$ denote the phenotype state. Let H_i = log(mean healthcare encounter count per month + 1) in patient i's record, a measure of healthcare utilization. Finally, we assume that $\boldsymbol{Y_i} = (Y_{i,1}, \dots, Y_{i,T_i})$ is annotated on a limited set of $n \ll N$ patients, but $\boldsymbol{C_i} = \left(\boldsymbol{C_{i,1}}, \dots, \boldsymbol{C_{i,T_i}}\right)$ is observed for all N patients.

Producing Patient-Timepoint Embeddings

SAMGEP leverages pre-trained feature embeddings $\boldsymbol{V}$ to compress the high-dimensional feature vector $\boldsymbol{C_{i,t}}$ to a low-dimensional patient-timepoint embedding $\boldsymbol{X_{i,t}}$, which can be efficiently modeled as a Gaussian process. We compute $\boldsymbol{X_{i,t}}$ as a weighted sum over feature embeddings:

$$\boldsymbol{X_{i,t}} = \boldsymbol{C_{i,t}}\boldsymbol{W}\boldsymbol{V}, \text{ for } i = 1, \dots, N \text{ and } t = 1, \dots, T_i, \tag{1}$$

where $\boldsymbol{W} = \begin{pmatrix} W_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & W_p \end{pmatrix}$, W_j is the unknown weight for the j^{th} feature, and $W_1 = 1$ to ensure identifiability. We choose $\boldsymbol{W}$ via L1-regularized linear discriminant analysis minimizing

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

where $\|\boldsymbol{W}\|_{\boldsymbol{1}}^{\boldsymbol{1}}$ denotes the L_1 norm of $\boldsymbol{W}$, $\lambda \geq 0$ is the tuning parameter,

$$\boldsymbol{\mu_y} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T_i}\boldsymbol{X_{i,t}}I(Y_{i,t} = y)}{\sum_{i=1}^{N}\sum_{t=1}^{T_i}I(Y_{i,t} = y)}, \text{and } \boldsymbol{\Sigma_X} = \frac{1}{\sum_{i=1}^{N}T_i}\sum_{i=1}^{N}\sum_{t=1}^{T_i}(\boldsymbol{X_{i,t}} - \boldsymbol{\mu_{Y_{i,t}}})\left(\boldsymbol{X_{i,t}} - \boldsymbol{\mu_{Y_{i,t}}}\right)'; \; y = 0,1.$$

We choose L_1 regularization over L_2 or other L_p to impose sparsity given that most input features are likely uninformative. We optimize $\boldsymbol{W}$ using projected gradient ascent, where without loss of generality we assume that the first feature is a known highly predictive feature. The step-size at each iteration of ascent is chosen by line search, and λ is optimized using 5-fold cross-validation within the labeled set.

Fitting MGP

MGP is a generative mixture-like model that combines two assumptions: 1) $\boldsymbol{Y_i}$ follows a discrete time Markov process, and 2) $\boldsymbol{X_i} = \left(\boldsymbol{X_{i,1}}, \dots, \boldsymbol{X_{i,T_i}}\right)|\boldsymbol{Y_i}$ follows a Gaussian process. This generative framework primes SAMGEP for the semi-supervised setting.

Discrete Time Markov Process Assumption

We assume a Markov process model for $\boldsymbol{Y_i}|H_i$ such that $P(Y_{i,t} = y|Y_{i,1}, \dots, Y_{i,t-1}, H_i) = P(Y_{i,t} = y|Y_{i,t-1}, H_i)$. This model is specified by two rules:

$$P(Y_{i,1} = 1|H_i) = \pi_{init}(H_i); \quad P(Y_{i,t} = y|Y_{i,t-1} = y_{t-1}, H_i) \equiv \pi_t(y_{t-1}, H_i) \text{ for } t > 1, \qquad (2)$$

where $\{\pi_{init}, \pi_t(y_{t-1}) \; \forall \; t > 1\}$ are unknown transition probabilities that fully specify the Markov model. We further assume that for some $\boldsymbol{\lambda}_{\text{markov}} = \{\lambda_{init}, \lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_{H0}, \lambda_H\}$,

$$\pi_{init}(H_i) = expit(\lambda_{init} + \lambda_{H0} H_i), \text{ and}$$

$$\pi_t(y_{t-1}|H_i) = expit(\lambda_0(1 - y_{t-1}) + \lambda_1 y_{t-1} + \lambda_2 t + \lambda_3 \log t + \lambda_H H_i)$$

where $expit(x) = \exp(x)/[1 + \exp(x)]$. We include both linear and log-linear time effects on $\pi_t(y_{t-1}|H_i)$ to better capture the temporal risk function without overfitting.

Gaussian Process Assumption

We assume the dense representations of patients' EHRs (i.e. patient embeddings) over time follow a Gaussian process:

$$\boldsymbol{X_i}|\boldsymbol{Y_i} \sim GP(\boldsymbol{\mu_i}(\boldsymbol{t}), \boldsymbol{\Sigma_i}(\boldsymbol{t})).$$

Since the feature embeddings $\boldsymbol{V}$ are engineered to approximately follow a multivariate normal distribution as described in the Supplementary Materials, it is reasonable to assume $\boldsymbol{X_{i,t}}$ to be a Gaussian process over time t. We further specify the mean and covariance functions $\boldsymbol{\mu_i(t)}$ and $\boldsymbol{\Sigma_i(t)}$ respectively. For some $\boldsymbol{\theta}_{\mathrm{GP}} = \{\boldsymbol{\mu_0}, \boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \boldsymbol{\mu_3}, \boldsymbol{\mu_4}, \boldsymbol{\mu_5}, \boldsymbol{\mu_H}, \boldsymbol{\mu_{YH}}, \sigma_k, \alpha_k, \tau_k, \rho_{kl}, k = 1, \dots, p; l = 1, \dots, p\}$, we assume:

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

$$\mathrm{Var}(X_{i,t,k}) = \sigma_k^2 \exp(2\alpha_k H_i), \ \mathrm{Cov}(X_{i,t,k}, X_{i,t,l}) = \rho_{kl}\sigma_k\sigma_l \exp\{(\alpha_k + \alpha_l) H_i\}.$$

In summary, we assume that patient i's expected embedding at time t, $\boldsymbol{\mu_i(t)}$, is a function of Y_{it}, H_i, and t. We assume that the marginal variance of embedding component k can be represented by some baseline σ_k^2 scaled by H_i. We denote the correlation between embedding components k and l as ρ_{kl}, which we assume to be constant over time. Between timepoints, we employ a first-order univariate autoregressive (AR(1)) kernel structure such that the residual at t, $\epsilon_{i,t,k} = X_{i,t,k} - E(X_{i,t,k}|\boldsymbol{Y_i}, H_i)$, is a linear function of its preceding value $\epsilon_{i,t-1,k}$ with autocorrelation coefficient τ_k:

$$E\left[\epsilon_{i,t,k} \middle| \epsilon_{i,t-1,k}\right] = r\tau_k \epsilon_{i,t-1,k}.$$

$r \in [0,1]$ is an autoregression regularization hyperparameter separately trained via 5-fold cross-validation: $r = 0$ ignores intertemporal correlation while $r = 1$ denotes undampened autoregression. We chose first-degree autoregression over higher-degree models due to computational ease and mitigation of overfitting. We provide a sensitivity analysis with respect to the choice of *k*-fold cross-validation in Supplementary Figure S2 that demonstrates no significant effect of *k* on predictive accuracy.

Implementation and Inference

MGP is fit via one iteration of an approximating expectation-maximization (EM) algorithm. We approximate the expected log-likelihood in the E-step using the marginal posterior of each latent phenotype state, $\widehat{Y}_{i,t}|\boldsymbol{X}_i \ \forall \ t \in \{1, \dots, T_i\}$, rather than the more complex joint posterior, $\widehat{\boldsymbol{Y}}_i|\boldsymbol{X}_i$. Before we fit MGP, we optimize r using 5-fold cross-validation on the labeled set. We then initialize the EM by optimizing the model parameters $\{\boldsymbol{\lambda}_{\mathrm{markov}}, \boldsymbol{\theta}_{\mathrm{GP}}\}$ on the labeled set and using this model to impute labels

for the unlabeled set. We refer to predictions using this initial model as MGP's supervised estimator $\widehat{\boldsymbol{p}}_{\boldsymbol{sup}}$. Finally, we re-optimize $\{\boldsymbol{\lambda}_{\text{markov}}, \boldsymbol{\theta}_{\text{GP}}\}$ using both observed and imputed labels. We refer to predictions using this re-trained model as MGP's semi-supervised estimator $\widehat{\boldsymbol{p}}_{\boldsymbol{semisup}}$. We execute one iteration of this EM procedure rather than running it to convergence for several reasons. First, since the EM is initialized at the consistent supervised estimator, the solution obtained after one iteration is guaranteed to be consistent assuming correct model specification while heuristically minimizing the influence of the unlabeled set. The one-step update also greatly reduces the computational cost. As Supplementary Figure S3 demonstrates, SAMGEP's performance is not sensitive to the maximum number of EM iterations allowed. Specific details of our fitting procedure are supplied in the Supplementary Materials.

Combining Semi-supervised and Supervised Predictions

Semi-supervised generative models such as MGP should benefit from the additional information in the unlabeled set if the model is correctly specified. However, semi-supervised predictors have been shown to be more sensitive to model misspecification than their supervised counterparts. To mitigate this effect, SAMGEP returns a weighted average of $\widehat{\boldsymbol{p}}_{\boldsymbol{sup}}$ and $\widehat{\boldsymbol{p}}_{\boldsymbol{semisup}}$, $\widehat{\boldsymbol{p}}_{\boldsymbol{final}} = \alpha\widehat{\boldsymbol{p}}_{\boldsymbol{semisup}} + (1-\alpha)\widehat{\boldsymbol{p}}_{\boldsymbol{sup}}$ **(3),** with weight α selected by 5-fold cross-validation maximizing the AUROC of $Y_{i,t}$ predictions. As we demonstrate in Supplementary Figure S1, $\widehat{\boldsymbol{p}}_{\boldsymbol{final}}$ is consistently equivalent or superior to the better of $\widehat{\boldsymbol{p}}_{\boldsymbol{sup}}$ and $\widehat{\boldsymbol{p}}_{\boldsymbol{semisup}}$.

Data and Metrics for Evaluation

Simulation Study

We generated datasets of $p = 150$ count features along with H for $N \in \{1000,5000,20000\}$ unlabeled patients, each with a mean of $E[T_i] = 25$ timepoints. To assess SAMGEP's robustness to various model misspecifications, we varied the following generative parameters: (i) $Y|T$ where 'independent' indicates $Y \perp T$, 'correct' follows SAMGEP's generative model, and 'complex' denotes

over-parametrization of $Y(T)$; and (ii) $\boldsymbol{C}|Y$ (marginally lognormal vs. log-t with 5 degrees of freedom). We considered $n = 100$ labeled patients and let the number of informative features vary from 5 to 100. Details of our simulation generative mechanisms are supplied in the *Simulation Data Generative Mechanisms* section of the Supplementary Materials.

Multiple Sclerosis (MS) Relapse and Heart Failure (HF) Onset Identification

We further validate SAMGEP's performance by classifying MS relapse and HF status over time using two EHR studies. Whereas HF is a chronic disease for which primary interest lies in the cumulative probability of disease over time, MS is a relapsing and remitting phenotype for which we seek to predict all relapses over time. For identification of MS relapse, we collected EHR data between January 1, 2006 and December 31, 2016 for 4,706 patients with at least one MS ICD code from the Research Patient Data Registry (RPDR) of the Mass General Brigham (MGB) health system in Boston, MA. We derived neurologist-confirmed MS relapse events and dates for 1,435 patients from the Comprehensive Longitudinal Investigation of Multiple Sclerosis (CLIMB) research registry. 57.2% of these patients had at least one relapse event, with a mean of 2.60 relapses per patient and 0.081 relapses per patient-month. For HF, we collected EHR data for 59,395 patients in the MGB RPDR with at least one ICD code for HF. We compiled HF status and onset dates for 300 randomly selected patients from this cohort via chart review by two independent cardiologists at MGB who jointly reconciled any differences in initial assessment. Among the 300, 60.7% developed HF during follow-up. The MGB IRB approved the use of both EHR and research registry data, and data were appropriately deidentified before use in accordance with relevant guidelines and regulations. Informed consent was obtained from all subjects and/or their legal guardians by RPDR investigators during collection of the data.

From the EHR dataset we extracted age, sex, and patient-level occurrences of ICD and CPT codes. We grouped ICD codes to PheCodes using the established PheWAS mapping.[34] From free-text clinical narratives we extracted mentions of clinical concepts via the Narrative Information Linear Extraction (NILE) natural language processing (NLP) method.[35] We binned all EHR features into

consecutive, non-overlapping 1-month time intervals such that $\boldsymbol{C_{i,t}}$ represents the counts of patient i's PheCodes, CPT codes, and NLP features between months t and $t+1$. As Supplementary Figure S4 demonstrates, SAMGEP attains higher accuracy with longer window lengths, suggesting that the user should generally employ the longest window length that achieves sufficient temporal precision for a given task. 155 features were selected for identification of MS relapse by a domain expert. For HF onset identification, we selected 121 features with embedding cosine similarities of 0.1 or above relative to the HF ICD code.

Benchmark Methods for Comparison

We considered as benchmarks three supervised methods using the labeled set alone: (i) long short term memory RNN (LSTM) [24,39,43,44] trained with $\boldsymbol{C_{i,t}}$, (ii) RETAIN [26] trained with raw EHR observations in the continuous time domain, and (iii) linear discriminant analysis (LDA) trained with patient-timepoint embeddings generated without weights ($\boldsymbol{X^0_{i,t} = C_{i,t}V}$), which we refer to as $\text{LDA}_{\text{embed}}$. $\text{LDA}_{\text{embed}}$ predicts $Y_{i,t}$ using only concurrent features without considering the time sequence. In addition, we considered a semi-supervised benchmark: HMM [26–29,45,46] with a multivariate gaussian emission trained on $\boldsymbol{X^0_{i,t}}$. As a baseline we also included predictions based only on the closest PheCodes (MS: 355; HF: 428). See the Supplementary Materials for details of our benchmark method implementations. In Supplementary Figure S5 we also include results for LASSO-penalized logistic regression [16,17,34,37–39], random forest (RF) [40,41], and LDA [45] trained with $\boldsymbol{C_{i,t}}$. These methods leverage neither the time sequence nor $\boldsymbol{V}$and achieve subpar predictive accuracy.

Evaluation Metrics

To evaluate methods' predictions for $Y_{i,t}$, we computed (i) AUC and (ii) F1 score choosing a cutoff that achieves 95% specificity. Let $N_i(t) = \sum_{k \le t} Y_{i,k}(1 - Y_{i,k-1})$ denote the observed all-event counting process, where $Y_{i,0} = 0$, and let $F_i(t) = 1 - \Pi_{k \le t}(1 - Y_{i,k})$ denote the observed first-event

process. We evaluated methods' phenotype counting process predictions by computing the area between $N_i(t)$ and the predicted counting process $\hat{N}_i(t) = \sum_{k \le t} I(\hat{\pi}_{i,k} \geq c)$, denoted as $\text{ABC}_{\text{count}}$, where $\hat{\pi}_{i,k}$ denotes a method's prediction of $P(Y_{i,k} = 1)$ and c is chosen such that in labeled set

$$\frac{\sum_{i=1}^{n} \sum_{k=1}^{T(i)} I(\hat{\pi}_{i,k} \geq c)}{\sum_{i=1}^{n} T(i)} = \frac{\sum_{i=1}^{n} \sum_{k=1}^{T(i)} Y_{i,k}}{\sum_{i=1}^{n} T(i)}.$$

Likewise, we evaluated methods' first-event cumulative probability (CDF) predictions by computing the area between $F_i(t)$ and the predicted CDF $\hat{F}_{i,t} = 1 - \prod_{k \le t}(1 - \hat{\lambda}_{i,k})$, denoted by ABC_{cdf}, where $\lambda_{i,k}$ denotes patient i's true hazard at time k and $\hat{\lambda}_{i,k}$ denotes a method's prediction thereof. In the absence of censoring, ABC_{cdf} is equivalent to the mean absolute difference between true and predicted event times. Since SAMGEP and HMM jointly model the outcome sequence $\{Y_{i,1}, \ldots, Y_{i,T(i)}\}$, we used these methods to directly estimate $\hat{F}_{i,t}$. Other methods only predict marginal probabilities $\hat{\pi}_{i,t}$, so we assumed that $\hat{\lambda}_{i,t} = \hat{\pi}_{i,t}$, or equivalently that event states are independent over time. Rather than report raw ABC quantities – whose scale can vary greatly across settings – we report methods' percent decrease below those of the null model that sets $\hat{\pi}_{ik}$ to the prevalence at time k:

$$\text{ABC}_{\text{cdf}}^{\text{Gain}} = \frac{(\text{ABC}_{\text{cdf,null}} - \text{ABC}_{\text{cdf,method}})}{\text{ABC}_{\text{cdf,null}}}, \text{and } \text{ABC}_{\text{count}}^{\text{Gain}} = \frac{(\text{ABC}_{\text{count,null}} - \text{ABC}_{\text{count,method}})}{\text{ABC}_{\text{count,null}}}. \quad (4)$$

We do not compute $\text{ABC}_{\text{count}}^{\text{Gain}}$ for HF onset identification since only the CDF is of interest for HF. Finally, for both MS relapse and HF onset identification, we qualitatively evaluate SAMGEP's feature selection and weighting mechanism by generating feature clouds using the product of SAMGEP's feature weights and emprirical feature standard devations: $w_j = W_{j,j} \times \hat{s}(\boldsymbol{C}_{,j})$.

REFERENCES

1. Kohane, I. S., Churchill, S. E. & Murphy, S. N. A translational engine at the national scale: informatics for integrating biology and the bedside. *J. Am. Med. Informatics Assoc.* **19**, 181–185 (2012).

2. Hripcsak, G. & Albers, D. J. Next-generation phenotyping of electronic health records. *J. Am. Med. Informatics Assoc.* **20**, 117–121 (2012).

3. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* **6**, 26094 EP (2016).

4. Liao, K. P. *et al.* Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res* **62**, 1120–1127 (2010).

5. Cipparone, C. W. *et al.* Inaccuracy of ICD-9 codes for chronic kidney disease: A study from two practice-based research networks (PBRNs). *J. Am. Board Fam. Med.* **28**, 678–682 (2015).

6. Uno, H. *et al.* Determining the Time of Cancer Recurrence Using Claims or Electronic Medical Record Data. *JCO Clin. Cancer Informatics* 1–10 (2018) doi:10.1200/cci.17.00163.

7. Hassett, M. J. *et al.* Detecting lung and colorectal cancer recurrence using structured clinical/administrative data to enable outcomes research and population health management. *Med. Care* **55**, e88–e98 (2017).

8. Chubak, J. *et al.* Administrative data algorithms to identify second breast cancer events following early-stage invasive breast cancer. *J. Natl. Cancer Inst.* **104**, 931–940 (2012).

9. Carroll, R. J. *et al.* Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J. Am. Med. Informatics Assoc.* **19**, e162--e169 (2012).

10. Liao, K. P. *et al.* Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. *PLoS One* **10**, e0136651 (2015).

11. Beaulieu-Jones, B. K., Greene, C. S. & others. Semi-supervised learning of the electronic health record for phenotype stratification. *J. Biomed. Inform.* **64**, 168–178 (2016).

12. Newton, K. M. *et al.* Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J. Am. Med. Informatics Assoc.* **20**, e147--e154 (2013).

13. Ananthakrishnan, A. N. *et al.* Improving case definition of Crohn's disease and ulcerative colitis

in electronic medical records using natural language processing: a novel informatics approach. *Inflamm. Bowel Dis.* **19**, 1411–1420 (2013).

14. Xia, Z. *et al.* Modeling disease severity in multiple sclerosis using electronic health records. *PLoS One* **8**, e78927 (2013).

15. Liao, K. P. *et al.* Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *bmj* **350**, h1885 (2015).

16. Kirby, J. C. *et al.* PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J. Am. Med. Informatics Assoc.* **23**, 1046–1052 (2016).

17. Halpern, Y., Choi, Y., Horng, S. & Sontag, D. Using anchors to estimate clinical state without labeled data. in *AMIA Annual Symposium Proceedings* vol. 2014 606 (2014).

18. Yu, S. *et al.* Enabling phenotypic big data with PheNorm. *J. Am. Med. Informatics Assoc.* **25**, 54–60 (2017).

19. Liao, K. *et al.* High-throughput Multimodal Automated Phenotyping (MAP) with Application to PheWAS. *J. Am. Med. Informatics Assoc.* **26**, 1255–1262 (2019).

20. Ahuja, Y. *et al.* sureLDA: A multidisease automated phenotyping method for the electronic health record. *J. Am. Med. Informatics Assoc.* **27**, 1235–1243 (2020).

21. Choi, E., Du, N., Chen, R., Song, L. & Sun, J. Constructing disease network and temporal progression model via context-sensitive hawkes process. *Proc. - IEEE Int. Conf. Data Mining, ICDM* **2016-Janua**, 721–726 (2016).

22. Kaji, D. A. *et al.* An attention based deep learning model of clinical events in the intensive care unit. *PLoS One* **14**, 1–17 (2019).

23. Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* **1**, 1–10 (2018).

24. Ruan, T. *et al.* Representation learning for clinical time series prediction tasks in electronic health records. *BMC Med. Inform. Decis. Mak.* **19**, 1–14 (2019).

25. Cheng, Y., Wang, F., Zhang, P. & Hu, J. Risk prediction with electronic health records: A deep

learning approach. *16th SIAM Int. Conf. Data Min. 2016, SDM 2016* 432–440 (2016) doi:10.1137/1.9781611974348.49.

26. Choi, E. *et al.* RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. *Adv. Neural Inf. Process. Syst.* **29**, 3512–3520 (2016).

27. Pivovarov, R. *et al.* Learning probabilistic phenotypes from heterogeneous EHR data. *J. Biomed. Inform.* **58**, 156–165 (2015).

28. Pivovarov, R. Electronic health record summarization over heterogeneous and irregularly sampled clinical data. (Columbia University, 2016).

29. Jackson, C. H., Sharples, L. D., Thompson, S. G., Duffy, S. W. & Couto, E. Multistate Markov models for disease progression with classification error. *Stat.* **52**, 193–209 (2003).

30. Sukkar, R., Katz, E., Zhang, Y., Raunig, D. & Wyman, B. T. Disease progression modeling using Hidden Markov Models. *Conf Proc IEEE Eng Med Biol Soc* 2845–2848 (2012).

31. Wang, X., Sontag, D. & Wang, F. Unsupervised learning of disease progression models. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 85–94 (2014) doi:10.1145/2623330.2623754.

32. Zhou, X., Kang, K. & Song, X. Two-part hidden Markov models for semicontinuous longitudinal data with nonignorable missing covariates. *Stat. Med.* **39**, 1801–1816 (2020).

33. Yu, S. *et al.* Surrogate-assisted feature extraction for high-throughput phenotyping. *J. Am. Med. Informatics Assoc.* **24**, e143–e149 (2017).

34. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1111 (2013).

35. Yu, S., Cai, T. & Cai, T. NILE: Fast Natural Language Processing for Electronic Health Records. *arXiv* 1–23 (2013).

36. Lin, C. *et al.* Automatic Prediction of Rheumatoid Arthritis Disease Activity from the Electronic Medical Records. *PLoS One* **8**, (2013).

37. Li, R. *et al.* Detection of bleeding events in electronic health record notes using convolutional

neural network models enhanced with recurrent neural network autoencoders: Deep learning approach. *J. Med. Internet Res.* **21**, 1–10 (2019).

38. Yang, Z., Dehmer, M., Yli-Harja, O. & Emmert-Streib, F. Combining deep learning with token selection for patient phenotyping from electronic health records. *Sci. Rep.* **10**, 1–18 (2020).
39. Sun, Z. *et al.* A probabilistic disease progression modeling approach and its application to integrated Huntington's disease observational data. *JAMIA Open* **2**, 123–130 (2019).
40. Verma, A., Powell, G., Luo, Y., Stephens, D. & Buckeridge, D. L. Modeling disease progression in longitudinal EHR data using continuous-time hidden Markov models. 1–5 (2018).
41. Castro, V. M. *et al.* Validation of Electronic Health Record Phenotyping of Bipolar Disorder and Controls. *Am. J. Psychiatry* **172**, 363–372 (2015).
42. Anderson, A. E. *et al.* Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study. *J. Biomed. Inform.* **60**, 160–168 (2016).
43. Garg, R., Dong, S., Shah, S. & Jonnalagadda, S. R. A Bootstrap Machine Learning Approach to Identify Rare Disease Patients from Electronic Health Records Division of Health and Biomedical Informatics , Department of Preventive Medicine , Division of Cardiology , Department of Medicine , Northwestern Unive. (2016).
44. Teixeira, P. L. *et al.* Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *J. Am. Med. Informatics Assoc.* **24**, 162–171 (2017).
45. Yang, S. *et al.* Early detection of disease using electronic health records and fisher's wishart discriminant analysis. *Procedia Comput. Sci.* **140**, 393–402 (2018).

Author Contributions: Y.A. conceptualized and developed SAMGEP, conducted experimentation and statistical analysis, and wrote the manuscript. J.W. conducted experimentation and statistical analysis and wrote the manuscript. C.H., Z.X., and S.H. compiled and pre-processed the datasets used to analyze

SAMGEP. T.C. provided supervision and critical guidance. All authors contributed meaningfully to revision of the manuscript.

Data Availability: The Electronic Health Record (EHR) and research registry data underlying this article were provided by Mass General Brigham (MGB) Research Information Sciences & Computing by permission, and with the approval of the MGB Institutional Review Board (IRB). These data will be shared on request to the corresponding author with permission of MGB Research Information Sciences & Computing as well as the MGB IRB.

Funding: This work was supported by the U.S. National Institutes of Health Grants T32-AR05588512, T32-GM7489714, and R21-CA242940.

Competing Interests: None.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SAMGEPsupplementfinalscientificreports.pdf