Flexible-Window Predictions on Electronic Health Records

Mehak Gupta[1], **Raphael Poulain**[1], **Thao-Ly T. Phan**[2], **H. Timothy Bunnell**[2], **Rahmatollah Beheshti**[1]

[1]University of Delaware, DE, USA
[2]Nemours Children's Health, DE, USA
{mehakg, rpoulain}@udel.edu, {thaoly.phan, tim.bunnell}@nemours.org, rbi@udel.edu

Abstract

Various types of machine learning techniques are available for analyzing electronic health records (EHRs). For predictive tasks, most existing methods either explicitly or implicitly divide these time-series datasets into predetermined observation and prediction windows. Patients have different lengths of medical history and the desired predictions (for purposes such as diagnosis or treatment) are required at different times in the future. In this paper, we propose a method that uses a sequence-to-sequence generator model to transfer an input sequence of EHR data to a sequence of user-defined target labels, providing the end-users with "flexible" observation and prediction windows to define. We use adversarial and semi-supervised approaches in our design, where the sequence-to-sequence model acts as a generator and a discriminator distinguishes between the actual (observed) and generated labels. We evaluate our models through an extensive series of experiments using two large EHR datasets from adult and pediatric populations. In an obesity predicting case study, we show that our model can achieve superior results in flexible-window prediction tasks, after being trained once and even with large missing rates on the input EHR data. Moreover, using a number of attention analysis experiments, we show that the proposed model can effectively learn more relevant features in different prediction tasks.

Introduction

As more healthcare systems adopt standardized methods of collecting health data in electronic health record (EHR) formats, unprecedented opportunities for applying AI/ML methods on these datasets have been arising. By informing various types of interventions (from prevention to treatment), the application of such data-driven methods offers great hopes to shift healthcare systems in almost every aspect and makes achieving the ultimate precision medicine goals more promising. A popular application of EHR datasets is in developing predictive models to estimate future clinical outcomes of interest. Due to the temporal nature of EHRs, existing predictive models commonly define one or multiple observations (input period) and prediction windows (output period) on the data. As an example of such an approach, consider a popular study by Liu, Zhang, and Razavian (2018) that uses a 12-month observation window and a 6-month prediction window for future disease prediction. Having pre-defined observation and prediction windows creates several challenges that include handling 1) the EHR data that is not available for the entire length of a defined observation window, 2) the needed predictions at different times for different patients, and 3) having patients with very different lengths of medical history. Generally, existing methods involve training separate models for different observation and prediction windows; and for any of the fixed-time predictions, only samples that have medical histories of the length of the observation window and have the output label at the prediction time are used. A possible solution can be using the rich body of methods for imputing EHRs, by learning to impute the data anywhere inside the observation or prediction windows and hence achieving flexible-window predictions, such as the work by Cao et al. (2018) and Luo et al. (2018). However, the performance of such methods is not as good as the earlier methods for fixed-window predictions, as the latter methods are not designed for prediction tasks. A major gap in the literature remains in having a clear and explicit strategy for creating flexible-window predictive models that do not need to be re-trained separately for different windows. A "flexible-window" prediction design not only has technical advantages but also has important clinical relevance in studying chronic diseases that require knowing the disease trajectories at various times before and following the disease onset.

In this study, we aim to address such limitations by presenting a general method for creating a predictive model that is trained once and can be later flexibly used for different observation and prediction window lengths. Following a sequence-to-sequence theme, we train a model on the entire

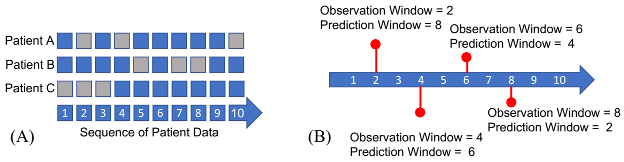


Figure 1: (A) An example of the observed and missing data configurations (gray shows missing). (B) Flexible observation and prediction windows for a patient with 10 yrs of data.

(training portion of the) EHR dataset, where different patients have different lengths of medical histories recorded at irregular intervals. By learning to work with different lengths of input sequences and predicting output labels at different times in the future, our model learns to generate flexible-window predictions at the time of test or deployment. As an example, consider a scenario where a 10-yr EHR dataset is available from a cohort of patients. During training, our model learns to predict the output labels for the complete 10-yr, while the complete 10-yr sequence might not be available for all patients at different timestamps (Figure 1-A). During the deployment, a user can set the observation window to any length; for instance, by using 2 yrs, the model only looks at 2 yrs of medical history (using the data different from the training data) and predicts the next 8 yrs of output labels. Other observation and prediction window configurations can be also defined, similarly (Figure 1-B).

Additionally, to handle the missing EHR data values (including labeled and unlabeled samples), we train the sequence-to-sequence model in a semi-supervised manner by using a generative adversarial network (GAN) architecture. Accordingly, the complete model presented in this study has two parts: (a) a sequence generator network, which generates a sequence of outputs from a sequence of medical history, and (b) a discriminator network, which distinguishes between the actual (observed) and fake outputs (missing in the data and generated by the sequence generator network). The primary contributions of this paper are as follows:

- We present a sequence-to-sequence model that learns jointly from the present and missing samples in the EHR dataset to predict desired clinical outcomes.
- Following the training, our model can be used with any desired size of observation and prediction windows. This approach not only reduces the needed resources for developing and maintaining the models but more importantly, gives complete flexibility to the end-users (e.g., providers) to choose the desired length of windows, facilitating a more successful deployment of our tools.
- We evaluate our models using two separate EHR datasets collected from different sites in the US with around 34,000 adult and 70,000 pediatric patients and show that our model outperforms several other baselines in flexible prediction tasks.

Related Work

Many studies have used EHR data to predict clinical outcomes, while machine learning and deep learning techniques are common choices for building such models. On the technical side, what our study adds to the current state of the field is a combination of RNN and GAN, where an RNN-based sequence-to-sequence model is enhanced to use the generative adversarial loss to learn from both labeled and unlabeled samples. RNN-type models have been a popular choice for creating predictive models using longitudinal data. An early example is the Deep patient model by Miotto et al. (2016) that uses RNN autoencoders to learn the patient representations to predict the disease outcomes at varying prediction windows. Similarly, Chen et al. (2018) used RNNs on ICU data for multi-task prediction, and Choi et al. (2017) used several different observation and prediction windows to predict heart failure using RNNs, among many examples of this kind (Ramazi et al. 2021). This type of study considers separate models trained for each prediction window using the labeled samples for that prediction window. In addition to creating an overhead, this configuration reduces the ability of the models to learn from all available patterns in a dataset. There are also some studies that use RNNs for multivariate time-series imputation tasks, such as using bidirectional RNNs to impute missing values in time-series (Cao et al. 2018; Luo et al. 2018). These studies also relate to our work, as they learn from the time-series data and the output is also in the form of time-series data. Though, these models are used for imputation and cannot be directly applied in predictive scenarios.

We also use a GAN-based design in our model to train our model using semi-supervised learning. GANs have been widely used on clinical data due to their generative capabilities to generate synthetic EHR longitudinal data (Baowaly et al. 2019; Lee et al. 2020; Chin-Cheong, Sutter, and Vogt 2019). For example, ehrGAN (Che et al. 2017) and SMOOTH-GAN (Rashidian et al. 2020) are used to generate labeled data by mimicking the real labeled patient records. Similar to our work, recently the generative capabilities of GANs have also been explored for semi-supervised learning. For instance, McDermott et al. (2018) use adversarial loss and cyclical reconstruction loss to predict individualized treatment effects. Instead of cyclical reconstruction loss, we use the adversarial loss to distinguish between real and fake labels for labeled and unlabeled samples.

Problem Setup

EHR datasets can contain various types of elements, including static information such as demographics, and dynamic information such as medical conditions, medications, measurements, and lab tests. While we formulate our problem using only two more commonly used elements (conditions and medications), the method should be generalizable to other EHR configurations, too.

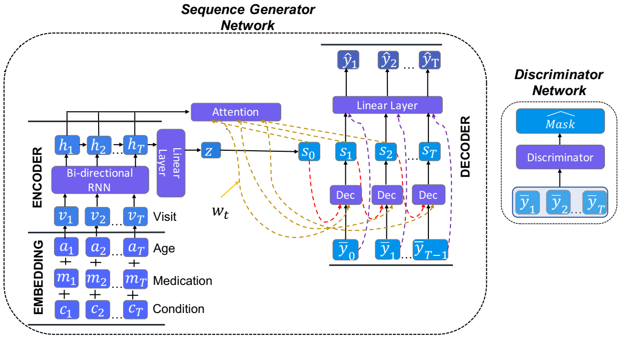

Sequence Generator Network
ENCODER
EMBEDDING
DECODER
Discriminator Network
$\hat{y}_1$ $\hat{y}_2$ ... $\hat{y}_T$
Linear Layer
Attention
h_1 h_2 ... h_T
Bi-directional RNN
Linear Layer
z
s_0 s_1 s_2 ... s_T
Dec Dec Dec
w_t
v_1 v_2 ... v_T Visit
a_1 a_2 ... a_T Age
m_1 m_2 ... m_T Medication
c_1 c_2 ... c_T Condition
$\overline{y}_0$ $\overline{y}_1$ ... $\overline{y}_{T-1}$
$\widehat{Mask}$
Discriminator
$\overline{y}_1$ $\overline{y}_2$ ... $\overline{y}_T$

Figure 2: Our model's structure. Red arrows show the last hidden state of the decoder. The yellow arrows show the weighted vector, given to the decoder, and the purple arrows show $\overline{y}_t$, given to the decoder's linear layer.

We first define three vectors: C, M, and A. Let $C = \{c_1,\ c_2 \ldots, c_T\} \in \mathbb{R}^{T \times N}$ be a multivariate time-series vector that represents the conditions recorded over a patient's visits, where $T = (1,\ 2, \ldots .T)$ shows the visit timestamps, and the t-th condition vector, $c_t \in (0,1)^N$ is a one-hot vector for N unique conditions. If the i-th condition is recorded in the t-th visit, then $c_{ti} = 1$, and 0 otherwise. Finally, let $M = \{m_1,\ m_2 \ldots, m_T\} \in \mathbb{R}^{T \times D}$ be a multivariate time-series vector that represents the medications recorded over a patient's visits, where the t-th medication vector $m_t \in (0,1)^D$ is a one-hot vector for D unique medications. If the j-th medication is observed in t-th visit, then $m_{tj} = 1$, and 0 otherwise. Let $A = \{a_1,\ a_2 \ldots, a_T\} \in \mathbb{R}^T$ be a univariate time-series vector that represents the patient's age, where $a_t \in \mathbb{R}$ is the age at the t-th visit.

Given a patient's data, C (conditions), M (medications), and A (age) recorded over T timestamps, during training, the model learns to predict the sequence of labels (showing the outcome of interest) over T timestamps. However, during testing, the model only looks at C, M, and A over observation window timestamps and predicts output labels for prediction window timestamps. We show the observed (actual) labels with $Y = \{y_1, y_2, \ldots,\ y_T\}$, and estimated labels as $\hat{Y} = \{\hat{y}_1,\ \ \hat{y}_2, \ldots,\ \hat{y}_T\}$. We align all the temporal data of the patients by the maximum length of medical history in the data and segment the time-series into disjoint time-periods. As an example, Figure 1-A shows 10 yrs of data segmented into one-yr bins. As in our semi-supervised learning setting, not all data samples have labels, we use a binary $mask$ vector to identify the observed $(= 1)$ and missing $(= 0)$ labels in the dataset. The model learns to predict $\hat{Y}$, by learning from both the labeled and unlabeled samples.

Model Architecture

The proposed model consists of two parts (as shown in Figure 2): (1) a sequence generator network (Seq–Gen), which consists of an encoder (Enc) and a decoder (Dec), that together generate a sequence of labels, and (2) a discriminator (Dis), which distinguishes between the observed (actual) and generated labels.

Sequence Generator Network – The two components of this network, Enc and Dec, use GRU (gated recurrent unit) layers, receiving the input time-sequence and generate another time-sequence. The input to Enc is an embedded vector (V) obtained by the summation of the embeddings of C, M, and A, such that:

$$v_t = Emb\left(c_t\right) + Emb\left(m_t\right) + Emb\left(a_t\right), \tag{1}$$

where, Emb is the embedding layer. V is then given to Enc, where bidirectional GRU layers are used to produce the output in the forward and backward directions:

$$\overrightarrow{h_t}, \overleftarrow{h_t} = Enc\left(v_t, h_{t-1}\right), \tag{2}$$

where, h_{t-1} is the previous hidden state, and $\overrightarrow{h_t}$, and $\overleftarrow{h_t}$ are the output of the bidirectional GRU layer in the forward and backward direction, respectively. This output is then concatenated into H (Enc's output):

$$h_t = |\overrightarrow{h_t}, \overleftarrow{h_t}|, \tag{3}$$

$$H = \{h_1, h_2, \ldots, h_T\}. \tag{4}$$

The last hidden state h_T is passed through a fully connected linear layer and then a $Tanh$ activation, to obtain a context vector z, creating the first hidden state s_0, and is given to Dec.

An attention mechanism is used to obtain the weighted sum of the encoded vector H, similar to other studies (Chen et al. 2018). The attention layer ($attn$) obtains the attention scores for the encoded vectors, and learns to assign a higher weight to the subset of vectors in H that are more important in predicting the output at each timestamp, as shown below:

$$attn\_score_t = Softmax\left(Tanh\left(attn\left(H,\ s_{t-1}\right)\right)\right). \tag{5}$$

This step calculates attention scores for H with each hidden state s_{t-1} of Dec. Applying $Softmax$ ensures that the attention scores lie between 0 and 1, and sum to 1. Finally, the weight vector w_t, which is the weighted sum of H, is obtained by using $attn\_score_t$ as the weights:

$$w_t = attn\_score_t \odot H. \tag{6}$$

Dec is trained to predict $\hat{y}_t$ by using w_t, s_{t-1}, and the embedded input at timestamp $t-1$, or $Emb\left(\overline{y}_{t-1}\right)$:

$$s_t = Dec\left(w_t,\ s_{t-1},\ Emb\left(\overline{y}_{t-1}\right)\right), \tag{7}$$

$$\overline{y}_{t-1} = mask\ \odot\ y_{t-1} + (1 - mask) \odot\ \hat{y}_{t-1}, \tag{8}$$

where, $\overline{y}_{t-1}$ is the contextual input obtained by using $mask$ at timestamp $t-1$. According to Eq. 8, $\overline{y}_{t-1}$ is equal to the present label y_{t-1}, if the labels are observed at timestamp $t-1$, and equal to the predicted label $\hat{y}_{t-1}$, if the labels are missing. Note that during the test phase, $\overline{y}_{t-1}$ will always be equal to the predicted label $\hat{y}_{t-1}$, as the model needs to predict the output sequence without the knowledge of the ground truth. Eq. 9, shows the calculation of the final output label $\hat{y}_t$ at time t using $\overline{y}_{t-1}$, w_t, and s_t, through the fully connected layer f with a linear activation.

$$\hat{y}_t = f\left(Emb\left(\overline{y}_{t-1}\right),\ w_t,\ s_t\right) \tag{9}$$

Discriminator – By adding the discriminator (Dis) to Seq–Gen presented so far, we implement a generative adversarial framework, where Seq–Gen acts as a generator and Dis learns to determine which labels are observed (actual) or fake (labels generated by Seq–Gen). This discrimination task amounts to predicting the mask vector ($mask$):

$$\widehat{mask} = Dis\left(\overline{y}_t\right), \tag{10}$$

where $\widehat{mask}$ is the predicted mask values by Dis. Using this generative adversarial setting improves the performance of the model by learning the overall distribution of data.

Loss Definitions – For training the entire model jointly on the labeled and unlabeled data, we define a two-part loss function. The first part ensures that Seq–Gen's generated labels for the observed values are close to the actually observed ones. This is the masked loss ($loss_M$), which is the supervised cross-entropy (CE) loss capturing the difference between the ground truth labels in the EHR data and the corresponding labels predicted by the Dec. It is calculated by masking the unlabeled samples:

$$loss_M = CE\left(mask \odot y_t, mask \odot \hat{y}_t\right). \tag{11}$$

The second part of the loss function is the generative adversarial loss, which ensures that the labels generated for the missing components can 'fool' the discriminator by producing labels close to the true underlying data distribution. Through this loss, Dis learns to maximize the probability of correctly identifying the observed and generated values, while (Seq–Gen) learns to minimize the probability that Dis correctly identifies the generated values. We define $loss_{\mathcal{D}}$ and $loss_G$ as follows, and train Dis to minimize $-loss_{\mathcal{D}}$ and Seq–Gen to minimize $loss_M + \ loss_G$.

$$loss_{\mathcal{D}} = mask\, log\left(\widehat{mask}\right) + (1 - mask)\, log\left(1 - \widehat{mask}\right) \tag{12}$$

$$loss_G = -\left(1 - mask\right)\ log\left(\widehat{mask}\right) \tag{13}$$

Obesity Case Study

To evaluate the performance of the proposed model, we performed a series of experiments on two large EHR datasets — All of Us and Nemours Pediatric. We used these datasets to predict the obesity status as indicated by the individuals' BMI (body mass index) values using the CDC's definition of obesity (CDC 2001). Using these datasets, we rigorously evaluate our models on adult and pediatric populations, both spanning across different healthcare sites. The entire All of Us workbench used in this project is available on the All of Us Research Program's platform (researchallofus.org) and can be accessed by any registered user. Processed data files for the Pediatric dataset can be made available upon signing a data use agreement. We present a brief overview of the two EHR datasets and the steps taken for extracting the data, but we refer readers to other studies for more details about our cohorts (Gupta et al. 2019, 2021; Gupta and Beheshti 2020) and our design (Pang et al. 2021).

All of Us – We used the EHR portion of the All of Us Research Program (Investigators 2019), which is a publicly available dataset collected from the data donations of over one million adult participants in the US. We included individuals over 20 yrs and with a minimum of 10 yrs medical history. Our final cohort had 11,152 males and 23,074 females, with 610 and 662 unique condition codes, and 1,099 and 1,279 medication codes, respectively.

Nemours Pediatric – We have also used an EHR dataset from Nemours Children Health, which is a large pediatric healthcare system in the US. Our work was approved by a local IRB. We included any child with a minimum 10-yr of medical history from the age of 1 to 10 yrs. Our cohort had 36,874 males and 30,904 females with 1,457 and 1,443 conditions, respectively, and 380 medication codes, for both.

For both datasets, all the conditions and medications were one-hot encoded. We segmented the 10 yrs of data into disjoint 3-month windows, obtaining 40 timestamps per patient. We combined all observations within each window and took the maximum BMI values for each of the segmented windows (as max-BMI carries more clinical information than other comparable aggregation methods). If a patient did not have any visit over a certain 3-month period, that period's timestamp entries were marked as missing. Table 1 shows the configuration of the timestamps for the data. Note that some remaining timestamps did not have corresponding BMI information. We use this data to predict the patients' obesity status for all 40 timestamps.

	All of Us		Nemours Pediatric	
	Male	Female	Male	Female
Total	446,080	922,960	1,474,960	1,236,160
w/ Ob.	13,953	33,603	52,993	44,038
w/o Ob.	19,698	33,738	236,460	200,648

Table 1: Number of timestamps with obesity (w/ Ob.) and without obesity (w/o Ob.)

We perform our experiments using 5-fold cross-validation with 80:20 train:test data ratio and report the average performances using the $loss_M$ (Eq. 11), which is the loss calculated for the prediction labels for the timestamps with existing labels in the data. We fix the best model on the validation data (5% of training data) and report the AUROC (area under the receiver operating curve) and AUPRC (area under the precision-recall curve) on the test data. AUROC represents the model's capability of distinguishing between classes and the AUPRC is a useful performance metric for imbalanced data. We also report the AUPRC baseline (ratio of the positive over total instances) which is a baseline performance for a random estimator. Higher the AUPRC score above its baseline better is the performance. Our code is available on GitHub at https://github.com/healthylaife/FlexPrediction.

Prediction performance analysis — For the classification task described above, we have compared our model to several popular predictive models for analyzing EHR datasets. These baseline models are built to learn from the labeled samples with fixed observation and prediction windows. The baseline models included RETAIN (Choi et al. 2016), tLSTM (Baytas et al. 2017), Dipole (Ma et al. 2017), and StageNet (Gao et al. 2020). We used Python library implementation of these baseline models (Zhao et al. 2021).

We compare the performance of our proposed model against the baselines for an observation window of 3-yrs and a prediction window of 7-yrs, where the target label is predicted for all timestamps in the 7-yr prediction window. All baseline models are built to make a one-time prediction (having or not having obesity). Therefore, to have a fairer comparison between the baselines and the sequence-to-sequence predictions in our proposed model, we train the baselines to predict the outputs at each timestamp in the prediction window. To do this, we train the baselines 28 times (for the predictions at each 3-month timestamp in the 7-yr window), collect the prediction results for all timestamps in the prediction window, and calculate the overall performance metric. Moreover, following an ablation analysis theme, we include another baseline (Seq-Gen) in our experiments, by leaving out the discriminator and removing $loss_G$ and $loss_{Dis}$ from our model. The results in Table 2 show that our proposed model outperforms the baselines. Additionally, the ablation analysis shows that the discriminator helps improve the performance of our model.

Model Variations	All of Us				Nemours Pediatric			
	Male		Female		Male		Female	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
RETAIN	0.60	0.52	0.60	0.60	0.50	0.22	0.53	0.25
	(0.002)	(0.001)	(0.001)	(0.002)	(0.01)	(0.02)	(0.02)	(0.02)
tLSTM	0.62	0.55	0.60	0.61(	0.52	0.22	0.55	0.24
	(0.002)	(0.002)	(0.002)	(0.002)	(0.01)	(0.02)	(0.01)	(0.02)
Dipole	0.63	0.56	0.62	0.63	0.52	0.23	0.59	0.27
	(0.004)	(0.001)	(0.002)	0(0.003)	(0.01)	(0.01)	(0.01)	(0.01)
StageNet	0.62	0.55	0.63	0.62	0.53	0.24	0.59	0.27
	(0.003)	0.55(0.002)	(0.004)	(0.003)	(0.02)	(0.01)	(0.01)	(0.01)
Seq-Gen	**0.71**	0.64	0.70	0.70	0.56	0.24	0.56	0.26
	(0.001)	(0.001)	(0.001)	(0.001)	(0.02)	(0.01)	(0.02)	(0.02)
Proposed	0.70	**0.64**	**0.71**	**0.71**	**0.59**	**0.26**	**0.62**	**0.31**
	(0.001)	**(0.001)**	**(0.004)**	**(0.005)**	**(0.01)**	**(0.01)**	**(0.01)**	**(0.01)**

Table 2: Performance comparison, using 0-3 yrs observation and 3-10 yrs prediction window. Mean (Std).

Obs/Pred	All of Us						Nemours Pediatric					
	Male			Female			Male			Female		
	AUROC	AUPRC	AUPRC Baseline	AUROC	AUPRC	AUPRC Baseline	AUROC	AUPRC	AUPRC Baseline	AUROC	AUPRC	AUPRC Baseline
3/7	0.70	0.64	0.42	0.71	0.71	0.47	0.59	0.26	0.17	0.62	0.31	0.18
5/5	0.70	0.64	0.40	0.74	0.71	0.46	0.63	0.38	0.18	0.67	0.45	0.20
7/3	0.81	0.79	0.49	0.80	0.79	0.49	0.68	0.46	0.20	0.74	0.60	0.22

Table 3: The performance of our model in three scenarios of flexible observation (Obs) and prediction (Pred) windows.

Flexible-window evaluations — We also evaluate the performance of our method for flexible-window predictions, by varying the observation and prediction window sizes as shown in Table 3. We use observation window sizes of 3, 5, and 7 yrs, and corresponding prediction window sizes of 7, 5, and 3 yrs. After the (one-time) training and during the test phase, we delete all data in the prediction window, and the model predicts the labels in the prediction window by looking only at the data in the observation window. We also evaluate the performance of our model for short- to long-term prediction using a fixed observation and varying prediction window sizes. In Table 4, we compare the performance of the model by keeping the observation window fixed at 3 yrs, and using different prediction window sizes ranging from 1 to 7 yrs ahead of the observation window. For the All of Us female dataset, the original cohort did not have any samples with labels for the 4th and 5th prediction yrs. These results demonstrate the capability of the model to reasonably deliver its anytime prediction promise. This is also shown by the performance boost compared to the AUPRC baselines.

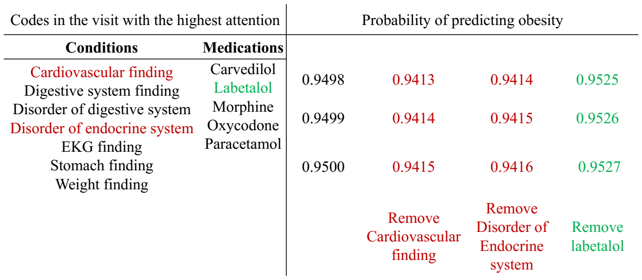
Attention Analysis — Following a practice used in similar studies (Luo et al. 2020), attention analysis was performed to study the "reasonableness" and increase the interpretability of our model. An example of local attention analysis is shown in Figure 3, where we study whether adding and removing medical codes that are known to be clinically relevant indeed change the probability of becoming obese or not. We analyze learned attention weights for one random sample (having obesity) from the All of Us dataset. For this case, we selected the visit with the highest attention weights and print the probability of the case being positive at different time points. We then remove several conditions and medication codes one at a time and check how the removal of the codes affects the probability of the case being positive. We observe that by removing the 'cardiovascular finding' and 'disorder of endocrine system' conditions from this sample, the model's estimate of predicting obesity decreases. In adults, cardiometabolic comorbidities is often associated with obesity (Alpert and Hashimi 1993; Isomaa et al. 2001). On the contrary, when we remove the medication code 'labetalol,' which is a treatment for the conditions 'cardiovascular finding,' the probability of predicting obesity increases. We have also performed a global attention analysis, where we visualize the normalized attention scores for all timestamps in the observation window of 3 yrs to predict the output for the next 7 yrs for all true positive samples. As Figure 4 shows, higher attention is given to the timestamps at the start of the observation window in the All of Us dataset, and towards the end of the observation window in the pediatric dataset. This may be related to the fact that the All of Us dataset contains mostly individuals above 40 yrs of age, where the progression of obesity is more gradual as compared to the pediatric dataset, where obesity develops rapidly due to more body changes (Ward et al. 2017).

Clinical Relevance and Deployment

Clinical implications of our proposed method can be significant and the technique is potentially applicable to a variety of healthcare domains beyond obesity for both short- and long-term predictions. Flexible-window prediction not only

Prediction Year	All of Us						Nemours Pediatric					
	Male			Female			Male			Female		
	AU-ROC	AU-PRC	AUPRC Baseline	AU-ROC	AU-PRC	AUPRC Baseline	AU-ROC	AU-PRC	AUPRC Baseline	AU-ROC	AU-PRC	AUPRC Baseline
4th	0.72	0.66	0.42	-	-	-	0.59	0.25	0.15	0.62	0.30	0.14
5th	0.70	0.66	0.43	-	-	-	0.56	0.23	0.15	0.59	0.30	0.17
6th	0.65	0.62	0.42	0.76	0.72	0.44	0.57	0.25	0.17	0.59	0.30	0.18
7th	0.68	0.64	0.40	0.72	0.68	0.45	0.58	0.28	0.18	0.55	0.30	0.20
8th	0.70	0.59	0.42	0.72	0.70	0.48	0.56	0.26	0.20	0.56	0.32	0.22
9th	0.78	0.64	0.31	0.74	0.77	0.50	0.56	0.29	0.22	0.53	0.30	0.23
10th	0.71	0.55	0.36	0.67	0.72	0.49	0.57	0.29	0.24	0.48	0.25	0.24

Table 4: Our model's performance with a fixed observation window of 3 yrs and varying prediction windows of 4th to 10th yrs.



Codes in the visit with the highest attention		Probability of predicting obesity			
Conditions	**Medications**				
Cardiovascular finding Digestive system finding Disorder of digestive system Disorder of endocrine system EKG finding Stomach finding Weight finding	Carvedilol Labetalol Morphine Oxycodone Paracetamol	0.9498 0.9499 0.9500	0.9413 0.9414 0.9415	0.9414 0.9415 0.9416	0.9525 0.9526 0.9527
			Remove Cardiovascular finding	Remove Disorder of Endocrine system	Remove labetalol

Figure 3: Attention analysis on one positive sample. Each row of numbers on the right shows the predicted probability of obesity. Removing the factors positively associated with obesity (show in red), decreases the probability of obesity. An opposite is seen when medication is removed (green).

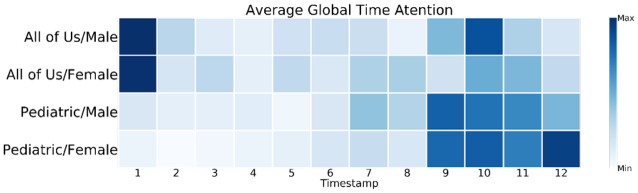


Figure 4: Average global time attention scores for the 12 timestamps in the 3-yr observation window.

can reduce the needed resources for developing and maintaining the predictive models (as it is trained once), but also offers flexibility to the end-users to run the queries of interest based on different lengths of available data. Short-term predictions alert clinicians about the need to intervene more quickly and with more intensive interventions. On the other side, long-term predictions allow providers to intervene as early as possible, increasing the likelihood of success.

We have started taking the necessary steps for deploying our models in actual clinical settings in Nemours Children Health. These steps include performing a preliminary outcome-action-pairing analysis based on the generated predictions of the model and studying the feasibility of the needed interventions for different scenarios. We are also developing a clinical decision support tool prototype to identify the primary care patients at risk for the development of obesity. This includes adding a small software module to the providers' dashboards, generating on-demand predictions. While our model can be used to create the common 'best practice alert' popups, we avoided such design, due to the concerns about the frequency of similar alerts in the clinical settings. This step will focus on capturing the providers' feedback, and in another similar thread, we are forming a small patient engagement group to assess the patients' viewpoints on the relevance of the generated predictions. Another remaining step relates to evaluating our models in a prospective (versus retrospective) setting. This step would be also related to a golden last step before the system-wide deployment that involves a small clinical trial to evaluate the effectiveness of the tool, by comparing the outcomes between a small group of patients that the tool was used on them versus the control group.

Conclusion

In this study, we presented a flexible-window model for obtaining *anytime* predictions using EHR data. To address EHR irregularities and missingness, we used a semi-supervised and adversarial design to improve the model's efficiency. Using two large EHR datasets from separate adult and pediatric populations in a case study aimed at predicting obesity, we showed that our proposed model outperforms several other popular baselines. Due to the flexibility of our model, it only needs to be trained once to then predict the desired output labels for all timestamps. After the training and during the deployment, the observation and prediction window can be set for any length and our model only looks at the data from the timestamps in the observation window of deployment data and then predicts the output label for each timestamp in the prediction window. While other existing methods are applicable to this problem, our method offers a more natural and effective way to study chronic diseases, where knowing the disease trajectories across various future periods can inform earlier and more effective interventions.

Acknowledgments

Our study was supported by NIH awards, P20GM103446 and P20GM113125, and RWJF award 76778. The All of Us Research Program is supported by multiple awards from NIH. In addition, this program would not be possible without the partnership of its participants.

References

Alpert, M. A.; and Hashimi, M. W. 1993. Obesity and the Heart. *The American Journal of the Medical Sciences*, 306(2): 117–123.

Baowaly, M. K.; Lin, C.-C.; Liu, C.-L.; and Chen, K.-T. 2019. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3): 228–241.

Baytas, I. M.; Xiao, C.; Zhang, X.; Wang, F.; Jain, A. K.; and Zhou, J. 2017. Patient subtyping via time-aware lstm networks. In *Proc. of the ACM SIGKDD conf. on knowledge discovery and data mining*, 65–74.

Cao, W.; Wang, D.; Li, J.; Zhou, H.; Li, L.; and Li, Y. 2018. BRITS: Bidirectional Recurrent Imputation for Time Series. In *Advances in Neural Information Processing Systems*, volume 31.

CDC. 2001. Center for Disease Control - Data Table of BMI-for-age Charts.

Che, Z.; Cheng, Y.; Zhai, S.; Sun, Z.; and Liu, Y. 2017. Boosting deep learning risk prediction with generative adversarial networks for electronic health records. In *IEEE Conf. on Data Mining*, 787–792.

Chen, W.; Wang, S.; Long, G.; Yao, L.; Sheng, Q. Z.; and Li, X. 2018. Dynamic illness severity prediction via multi-task rnns for intensive care unit. In *IEEE Conf. on Data Mining*, 917–922.

Chin-Cheong, K.; Sutter, T.; and Vogt, J. E. 2019. Generation of heterogeneous synthetic electronic health records using GANs. In *Workshop on Machine Learning for Health at the Conf. on Neural Information Processing Systems*.

Choi, E.; Bahadori, M. T.; Kulas, J. A.; Schuetz, A.; Stewart, W. F.; and Sun, J. 2016. RETAIN: An Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism. In *Proc. of the Conf. on Neural Information Processing Systems*, 3512–3520. ISBN 9781510838819.

Choi, E.; Schuetz, A.; Stewart, W. F.; and Sun, J. 2017. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2): 361–370.

Gao, J.; Xiao, C.; Wang, Y.; Tang, W.; Glass, L. M.; and Sun, J. 2020. Stagenet: Stage-aware neural networks for health risk prediction. In *Proc. of The Web Conf.*, 530–540.

Gupta, M.; and Beheshti, R. 2020. Time-series Imputation and Prediction with Bi-Directional Generative Adversarial Networks. arXiv:2009.08900.

Gupta, M.; Phan, T.-L. T.; Bunnell, H. T.; and Beheshti, R. 2021. Concurrent Imputation and Prediction on EHR data using Bi-Directional GANs. In *Proc. of the ACM Conf. on Bioinformatics, Computational Biology, and Health Informatics*, 1–9.

Gupta, M.; Phan, T.-L. T.; Bunnell, T.; and Beheshti, R. 2019. Obesity Prediction with EHR Data: A deep learning approach with interpretable elements. *arXiv preprint arXiv:1912.02655*.

Investigators, A. o. U. R. P. 2019. The “All of Us” research program. *New England Journal of Medicine*, 381(7): 668–676.

Isomaa, B.; Almgren, P.; Tuomi, T.; Forsén, B.; Lahti, K.; Nissén, M.; Taskinen, M.-R.; and Groop, L. 2001. Cardiovascular Morbidity and Mortality Associated With the Metabolic Syndrome. *Diabetes Care*, 24(4): 683–689.

Lee, D.; Yu, H.; Jiang, X.; Rogith, D.; Gudala, M.; Tejani, M.; Zhang, Q.; and Xiong, L. 2020. Generating sequential electronic health records using dual adversarial autoencoder. *Journal of the American Medical Informatics Association*, 27(9): 1411–1419.

Liu, J.; Zhang, Z.; and Razavian, N. 2018. Deep ehr: Chronic disease prediction using medical notes. In *Machine Learning for Healthcare Conf.*, 440–464. PMLR.

Luo, J.; Ye, M.; Xiao, C.; and Ma, F. 2020. HiTANet: Hierarchical Time-Aware Attention Networks for Risk Prediction on Electronic Health Records. In *Proc. of the ACM SIGKDD Conf. on Knowledge Discovery & Data Mining*, 647–656.

Luo, Y.; Cai, X.; Zhang, Y.; Xu, J.; and Yuan, X. 2018. Multivariate time series imputation with generative adversarial networks. In *Proc. of the Conf. on Neural Information Processing Systems*, 1603–1614.

Ma, F.; Chitta, R.; Zhou, J.; You, Q.; Sun, T.; and Gao, J. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proc. of the ACM SIGKDD conf. on knowledge discovery and data mining*, 1903–1911.

McDermott, M.; Yan, T.; Naumann, T.; Hunt, N.; Suresh, H.; Szolovits, P.; and Ghassemi, M. 2018. Semi-supervised biomedical translation with cycle wasserstein regression GANs. In *Proc. of the AAAI Conf.*, volume 32.

Miotto, R.; Li, L.; Kidd, B. A.; and Dudley, J. T. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1): 1–10.

Pang, X.; Forrest, C. B.; Lê-Scherban, F.; and Masino, A. J. 2021. Prediction of early childhood obesity with machine learning and electronic health record data. *International Journal of Medical Informatics*, 150: 104454.

Ramazi, R.; Perndorfer, C.; Soriano, E. C.; Laurenceau, J.-P.; and Beheshti, R. 2021. Predicting progression patterns of type 2 diabetes using multi-sensor measurements. *Smart Health*, 21: 100206.

Rashidian, S.; Wang, F.; Moffitt, R.; Garcia, V.; Dutt, A.; Chang, W.; Pandya, V.; Hajagos, J.; Saltz, M.; and Saltz, J. 2020. SMOOTH-GAN: Towards Sharp and Smooth Synthetic EHR Data Generation. In *Conf. on Artificial Intelligence in Medicine*, 37–48.

Ward, Z. J.; Long, M. W.; Resch, S. C.; Giles, C. M.; Cradock, A. L.; and Gortmaker, S. L. 2017. Simulation of Growth Trajectories of Childhood Obesity into Adulthood. *New England Journal of Medicine*, 377(22): 2145–2153. PMID: 29171811.

Zhao, Y.; Qiao, Z.; Xiao, C.; Glass, L.; and Sun, J. 2021. PyHealth: A Python Library for Health Predictive Models. *arXiv preprint arXiv:2101.04209*.