A GPT-based EHR Modeling System for Unsupervised Novel Disease Detection

Authors: Boran Hao[1], M.S., Yang Hu[1], B.S., William G. Adams[2, 4], M.D., Sabrina A. Assoumou[3, 4], M.D., M.P.H., Heather E. Hsu[2, 4], M.D., M.P.H., Nahid Bhadelia[4, 5], M.D., and Ioannis Ch. Paschalidis[1,6], Ph.D.

[1] Department of Electrical and Computer Engineering, Boston University, Boston, MA, USA

[2] Department of Pediatrics, Boston Medical Center, Boston, MA, USA

[3] Department of Medicine, Boston Medical Center, Boston, MA, USA

[4] Chobanian & Avedisian School of Medicine, Boston University, Boston, MA, USA

[5] Center for Emerging Infectious Diseases Policy and Research, Boston University, Boston, MA, USA

[6] Department of Biomedical Engineering, Division of Systems Engineering, Faculty of Computing & Data Sciences, and Hariri Institute for Computing and Computational Science & Engineering, Boston University, Boston, MA, USA

Word Count: 3934 words

Corresponding Author:

Ioannis Ch. Paschalidis
Rafik B. Hariri Institute for Computing and Computational Science & Engineering,
Boston University,
665 Commonwealth Ave.,
Boston, MA 02215, USA
e-mail: yannisp@bu.edu, Tel: 617-353-0434
http://sites.bu.edu/paschalidis

ABSTRACT

Objective: To develop an *Artificial Intelligence (AI)*-based anomaly detection model as a complement of an “astute physician” in detecting novel disease cases in a hospital and preventing emerging outbreaks.

Methods: Data included hospitalized patients (n=120,714) at a safety-net hospital in Massachusetts. A novel *Generative Pre-trained Transformer (GPT)*-based clinical anomaly detection system was designed and further trained using *Empirical Risk Minimization (ERM)*, which can model a hospitalized patient's *Electronic Health Records (EHR)* and detect atypical patients. Methods and performance metrics, similar to the ones behind the recent *Large Language Models (LLMs)*, were leveraged to capture the dynamic evolution of the patient's clinical variables and compute an *Out-Of-Distribution (OOD)* anomaly score.

Results: In a completely unsupervised setting, hospitalizations for *Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)* infection could have been predicted by our GPT model at the beginning of the COVID-19 pandemic, with an Area Under the Receiver Operating Characteristic Curve (AUC) of 92.2%, using 31 extracted clinical variables and a 3-day detection window. Our GPT achieves individual patient-level anomaly detection and mortality prediction AUC of 78.3% and 94.7%, outperforming traditional linear

models by 6.6% and 9%, respectively. Different types of clinical trajectories of a SARS-CoV-2 infection are captured by our model to make interpretable detections, while a trend of over-pessimistic outcome prediction yields a more effective detection pathway. Furthermore, our comprehensive GPT model can potentially assist clinicians with forecasting patient clinical variables and developing personalized treatment plans.

Conclusion: This study demonstrates that an emerging outbreak can be accurately detected within a hospital, by using a GPT to model patient EHR time sequences and labeling them as anomalous when actual outcomes are not supported by the model. Such a GPT is also a comprehensive model with the functionality of generating future patient clinical variables, which can potentially assist clinicians in developing personalized treatment plans.

Keywords: Novel disease detection; Pandemic prevention; EHR modeling; Deep learning; GPT.

1. INTRODUCTION

Pandemics can be devastating, as the COVID-19 pandemic plainly demonstrated.[1,2] An effective, early novel infectious disease detection system could potentially be used to trigger a response and implement mitigation measures within a hospital, which is essential to prevent a future pandemic. Manual detection of a new disease cluster depends on an *astute physician* realizing the atypical nature of encountered cases, which requires expertise, time, and inquisitiveness.[3,4] *Artificial Intelligence (AI)* processes may provide a scalable solution with the potential to yield more rapid and accurate detection of emerging threats.

The *transformer*[5] architecture is responsible for the most recent successes of AI,[6–8] revolutionizing clinical *Natural Language Processing (NLP)*,[6,9] clinical image analysis,[10,11] *Electronic Health Records (EHR)* modeling,[12] and protein structure prediction.[13] Specifically, the *Generative Pre-trained Transformer (GPT)*[14] is a special type of unidirectional transformer which can generate new sequence instances. A GPT trained on large corpora using *Causal Language Modeling (CLM)* gave rise to *Large Language Models (LLMs)* such as ChatGPT.[8] GPT-based generative AI models do not simply memorize associations between words,[15] but are able to represent the semantics underlying the training data.[16,17]

Despite GPT's outstanding analytical and generative ability showed in a variety of biomedical applications,[18–21] to the best of our knowledge, it has not been applied to structured EHR sequence modeling for *unsupervised novel disease detection*. Current transformer-based EHR anomaly detection and modeling systems were trained using BERT-like methods,[12,22] and they were not successfully applied to detect novel diseases and prevent pandemics. Instead, recent disease detection systems utilized *Convolutional Neural Networks (CNN),*[23–25] and *Recurrent Neural Networks (RNN)*[26] using different types of data and trained in a supervised manner, which achieved decent detection accuracy. However, supervised training limits their generalizability to unknown diseases. Unsupervised COVID-19 detection models based on CNN and variational autoencoders were designed,[4,27,28] though these models relied on lung X-ray or CT scans, which limits their applicability. Pandemic surveillance strategies were based on the population distribution[29] and admission number time series,[30] but these methods cannot detect individual patient anomalies. An unsupervised NLP method was proposed to detect an outbreak in dogs from EHRs,[31] but this approach only used free text and was not applied to human subjects. An *One-Class Support Vector Machine (OCSVM)* approach was applied to spike protein sequences to detect SARS-CoV-2 variants,[32] but did not use EHR data or deep neural networks. The availability of GPT and EHR data creates an opportunity for innovation by training a comprehensive GPT for

EHR modeling, similar to how ChatGPT was developed for natural language tasks.

This study leverages GPT, language modeling schemes, and *Out-Of-Distribution (OOD)* detection to exploit the evolutionary characteristics in EHR time series, seeking to detect anomalies in EHRs associated with a potential new disease or emerging pandemic. Clinical variables in the EHRs were separated into consecutive periods to form a sequence. We used a *Causal (electronic health) Record Modeling (CRM)* method based on the CLM for GPT *Language Models (LM)*. To compute OOD anomaly scores for each patient, we modified the perplexity metric in language modeling to evaluate how well a new EHR sequence fits models pre-trained on the in-distribution EHR.

Statement of Significance

Issue	Rapid detection of a novel disease is essential in pandemic prevention, though traditional detection requires much expertise, time and inquisitiveness.
What is already known	Transformer-based AI models show great strength in biomedical tasks including *Electronic Health Record (EHR)* modeling, while they have hardly been applied to real-world novel disease detection.

What this paper adds	This study aims to build a *Generative Pre-trained Transformer (GPT)*-based structured EHR modeling system for real-time, unsupervised novel disease detection and hospital-level outbreak surveillance, which requires no prior knowledge of new diseases. This comprehensive EHR model can also potentially support a variety of EHR-based prediction tasks.

2. MATERIAL AND METHODS

2.1. Data description

Our data set contains de-identified EHRs of 120,714 hospitalized patients from the *Boston Medical Center (BMC)*, which is the largest safety-net hospital in New England. Each patient may have multiple admissions; overall 230,026 admissions from January 1, 2016 to January 24, 2023 were included. For each admission, 31 clinical variables were extracted, including SARS-CoV-2 test results, vital signs, demographics, past medical history, hospitalization status, *Intensive Care Unit (ICU)* status, mechanical ventilation records, and death (see Supplement for details).

2.2. Clinical rationale

Predictive modeling exploits correlations between clinical variables and outcomes of interest.[33,34] Earlier work[35,36] found that patient outcomes such as ICU admission and death can be predicted at an early stage of the infection by a few predictors, including vital signs and laboratory results. More complex

models that captured the evolution of a patient's clinical condition showed even stronger predictive power.[37] Given that each disease is associated with specific patterns, modeling these patterns in a structured EHR time series with sequence modeling algorithms has the potential to detect a new disease by identifying cases that do not fit the model.

2.3. EHR sequentialization

We defined a *period* to correspond to τ consecutive hours and separated the patient's EHR into a sequence of periods. We set $\tau = 4$ in this study, since the vital signs were typically collected every 4 hours in the hospital. For a hospitalization with H hours in total, $n = \lceil H/\tau \rceil$ consecutive periods $\{t_1, \dots, t_n\}$ were defined starting from the admission time. In each period, continuous variables (e.g., vital signs) were averaged, while categorical variables (e.g., ICU status) were binarized into $0/1$ indicator variables. Hence, for each variable f out of the m total clinical features/outcomes in a set $\mathcal{M} = \{f_1, \dots, f_m\}$, its value can be extracted for a period t_i, denoted as $x_{t_i}^{f}$, and $\mathbf{x}_{t_i} = \left(x_{t_i}^{f_1}, \dots, x_{t_i}^{f_m}\right) \in \mathbb{R}^m$ is a (column) feature vector for t_i. Thus, each admission is represented by a sequence $\mathbf{x}_{t_1:t_n} = \left\{\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_n}\right\}$. We exclude a specific patient sequence if more than 20% of the continuous variables are missing, and we impute the missing values for a continuous variable using the median of the non-missing values in the training set. Finally, all continuous variables were

standardized using the mean and standard deviation computed from the training set.

2.4. Causal EHR modeling

CLM is a powerful language modeling scheme, which we modify to develop our CRM method for EHR modeling. In NLP, for a text sequence with tokens $\{y_1, \dots, y_n\}$, the CLM training for an LM aims at maximizing the conditional probability of the correct next-token given the previous tokens, i.e., $P(y_i|y_1, \dots, y_{i-1}, \boldsymbol{\theta}_{LM})$, where $\boldsymbol{\theta}_{LM}$ are the parameters of the model. Therefore, the trained LM can generate new text by predicting the next token in an autoregressive way. For EHR modeling, rather than predicting the next token, we use a GPT with parameters $\boldsymbol{\theta}$ to define a discriminative probability model $\mathrm{P}\left(\mathbf{x}_i|\mathbf{x}_{t_1:t_{i-1}}, \boldsymbol{\theta}\right)$ and predict the next-period clinical variables, given only the current and past periods. We make the assumption that due to the EHR data correlations we discussed in Sec. 2.2, all the next-period variables are predictable given the historical and current-period EHR, i.e., a statistical assumption that all the future variables are *conditionally* independent given the past and present. As a result, the parameters can be learned by maximizing the likelihood on all variables and periods (with a history) using a training set of N admissions from the EHR dataset:

$$\max_{\boldsymbol{\theta}} \prod_{k=1}^{N} \prod_{i=2}^{n} \prod_{f \in \mathcal{M}} P\left(x_{k,t_i}^{f}|\mathbf{x}_{k,t_1:t_{i-1}}, \boldsymbol{\theta}\right),$$

where the index k denotes the k-th admission sample. By taking the negative logarithm, we can formulate our CRM problem as an *Empirical Risk Minimization (ERM)* with a loss function defined by the discriminative model:

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^{N} \sum_{i=2}^{n} \sum_{f \in \mathcal{M}} -\log P\left(x_{k,t_i}^{f} | \mathbf{x}_{k,t_1:t_{i-1}}, \boldsymbol{\theta}\right). \qquad (1)$$

Note that our causal EHR modeling is a framework inspired by Causal Language Modeling (CLM) and further trained using ERM, rather than a traditional causal inference method in clinical study. We use GPT to build this discriminative model, which is illustrated in Figure 1. Each $\mathbf{x}_{t_i}$ is first mapped into the initial hidden state $\mathbf{h}_{0,t_i}$, where the positional embedding is also added to discriminate different periods during the hospitalization. For each self-attention layer in GPT, the previous layer's output hidden states will be first mapped into Query, Key and Value vectors $\mathbf{Q}_{t_1}, \mathbf{K}_{t_1}$ and $\mathbf{V}_{t_1}$ for each period, and the attention score of t_j w.r.t t_i is computed as:

$$A_{t_i,t_j} = \begin{cases} \dfrac{\langle \mathbf{Q}_{t_i}, \mathbf{K}_{t_j} \rangle}{\sqrt{\dim(\mathbf{K}_{t_i})}}, & i \geq j, \\ -\infty, & i < j, \end{cases}$$

where $\langle \cdot, \cdot \rangle$ denotes inner product and $\dim(\cdot)$ the dimension of a vector. Therefore, in the updated hidden state $\mathbf{h}'_{t_i} = \sum_{j=1}^{n} \left(\frac{\exp\left(A_{t_i,t_j}\right)}{\sum_{z=1}^{n} \exp\left(A_{t_i,t_z}\right)} \right) \mathbf{V}_{t_j}$, the weights for the future periods after t_i become 0, which forces the multi-head attention mechanism to be unidirectional and implemented only across the current and past periods. The attention layers in GPT exploit the dependency of the clinical variables from different periods, and the hidden state $\mathbf{h}_{t_i}$ for

period t_i output by the last attention layer utilizes all the present/past but no future information, which naturally suits the definition of the loss function in CRM (cf. Eq. (1)). For each continuous variable f, a linear regression layer with parameters $\mathbf{w}_f \in \mathbb{R}^d, b_f \in \mathbb{R}$ is defined to predict the next-period $\hat{x}_{t_i}^f(\boldsymbol{\theta}) = \langle \mathbf{w}_f, \mathbf{h}_{t_{i-1}} \rangle + b_f$, where $\boldsymbol{\theta}$ are all parameters in the entire GPT model (including $\mathbf{w}_f$ and b_f). If f is binary, a softmax classifier is defined instead to derive the predicted positive probability $\hat{x}_{t_i}^f(\boldsymbol{\theta})$. The loss function L is defined by the absolute error and cross-entropy loss between the target and GPT prediction:

$$L_{t_i}^f(\boldsymbol{\theta}) = \begin{cases} \left| x_{t_i}^f - \hat{x}_{t_i}^f(\boldsymbol{\theta}) \right|, & \text{if } f \text{ continuous,} \\ -x_{t_i}^f \log \hat{x}_{t_i}^f(\boldsymbol{\theta}) - \left(1 - x_{t_i}^f\right) \log\left(1 - \hat{x}_{t_i}^f(\boldsymbol{\theta})\right), & \text{otherwise.} \end{cases}$$

Finally, the CRM training for a GPT model amounts to minimizing the total loss:

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^{N} \sum_{i=2}^{n} \sum_{f \in \mathcal{M}} L_{k,t_i}^f(\boldsymbol{\theta}).$$

In practice, we do not predict the past medical history and demographics, but only use them as features to predict other variables, which is because these features are typically time-invariant within an admission.

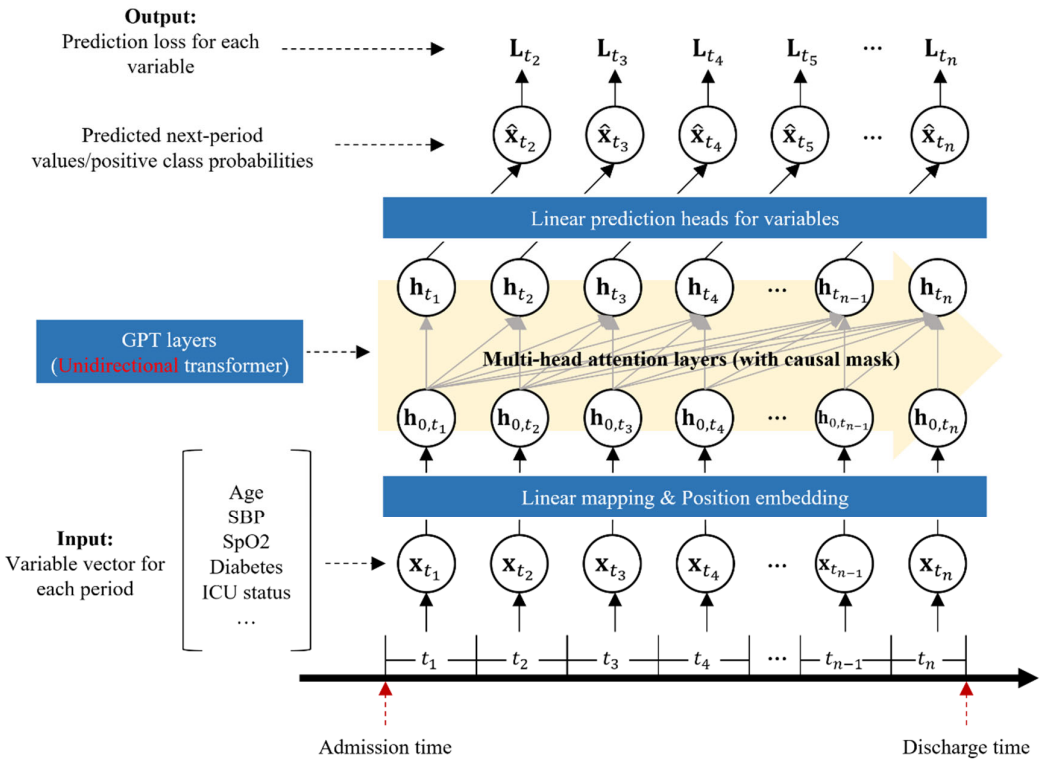


Figure 1. GPT for causal EHR modeling.

2.5. Out-Of-Distribution (OOD) anomaly detection

We regard novel disease detection as an OOD anomaly detection task, which detects data instances that are drawn far away from the training distribution. We use all the EHR data before January 1, 2019 as the in-distribution training data, as COVID-19 that is regarded as the novel disease to detect in this work, did not appear before 2019. Therefore, a GPT pre-trained on normal EHR using our CRM method can represent the in-distribution patterns and be used to detect anomalies in a new EHR.

Transformer-based LMs have shown significant distribution-specific characteristics,[9,38,39] with perplexity[38] being a key metric in evaluating how an LM fits the text. We define our perplexity metric to evaluate how our GPT model fits the EHR of a hospitalized patient w.r.t a certain clinical variable f:

$$\mathrm{PPL}^{f}(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{n-1}\sum_{i=2}^{n} L_{t_i}^{f}(\boldsymbol{\theta}).$$

A higher perplexity of an admission indicates a higher chance of OOD, i.e., a potential novel disease. We use the perplexities w.r.t ICU care, mechanical ventilation, and death as the main anomaly scores. To make our detection interpretable, we further designed two types of perplexities to discriminate different abnormal clinical trajectories. We define the *type-1 perplexity* as the perplexity computed only from periods with label 0, which measures the chance of mistakenly predicting a negative outcome as positive (a.k.a type-1 error, or false alarm). Similarly, the *type-2 perplexity* is computed only from the periods with label 1, which measures the chance of mistakenly predicting a positive outcome as negative (a.k.a type-2 error, or miss).

We define a γ-day *interval* T as γ consecutive days, so that the time axis can be separated into a sequence of $\{T_i, i = 1, \dots\}$. To mitigate noise, we further average the perplexities of all the admission cases whose admission time were within the same interval T_i, hence, computing an interval-level perplexity $\mathrm{PPL}_{T_i}^{f}(\boldsymbol{\theta})$. As long as at least one anomaly admission exists in T_i, we regard T_i as an anomaly interval to detect. We design two statistical methods to detect

the anomaly intervals in the time series $\{\mathrm{PPL}_{T_i}^{f}(\boldsymbol{\theta})\}$. In Method 1, we simply assume that during the normal years without a pandemic, the time series $\{\mathrm{PPL}_{T_i}^{f}(\boldsymbol{\theta})\}$ is stationary, where all $\mathrm{PPL}_{T_i}^{f}(\boldsymbol{\theta})$ are i.i.d Gaussian observations. Therefore, we directly use the mean μ and standard deviation σ of all $\mathrm{PPL}_{T_i}^{f}(\boldsymbol{\theta})$ in the training set to define the threshold, and detect all the T_i in the test set with $\mathrm{PPL}_{T_i}^{f}(\boldsymbol{\theta}) > \mu + \delta\sigma$ as anomalies, where δ is a scalar affecting the confidence. In Method 2, the stationary assumption does not hold, so we train a differenced autoregressive model (i.e., $\mathrm{ARIMA}(p, d, 0)^{40}$) on the training set to forecast every $\mathrm{PPL}_{T_i}^{f}(\boldsymbol{\theta})$ using the previous intervals, and the prediction absolute residual time series $\{|R_{T_i}^{f}(\boldsymbol{\theta})|\}$ is used to compute appropriate μ, σ and detect the anomaly intervals in the same manner as in Method 1.

3. RESULTS

We regard COVID-19 as the novel disease to detect. Based on whether the patient has a positive COVID-19 test result within 7 days before or after the admission time, we label each admission as 1 (anomaly) or 0 (normal). Table 1 shows characteristics of admitted individuals.

Table 1. Characteristics of individuals admitted.

	All admissions	**Admissions before 1/1/2020**	**Admissions after 1/1/2020**
Mean age	48.56	46.91	51.21
Female %	53.82%	54.08%	53.41%
Race - Black %	42.37%	42.12%	42.77%

Ethnicity - Hispanic or Latino %	23.63%	23.17%	24.35%
Intensive care %	15.88%	16.73%	14.53%
Mechanical ventilation %	4.28%	4.22%	4.39%
Death %	1.63%	1.37%	2.03%
SARS-CoV-2 infection %	2.39%	0.00%	6.21%

The EHR data between January 1, 2019 and May 15, 2020 were used as the test set, which covered the epidemic and initial pandemic stage of COVID-19. Table 2 shows the individual-level mortality prediction and COVID-19 detection *Area Under the Receiver Operating Characteristic Curve (AUC)* of different methods on the test set, as well as the 1-day interval-level ICU anomaly detection AUC. f indicates the variable used for perplexity computing. The GPT predicts mortality with a high AUC of 94.7% on the test set. Using the general perplexity, an unsupervised detection AUC of 77.7% is obtained by GPT when the intubation outcome is chosen, while the best individual-level COVID-19 detection AUC of 78.3% is achieved by GPT when all variables and the type-1 perplexity of the ICU outcome were used. When vital signs are excluded from variables and the type-2 perplexity computed from ICU prediction is used, a COVID-19 detection AUC of up to 65.7% can be obtained. The ICU type-1 perplexity achieves a high interval-level detection AUC of 87.2% when a 1-day interval is applied for COVID detection, which indicates a strong detection power for novel diseases.

Table 2. Model performance for mortality prediction and COVID-19 detection.

Model	Variables	Anomaly score	Mortality prediction AUC	COVID detection AUC			
				f = Death	f = Intubation	f = ICU	f = ICU (daily)
Kernel OCSVM	All	- Distance to decision boundary	-	0.735	0.739	0.737	0.768
Linear regression / classification	w/o vitals	PPL	0.722	0.570	0.618	0.582	0.714
		PPL (type-1)		0.521	0.513	0.469	0.462
		PPL (type-2)		0.605	0.613	0.648	0.688
	All	PPL	0.857	0.658	0.668	0.696	0.715
		PPL (type-1)		0.623	0.601	0.717	0.787
		PPL (type-2)		0.535	0.584	0.447	0.419
GPT	w/o vitals	PPL	0.750	0.615	0.624	0.595	0.715
		PPL (type-1)		0.573	0.505	0.478	0.404
		PPL (type-2)		0.614	0.625	0.657	0.675
	All	PPL	**0.947**	**0.772**	**0.777**	0.722	0.807
		PPL (type-1)		0.754	0.775	**0.783**	**0.872**
		PPL (type-2)		0.487	0.457	0.372	0.341

By setting different thresholds δ (cf. Sec. 2.5), we further evaluate the decisions for 3-day intervals in Table 3. Among the 166 intervals in the test set, 21 of them show an anomaly. By using 3-day intervals, COVID-19 detection AUC is overall higher than the individual-level AUC. The best AUC of 92.2% and weighted F1-score of 93% indicates that our GPT model accurately detects COVID-19 as a novel disease at the beginning of the pandemic, without raising serious false alarms before that time.

Table 3. AUC and weighted F1-score of 3-day interval-level COVID-19 detection using different decision thresholds.

Variable for PPL (Type 1)	Threshold	AUC	Weighted F1-score
ICU	$\delta = 3$	0.922	0.930
	$\delta = 4$		0.912
Intubation	$\delta = 3$	0.889	0.941
	$\delta = 4$		0.891
Death	$\delta = 3$	0.861	0.896
	$\delta = 4$		0.909

To better assess the effectiveness of GPT in detecting the COVID-19 anomalies, in Figure 2 we plot perplexity over time. Specifically, Figure 2(a) plots the weekly COVID-19 admissions (as a percentage of overall admissions) and Figure 2(b) plots weekly interval-level type-1 perplexities for the three variables reported in Table 2, along with detection thresholds for Methods 1 and 2 and the ARIMA forecast used in Method 2. As all the EHR data before January 1, 2019 were used as in-distribution training data, the corresponding weekly perplexities are relatively low and stable, and an anomaly week was hardly detected as a false alarm. Between January 1, 2019 and January 1, 2020, COVID-19 was not present in the U.S., and the weekly anomaly scores for all three outcomes also remained low, indicating that our GPT model pre-trained on a large amount of EHR data can generalize well to new EHRs presented to it.

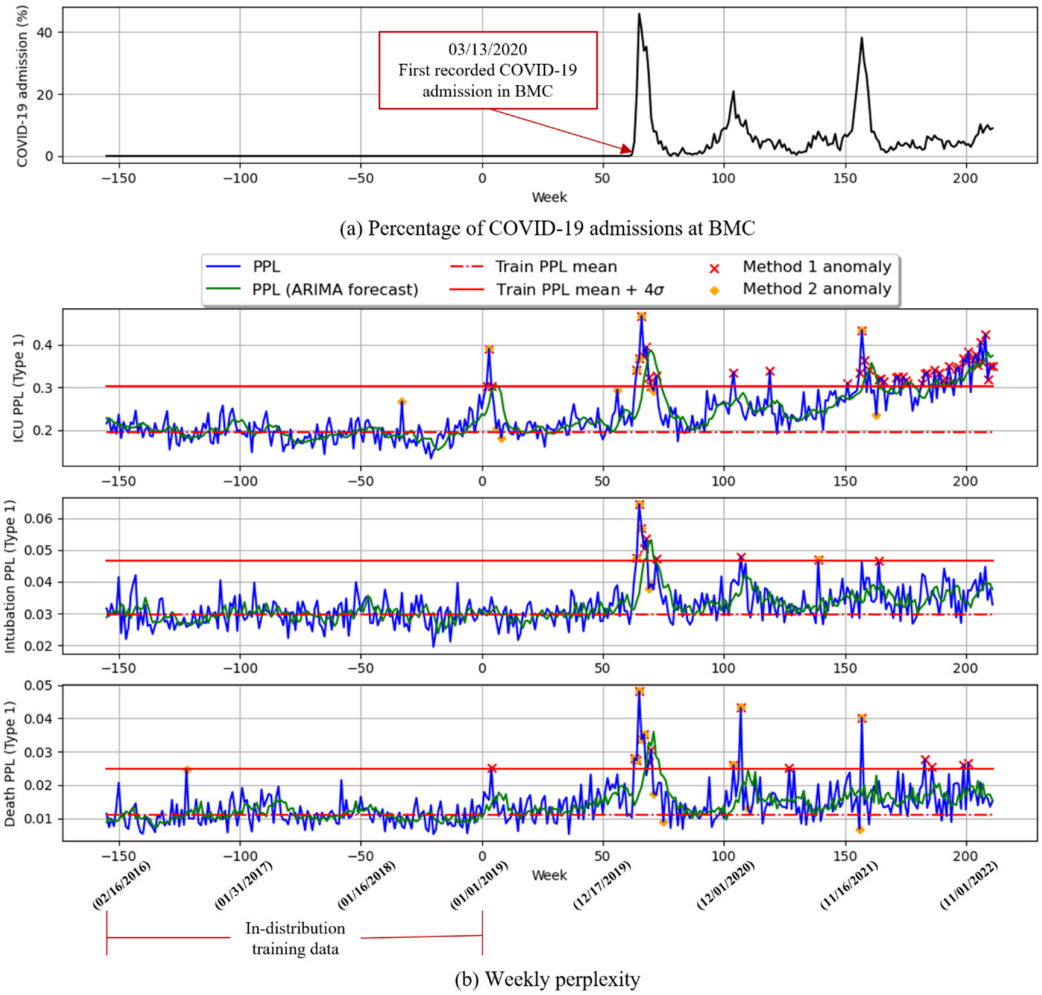


Figure 2. Weekly anomaly scores and detection.

Starting from March 2020, both weekly COVID admission percentage and weekly perplexity began to rise sharply, and a peak emerges. Given that the first COVID-19 admission in BMC was on March 13, 2020 (62.3-th week) and the first anomaly week around this time was the 63-th week detected by the mortality perplexity, our model successfully identified the initial anomaly weeks in the plot, without showing many false alarms. The perplexity peak and COVID admission percentage fell together around June, 2020, potentially due to the

improvements in prevention and treatments, such as mask orders and use of an IL-6 inhibitor (tociluzimab)[41] and dexamethasone.[42] Nevertheless, the weekly perplexity continued showing an overall higher level compared to the time before 2020, and anomaly weeks could still be detected, especially by the ICU perplexity.

4. DISCUSSION

The best mortality prediction and interval-level novel disease detection AUCs achieved by our novel GPT models are 94.7% and 92.2%, respectively, showing that our method can successfully model EHRs and handle both supervised and unsupervised predictive tasks. Note that training our GPT and computing the perplexities require no prior knowledge or labels for a novel disease. Therefore, for practical outbreak surveillance, monitoring the perplexity plots shown in Figure 2 can help detect a new infectious disease and trigger steps to prevent wider spread of infection. Traditional EHR clinical variable analysis typically uses linear models and the latest variables to make interpretable predictions.[35,37] In Table 2, we considered a baseline linear model where we replace the GPT by a simple linear regression/classification layer. GPT outperformed the best linear model by 9% and 6.6% for mortality prediction and individual-level COVID-19 detection, respectively. We also compared our method to Kernel One-Class SVM (OCSVM), which is one of the

most powerful non-linear anomaly detection methods;[32] GPT outperformed OCSVM by 4.4%.

The type-1 perplexity explains most of the detection power of our GPT model when all variables are used, achieving the highest individual-level COVID detection AUC of 78.3% and outperforming the best AUC achieved by type-2 perplexity by 12.6%. The detection ability driven by both type-1 (false alarm) and type-2 (miss) perplexity can be interpreted by the characteristics of COVID-19. On the one hand, although COVID-19 can cause severe symptoms such as high fever and dyspnea which can be reflected in the vital signs, its mortality rate is relatively low (about 3.4% in the U.S. in 2020[43]) compared to other infectious diseases (e.g., Ebola, MERS). Therefore, for a significant proportion of patients, mortality outcomes and related ICU/intubation status can be more favorable than what the historical patterns for patients with similar clinical condition may suggest, which “surprises” the GPT and leads to a high type-1 perplexity. To illustrate the above argument, we present the following case:

“Patient 156932, a 76 y.o. Black female with a past medical history of hypertension, chronic kidney disease and chronic heart disease, was admitted during the 64-th week in Figure 2 (an anomaly week) with a positive SARS-CoV-2 RT-PCR test sampled on the same day. The patient showed an SpO2 93% - 94% at admission and was admitted to ICU and intubated immediately.

In the first 72 hours, the patient remained intubated, while the averaged respiratory rate was mostly above 20. Three days later, the respiratory rate was up to 30, while SpO2 was mostly below 95%. The patient's condition improved after two additional days and she was discharged from the hospital 6 days after admission. Our GPT model for death prediction analyzes the situation in the first 72 hours and predicts that the patient has a high likelihood of dying, so when she was discharged, GPT showed a high type-1 perplexity of 0.31, among the top 3% in the same anomaly week."

On the other hand, patient characteristics such as age, male gender, and underlying conditions such as diabetes are risk factors for severe SARS-CoV-2 infection.[35] When the GPT was trained without vital signs, such demographic information and underlying conditions became the major variables used by the model. Some admissions can surprisingly lead to severe outcomes if the patient is infected by SARS-CoV-2 (e.g., an old male without any prior conditions[35]), which explains the model performance based on type-2 perplexity. Our GPT-based EHR modeling system offers an unsupervised learning framework for detecting novel disease clusters. In practice, depending on the disease, abnormal admissions may be indicated by type-1 or type-2 perplexity or both. Our detection method based on different types of outcome prediction errors and feature sets also injects interpretability and transparency into the system, which

helps clinicians to understand the basic characteristics and trajectories of a novel disease in real-world outbreak surveillance.

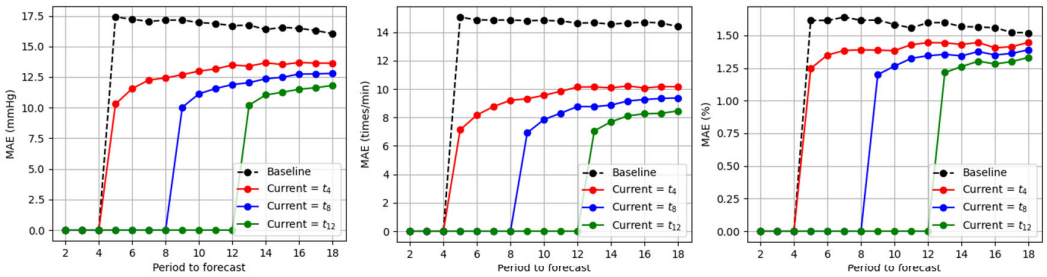
The issue of false alarms is a critical aspect of any outbreak surveillance system. As shown in Figure 2, almost all the anomaly points detected were after March 2020, and few anomalies were detected before COVID-19 appeared in the U.S., which indicates that our GPT-based system has a low false-alarm rate for novel disease detection. Still, one relatively obvious false alarm can be observed around January 2019. This could be explained by the flu season, with the 2018 – 2019 season being unusually long and marked by two separate waves of influenza A.[44] An unusually significant occurrence of a known disease might be detected as a novel disease outbreak as well, which is a cause of false alarms in our system. Nevertheless, Figure 2 and the high weighted F1-score of 93% establish that our current system has a satisfying balance between novel disease detection precision and recall, which effectively reduces false alarms.

Although we trained GPT for novel disease detection, it can also be viewed as a comprehensive EHR predictive model which can, potentially, be used for a variety of clinical purposes. In addition to the accurate mortality prediction we have shown, our GPT model can be applied to generate new EHR sequences and forecast future clinical variables for any given patient. This can be useful in

anticipating the patient's future clinical status and inform a personalized treatment plan.

More specifically, and similarly to the text generation strategy in ChatGPT,[8] we can generate future clinical variables in an autoregressive manner. To initially assess the accuracy of the model's predictions, we tested all admissions between January 1, 2019 and January 1, 2020, with a hospitalization duration of at least 18 periods. In Figure 3, we forecast the sequence of future vital signs including *Systolic Blood Pressure (SBP)*, pulse, and SpO2 at different times during the hospitalization. The later the period we consider as the current period, the longer the history we can use to make the predictions, which, as seen in Figure 3, reduces the prediction error. To quantify this error, we use the *Mean Absolute Error (MAE)* overall tested admissions. Using the mean of the feature values in the training set as the prediction serves as a simple baseline, and GPT predictions always outperform this baseline. Forecasting further into the future with GPT is more difficult than the near future. Moreover, while using a longer history is helpful, the benefit is less significant when predicting the next period. These observations indicate that our model is utilizing the dependencies underlying different periods and variables to make a better long-term prediction. Nevertheless, in this study, the major purpose of developing our GPT architecture and CRM framework is to detect novel diseases, and the extra functionalities (e.g., clinical variable forecasting ability) can be seen as side

benefits brought by our approach. However, more validation is needed for this additional functionality to become useful in a clinical setting.



(a) SBP forecasting error. (b) Pulse forecasting error. (c) SpO2 forecasting error.

Figure 3. Error of future EHR sequences generated by GPT.

Our model is highly applicable since the 31 generic variables we used are available across multiple EHR systems. Compared with previous novel disease detection works[4,24,25,27,28] based on chest radiology data, our model can also capture non-respiratory illnesses. There is a trade-off between using a limited number of variables (which makes the model easier to use and more generally applicable) and training with more variables (e.g., laboratory results and symptoms from clinical notes) larger GPT models (which may improve detection ability). In practice, hospitals can select which features to include in GPT and flexibly determine the cut-off year for EHR training data collection, based on the availability of clinical variables and the intended use. Although EHR data are widely accessible as the input to our model in practical pandemic surveillance, for research purposes, obtaining the accurate timestamps for

variables is challenging due to deidentification requirements. Therefore, the EHR data from other hospital systems were not collected for a multi-center analysis, which is a limitation of this work. Nevertheless, collected from the largest safety-net system in New England, our real-world data set contains a large and racially diverse patient cohort over a span of seven years, which is sufficient to support our method's effectiveness on a hospital-level outbreak surveillance. When more EHR data from multiple hospitals are available in the future, our work can be expanded to a multi-center analysis for broader, population-level outbreak detection.

5. CONCLUSION

In this study, we designed a causal EHR modeling method to build GPT models for EHR-based novel disease detection within a hospital. Our GPT models can potentially be implemented to assist a variety of other clinical tasks, such as EHR sequence embedding, patient outcome prediction and clinical variable sequence forecasting. Future work may consider further validating our GPT model's practical capability in these extra scenarios above, and expanding our hospital-level outbreak detection to a broader population level by introducing EHR data from multiple hospitals.

Acknowledgments: This research was partially supported by the NSF under grants ECCS-2317079, CCF-2200052, DMS-1664644, and IIS-1914792, by

the ONR under grant N00014-19-1-2571, by the DOE under grant DE-AC02-05CH11231, by the NIH under grant UL54 TR004130 and 1UL1TR001430, and by Boston University. We thank Melissa Hofman, MSIS and the BMC Clinical Data Warehouse for Research team for preparing the dataset for our analysis and answering many related questions.

Competing interests: All authors declare no financial or non-financial competing interests.

Author contributions: BH co-processed the data, developed the models, obtained results, and co-wrote the manuscript. YH co-processed the data. WGA, SAA, HEH, and NB provided access to data, offered medical insights, contributed to writing the manuscript, and reviewed the manuscript. ICP designed/led the study, contributed to model development, and co-wrote the manuscript.

Data availability: The de-identified data set used in this study was considered a *limited* dataset for HIPAA purposes. The study was approved by the pertinent *Institutional Review Board (IRB)*, which also waived informed consent. The data that support the findings of this study are available from the Boston Medical Center (BMC) in Massachusetts but restrictions apply to the availability of these data, which were used under license for the current study, and so are not

publicly available. Data are however available from the authors upon reasonable request and with permission of the Boston Medical Center.

Code availability: The underlying code for this study can be accessed via this link https://github.com/noc-lab/gpt_anomaly_detection.

References

1. WHO Coronavirus (COVID-19) Dashboard. Published online 2023. https://covid19.who.int/
2. Hlávka J, Rose A. COVID-19's total cost to the U.S. economy will reach $14 trillion by end of 2023. Published online May 16, 2023. https://healthpolicy.usc.edu/article/covid-19s-total-cost-to-the-economy-in-us-will-reach-14-trillion-by-end-of-2023-new-research/
3. Ajagbe SA, Adigun MO. Deep learning techniques for detection and prediction of pandemic diseases: a systematic literature review. *Multimed Tools Appl*. Published online 2023:1-35.
4. Chharia A, Upadhyay R, Kumar V, et al. Deep-precognitive diagnosis: Preventing future pandemics by novel disease detection with biologically-inspired conv-fuzzy network. *IEEE Access*. 2022;10:23167-23185.
5. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. *ArXiv170603762 Cs*. Published online December 5, 2017. Accessed May 15, 2021. http://arxiv.org/abs/1706.03762
6. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Prepr ArXiv181004805*. Published online 2018.
7. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv Prepr ArXiv201011929*. Published online 2020.
8. OpenAI. GPT-4 Technical Report. *ArXiv*. 2023;abs/2303.08774.
9. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. *ArXiv Prepr ArXiv190403323*. Published online 2019.

10. Felsch M, Meyer O, Schlickenrieder A, et al. Detection and localization of caries and hypomineralization on dental photographs with a vision transformer model. *NPJ Digit Med*. 2023;6(1):198.

11. Yun D, Yang HL, Kwon S, et al. Automatic segmentation of atrial fibrillation and flutter in single-lead electrocardiograms by self-supervised learning and Transformer architecture. *J Am Med Inform Assoc*. 2024;31(1):79-88.

12. Li Y, Rao S, Solares JRA, et al. BEHRT: transformer for electronic health records. *Sci Rep*. 2020;10(1):7155.

13. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589.

14. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. Published online 2018.

15. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ; 2021:610-623.

16. Jin C, Rinard M. Evidence of Meaning in Language Models Trained on Programs. *ArXiv Prepr ArXiv230511169*. Published online 2023.

17. Gurnee W, Tegmark M. Language Models Represent Space and Time. Published online 2023.

18. Frei J, Kramer F. Annotated dataset creation through large language models for non-english medical NLP. *J Biomed Inform*. 2023;145:104478.

19. Guo Y, Qiu W, Leroy G, Wang S, Cohen T. Retrieval augmentation of large language models for lay language generation. *J Biomed Inform*. 2024;149:104580.

20. Guevara M, Chen S, Thomas S, et al. Large language models to identify social determinants of health in electronic health records. *Npj Digit Med*. 2024;7(1):6.

21. Chen A, Chen DO, Tian L. Benchmarking the symptom-checking capabilities of ChatGPT for a broad range of diseases. *J Am Med Inform Assoc*. Published online 2023:ocad245.

22. Niu H, Omitaomu OA, Langston MA, et al. EHR-BERT: A BERT-based model for effective anomaly detection in electronic health records. *J Biomed Inform*. Published online 2024:104605.

23. Jain S, Sindhwani N, Anand R, Kannan R. COVID Detection Using Chest X-Ray and Transfer Learning. In: *International Conference on Intelligent Systems Design and Applications*. Springer; 2021:933-943.

24. Kundu R, Singh PK, Mirjalili S, Sarkar R. COVID-19 detection from lung CT-Scans using a fuzzy integral-based CNN ensemble. *Comput Biol Med*. 2021;138:104895.

25. Shaik NS, Cherukuri TK. Transfer learning based novel ensemble classifier for COVID-19 detection from chest CT-scans. *Comput Biol Med*. 2022;141:105127.

26. Amin S, Uddin MI, Hassan S, et al. Recurrent neural networks with TF-IDF embedding technique for detection and classification in tweets of dengue disease. *IEEE Access*. 2020;8:131522-131533.

27. Mansour RF, Escorcia-Gutierrez J, Gamarra M, Gupta D, Castillo O, Kumar S. Unsupervised deep learning based variational autoencoder model for COVID-19 diagnosis and classification. *Pattern Recognit Lett*. 2021;151:267-274.

28. Scarpiniti M, Ahrabi SS, Baccarelli E, Piazzo L, Momenzadeh A. A novel unsupervised approach based on the hidden features of Deep Denoising Autoencoders for COVID-19 disease detection. *Expert Syst Appl*. 2022;192:116366.

29. Bhatia S, Lassmann B, Cohn E, et al. Using digital surveillance tools for near real-time mapping of the risk of infectious disease spread. *NPJ Digit Med*. 2021;4(1):73.

30. Fox SJ, Lachmann M, Tec M, et al. Real-time pandemic surveillance using hospital admissions and mobility data. *Proc Natl Acad Sci*. 2022;119(7):e2111870119.

31. Noble PJM, Appleton C, Radford AD, Nenadic G. Using topic modelling for unsupervised annotation of electronic health records to identify an outbreak of disease in UK dogs. *Plos One*. 2021;16(12):e0260402.

32. Nicora G, Salemi M, Marini S, Bellazzi R. Predicting emerging SARS-CoV-2 variants of concern through a One Class dynamic anomaly detection algorithm. *BMJ Health Care Inform*. 2022;29(1).

33. Fang L, Xie H, Liu L, et al. Early predictors and screening tool developing for severe patients with COVID-19. *BMC Infect Dis*. 2021;21(1):1-8.

34. Gallo Marin B, Aghagoli G, Lavine K, et al. Predictors of COVID-19 severity: a literature review. *Rev Med Virol*. 2021;31(1):1-10.

35. Hao B, Sotudian S, Wang T, et al. Early prediction of level-of-care requirements in patients with COVID-19. *eLife*. 2020;9:e60519. doi:10.7554/eLife.60519

36. Hu CA, Chen CM, Fang YC, et al. Using a machine learning approach to predict mortality in critically ill influenza patients: a cross-sectional retrospective multicentre study in Taiwan. *BMJ Open*. 2020;10(2):e033898.

37. Hao B, Hu Y, Sotudian S, et al. Development and validation of predictive models for COVID-19 outcomes in a safety-net hospital population. *J Am Med Inform Assoc*. 2022;29(7):1253--1262. doi:10.1093/jamia/ocac062

38. Yang X, Bian J, Hogan WR, Wu Y. Clinical concept extraction using transformers. *J Am Med Inform Assoc*. 2020;27(12):1935-1942.

39. Hao B, Zhu H, Paschalidis ICh. Enhancing Clinical BERT Embedding using a Biomedical Knowledge Base. In: *Proceedings of the 28th International Conference on Computational Linguistics*. ; 2020:657-661.

40. Box GE, Jenkins GM, Reinsel GC, Ljung GM. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons; 2015.

41. Sinha P, Mostaghim A, Bielick CG, et al. Early administration of interleukin-6 inhibitors for patients with severe COVID-19 disease is associated with decreased intubation, reduced mortality, and increased discharge. *Int J Infect Dis*. 2020;99:28-33. doi:10.1016/j.ijid.2020.07.023

42. Ledford H. Coronavirus breakthrough: dexamethasone is first drug shown to save lives. *Nature*. 2020;582(7813):469-470.

43. Karmakar M, Lantz PM, Tipirneni R. Association of social and demographic factors with COVID-19 incidence and death rates in the US. *JAMA Netw Open*. 2021;4(1):e2036462-e2036462.

44. Scutti S. Longer than usual and M-shaped: CDC says 2018-19 flu season was odd but not as severe as the previous deadly season. Published online June 20, 2019.