Few-Shot Learning with Semi-Supervised Transformers for Electronic Health Records

Raphael Poulain RPOULAIN@UDEL.EDU
Mehak Gupta MEHAKG@UDEL.EDU
Rahmatollah Beheshti RBI@UDEL.EDU
University of Delaware

Abstract

With the growing availability of Electronic Health Records (EHRs), many deep learning methods have been developed to leverage such datasets in medical prediction tasks. Notably, transformer-based architectures have proven to be highly effective for EHRs. Transformer-based architectures are generally very effective in "transferring" the acquired knowledge from very large datasets to smaller target datasets through their comprehensive "pre-training" process. However, to work efficiently, they still rely on the target datasets for the downstream tasks, and if the target dataset is (very) small, the performance of downstream models can degrade rapidly. In biomedical applications, it is common to only have access to small datasets, for instance, when studying rare diseases, invasive procedures, or using restrictive cohort selection processes. In this study, we present CEHR-GAN-BERT, a semi-supervised transformer-based architecture that leverages both in- and out-of-cohort patients to learn better patient representations in the context of few-shot learning. The proposed method opens new learning opportunities where only a few hundred samples are available. We extensively evaluate our method on four prediction tasks and three public datasets showing the ability of our model to achieve improvements upwards of 5% on all performance metrics (including AUROC and F1 Score) on the tasks that use less than 200 annotated patients during the training process[1].

1. Introduction

Recent breakthroughs in foundation models (Bommasani et al., 2021) have enabled new opportunities for healthcare applications, notably using transformer-based models to leverage the longitudinal aspect shared by the natural language and Electronic Health Records (EHRs). Transformer-based models are generally built on the transfer learning paradigm, utilizing pre-training steps (such as masked language modeling) on large datasets of medical records to learn the bidirectional characteristics of EHRs. The knowledge acquired during the pre-training stage is then leveraged to fine-tune the model for the downstream tasks. While the mechanism of acquiring patterns from other (larger) sources has helped many transformer-based models achieve state-of-the-art results for EHRs (Li et al., 2020; Pang et al., 2021), the effectiveness of this mechanism can quickly drop on small datasets. As a result, one of the major challenges of most clinical applications using such methods is the lack of labeled data specific to the problem of interest in the downstream task. This is

1. Our code is available at github.com/healthylaife/CEHR-GAN-BERT.

© 2022 R. Poulain, M. Gupta & R. Beheshti.

a pervasive issue in many health applications, as the cohort-definition process can lead to datasets that are orders of magnitude smaller than the starting size of the data. As an example, consider the problem we later study in our experiments aiming to predict the survival patterns following a lung transplant procedure, where, from an original cohort of 120,000 patients, more than 119,000 patients had to be excluded. This would result in a drastic decrease (about 99.81%) in the dataset. Note that existing transformer-based methods do use the entire data, but only during the pre-training phase. Therefore, for downstream tasks, they still rely on larger datasets. It is pivotal to develop enhanced transformer-based methods for healthcare applications in the context of few-shot learning (where only a small amount of data is available). Specifically, few-shot learning aims at using the knowledge from other tasks that can generalize well to downstream predictive tasks.

In this study, we propose CEHR-GAN-BERT to overcome this ubiquitous challenge in machine learning for healthcare by leveraging both in-cohort and out-of-cohort patients' knowledge in an adversarial learning setting. Our model uses BERT (Devlin et al., 2019), which is one of the most popular transformer architectures recently adopted by many for EHR applications (Li et al., 2020; Pang et al., 2021; Prakash et al., 2021). CEHR-GAN-BERT is a BERT-based model coupled with a generative adversarial network (GAN) architecture. Though generative methods like GANs have been very successful in generating data modalities such as images, similar principles cannot be applied directly to EHRs as just one out-of-context medical token could change the entire clinical interpretation of a medical record. For example, as opposed to the imperceptible difference resulting from the changes of some pixels of an image, changing a fever finding to a heart failure disorder can completely change the clinical interpretation of a patient's record. In our proposed method, we train a generator network to reproduce EHR representations as close to the distribution of the ones produced by a BERT architecture and train a discriminator to differentiate between the representations produced by the generator and BERT, identifying the elements that stem out of real EHR data. The proposed architecture allows us to leverage the information of the patients that have been filtered out during the cohort selection process and forces BERT to produce more intricate representations of medical records, yielding new state-of-the-art results on few-shot learning on several popular prediction tasks. Specifically, the contributions of our work can be formulated as follows:

- We introduce a novel semi-supervised transformer-based model that can exploit both in-cohort and out-of-cohort EHR data, allowing the model to learn better data representations when very small patient data is available.
- We demonstrate the efficacy of our method by comparing it against several baselines in a series of extensive experiments involving four prediction tasks and three public datasets.

Generalizable Insights about Machine Learning in the Context of Healthcare

While using popular power analysis methods in machine learning for healthcare is less common, a known "Achilles' heel" of using advanced machine learning methods to achieve competitive performances is access to large datasets. Having access to only a few tens or hundreds of samples (versus thousands and millions in outside healthcare applications) is

very common. The small data size can be the result of the cohort definition process (e.g., by restricting the data to certain ages or having certain comorbidities) or the low prevalence of the condition/disease of interest. This issue leads to a subtle but fundamental bias in machine learning for healthcare applications, where the models' performance degrades on smaller samples, many of which are related to underrepresented populations.

In this study, we aim to investigate the hypothesis that using the data from out-of-cohort as part of the downstream task will improve the performance of transformer-based methods on smaller-size datasets. With the aid of both pre-training and our specific architecture, we demonstrate that promising results can be achieved in downstream medical predictive tasks with as few as 100 annotated patients, opening new opportunities where only little data is available.

2. Related Work

Among a large body of related studies to our work, here, we discuss a non-exhaustive group of studies that are most relevant to the proposed method, specifically the applications of BERT-based models and the semi-supervised GAN-based methods on EHRs.

BERT-based models for EHR – Amongst the most popular transformer methods that have been applied to EHRs is the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). One of the first major EHR applications of BERT is the BEHRT model (after BERT+EHR), where the medical observations were considered as words, visits as sentences, and the records of each patient as a document (Li et al., 2020). This method remains limited in a few aspects: it only considers the diagnosis codes, omitting other data modalities like medications or procedures and it does not include any demographic information. Following the promising success of BEHRT, other studies have extended this architecture. To name a few, Hi-BEHRT (Li et al., 2021) proposed a hierarchical BEHRT model expanding the receptive field of BEHRT to extract associations from longer medical sequences. Med-BERT (Rasmy et al., 2021) removed the age and segment embeddings as well as the use of the [CLS] and [SEP] tokens as used in BERT and BEHRT. The rationale behind such alterations was that the position embeddings should be able to differentiate between the visits. Alternatively, this group has introduced the prolonged length of stay prediction as an additional learning objective to be used alongside Masked Language Modeling (MLM) during the pre-training phase. Though this study outperformed BEHRT in some tasks, no demographics or temporal information has been given to the model. RareBERT (Prakash et al., 2021) has applied Med-BERT's architecture by reintroducing the [CLS] token and developing a temporal embedding capturing the time difference between an event and an anchor date. RareBERT has been trained for feature representation in highly imbalanced datasets to identify patients with a rare disease (X-linked hypophosphatemia, XLH). CEHR-BERT (Pang et al., 2021) has introduced a new embedding process by concatenating the age, temporal and concept embeddings and then using a fully-connected linear layer to create temporal embeddings instead of simply summing all embeddings like the previous BERT-based EHR applications. Both RareBERT and CEHR-BERT have not used demographic information. Meng et al. (2021) presented a way to add gender embeddings and Poulain et al. (2021) added race and ethnicity as well. In this study, we leverage CEHR-BERT's temporal embeddings and input representation

processes with the addition of three demographics embeddings covering gender, race, and ethnicity.

GANs for EHR – While many studies have used GANs for generating synthetic EHRs (Rashidian et al., 2020; Weldon et al., 2021; Sun et al., 2021), using GANs in a semi-supervised learning paradigm has seen growing interest. To name a few Gupta et al. (2021) have used an adversarial learning approach for concurrent imputation and prediction on EHR data, and, based on the same approach developed a semi-supervised learning method for flexible window medical predictions (Gupta et al., 2022). Yang et al. (2019a) have used GANs for categorical EHR imputation, and Che et al. (2017) have introduced ehrGAN to enhance risk prediction performances by generating patients close to the real ones in a semi-supervised learning setting. Similar to our work, Yu et al. (2020) have used unlabeled and labeled patients in a semi-supervised learning setting with RNNs to perform downstream medical tasks. While, to the best of our knowledge, BERT-based architectures and GANs have not been applied together to EHR data, GAN-BERT (Croce et al., 2020) has introduced a semi-supervised architecture with BERT for NLP tasks and has demonstrated state-of-the-art results in few-shot learning while remaining on par with BERT in settings where more labeled data is accessible.

3. Methods

In this study, we propose a novel architecture for semi-supervised learning of medical prediction tasks in an adversarial setting, using both labeled and unlabeled data which proved to be key in few-shot learning. Our method follows the same common training process for transformer-based methods, that is, our model is first pre-trained using Masked Language Modeling (MLM) learning to predict the original content of some tokens that have been masked or replaced. This allows us to leverage larger datasets as this process does not need any supervision, and train our model to better understand the intricacies of medical records. Following this process, the model is fine-tuned for the desired downstream task. This paradigm, greatly popularized by the foundation (language) models, can be directly adapted for EHR data by considering the medical codes as words, visits as sentences, and medical records as documents. Notably, we based our architecture on several BERT-based models for EHR covered above, namely BEHRT (Li et al., 2020) and CEHR-BERT (Pang et al., 2021), as well as GAN-BERT (Croce et al., 2020) which introduced the idea of adding a GAN network to BERT. Figure 1 shows the proposed architecture.

3.1. Temporal Embeddings

We consider a patient's visit as a sequence of medical tokens representing a medical observation (condition, medication, or procedure). Each visit starts and ends with the [VS] and [VE] tokens, respectively. Additionally, each visit is separated by a [ATT] token determining the time gap between the two visits. As an example, consider the sequence below for a fictional patient:

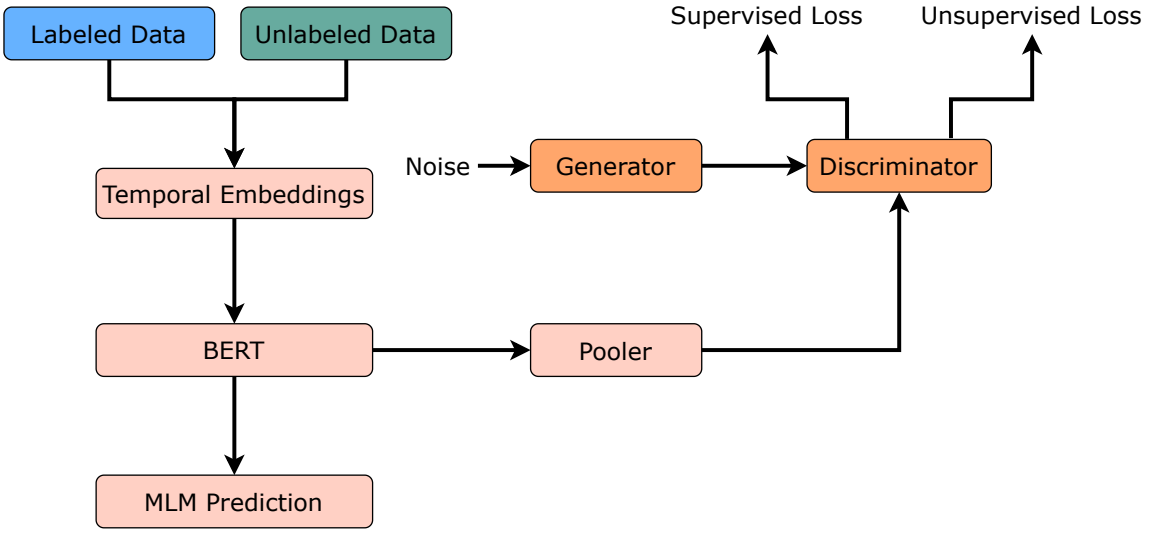


Figure 1: The proposed model's architecture. BERT is the original transformer presented by Devlin et al. (2019), and MLM refers to the masked language modeling used for pre-training. The fine-tuning phase continues with the pooler extracting the representation of the first [VS] token. Simultaneously, the generator tries to mimic the output of the pooler from a noise vector, and the discriminator is trained to perform two tasks: a medical predictive task and differentiation between the real and generated data. "Temporal Embeddings" is shown in more details in Figure 2.

$$VS, C_1, P_2, VE, ATT,$$
$$VS, C_2, M_1, VE, ATT,$$
$$VS, P_1, M_1, VE$$

where C, P, M refer to the condition, procedure, and medication, respectively, and the associated number refers to the index of the given medical code in its respective vocabulary (C_1 would be the first entry of the condition vocabulary). Therefore, this patient has 3 visits recorded with a total medical sequence length of 14.

We include demographics (gender, race, and ethnicity) tokens as numerical values repeated throughout the input sequence. The position tokens define the visit number each token belongs to, and the segment tokens alternate between 0 and 1 from one visit to the other to give the model clearer visit boundaries. We then feed these 6 sequences (medical, gender, race, ethnicity, position, and segment) into separate embedding layers and sum the resulting embeddings to form the concept embeddings as depicted by the blue block in Figure 2. In parallel, we define the age and time tokens to be respectively the age and the day of the year (1-366) each medical concept has been observed. This allows the model to learn patterns of age-related diseases, such as cardiovascular diseases, and seasonal diseases like the flu. Finally, we concatenate the concept, age, and time embeddings to be used as

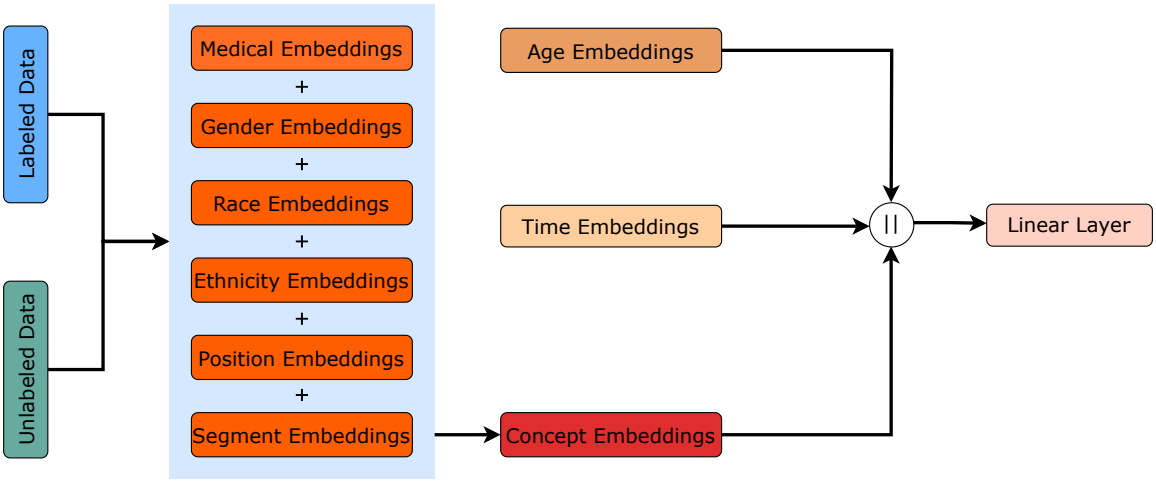


Figure 2: Architecture of the Temporal Embeddings process. Each sequence is fed into a separate embedding layer. The non-temporal embeddings are then summed up together (shown in the blue block) to create the concept embeddings. This summation is then concatenated (shown by the $\|$ symbol) with the Age and Time embeddings resulting in an embedding of size $3H$, where H is the embedding size. This final embedding is then fed into a Linear Layer to reduce its size back to H.

input of a linear layer in order to bring the dimension of the concatenation back to the embeddings' outputs. We define the Temporal Embeddings (TE) to be the output of the linear layer, which will be used as the starting point of the BERT-based architecture. Figure 2 shows the entire process.

3.2. Pre-Training

We pre-train our model using Masked Language Modeling (MLM) (Devlin et al., 2019) by masking some medical tokens to be predicted by the model. More precisely, we randomly chose 15% of the medical tokens to be predicted during this phase. We then replace 80% of those tokens with the [MASK] token (12% of all tokens), 10% of them with another randomly chosen medical token (1.5% of all tokens), and leave the remaining ones unchanged. By doing so, we not only train our model to learn the bidirectional contexts of the medical observations by predicting the [MASK] token but also introduce some noise that forces the model to learn beyond the neighboring conditions of each token by taking into consideration the entire medical record. We pre-trained our model with MLM using the cross-entropy loss on the entirety of the datasets available or generated (before any cohort selection process).

3.3. Fine-tuning in an adversarial setting

During the fine-tuning process, BERT's output is fed to the Pooler that extracts the representation of the first token of the sequence (here the first [VS] token, as opposed to the [CLS] token in BERT). Let us define r_{real} to be the vector of size E after the pooling pro-

cess, containing the representations produced by BERT on real data. The generator G is a multi-layer perceptron (MLP) that takes a noise vector of size N and outputs r_{fake}, which is a vector of fake representations of size E for each patient. Similarly, the discriminator D is another fully connected MLP taking both BERT's and G's outputs ($r_{real} \| r_{fake}$). D is trained to perform two tasks: differentiating between real and fake medical representations, and medical prediction.

As in GAN-BERT (Croce et al., 2020), let us denote $\hat{y}_{fake}$ and $\hat{y}_{real}$ to be the prediction made by D for the fake and real representations, respectively. D is trained to predict $\hat{y}_{fake}$ to approximate 0 and $\hat{y}_{real}$ 1. Thus, we define the discriminator's unsupervised loss to be:

$$\mathcal{L}_{D_{unsup}} = \frac{1}{2}\Big[BCE(\hat{y}_{fake}, 0) + BCE(\hat{y}_{real}, 1)\Big],$$

where BCE is the binary cross-entropy loss. Alternatively, we also train D for the predictive tasks in the form of binary classification for the supervised part of the network. As we leverage both in- and out-of-cohort patients, we introduce a masking vector, $mask = \{m_1, m_2, ..., m_{\mathcal{P}}\}$, where $\mathcal{P}$ stands for the number of patients (in- and out-of-cohort). $mask$ will allow our supervised loss function to only account for the in-cohort patients as follows:

$$m_p = \begin{cases} 1, & \text{if patient } p \text{ is in-cohort} \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, considering $\hat{y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_{\mathcal{P}}\}$ as the estimates vector for every patient, and $y = \{y_1, y_2, ..., y_{\mathcal{P}}\}$ as the ground truth, we define D's supervised loss as the mean binary cross-entropy loss across all patients:

$$\mathcal{L}_{D_{sup}} = \frac{1}{\sum mask} \sum_p BCE(\hat{y}_p, y_p) \odot m_p.$$

The overall discriminator loss $\mathcal{L}_D$ is then calculated as: $\mathcal{L}_D = \alpha \mathcal{L}_{D_{unsup}} + \mathcal{L}_{D_{sup}}$, where α is a hyperparameter to adjust the role of the two objectives (supervised and unsupervised). As we show later, this hyperparameter has a key effect on stabilizing the model on small datasets and achieving competitive results. We note that only the real labeled data is used for the calculation of $\mathcal{L}_{D_{sup}}$ with the addition of $mask$ (an unlabeled patient would have a mask equal to 0, and thus will not affect $\mathcal{L}_{D_{sup}}$), leaving the real unlabeled data and the generated data to be used solely for the unsupervised part of our model.

Simultaneously, the generator's goal is to maximize the discriminator's unsupervised loss by generating samples that are close to the real distribution of the data produced by BERT noted as p_d. To do so, we use two different losses for our generator. Let us first define the generator's unsupervised loss:

$$\mathcal{L}_{G_{unsup}} = BCE(\hat{y}_{fake}, 1)$$

In other words, we train G to persuade D that the fake representations are real. Additionally, we use the feature matching (FM) loss as described in Salimans et al. (2016):

$$\mathcal{L}_{G_{FM}} = ||\mathbb{E}_{x \sim p_d} f(x) - \mathbb{E}_{x \sim G} f(x)||_2^2,$$

Cohort	Inclusion Criteria	Observation*	Prediction*	Outcome
Mortality-ICU	all w/t ICU stay & min 2 hospital visits	unbounded	days in ICU	death
Mortality-Disch	all w/t an inpatient visit	unbounded	365 days	death
Heart Failure	all from 40 to 85 yrs of age	3 years	365 days	heart failure
Lung Transplant	all w/t a recorded lung transplant	unbounded	3 years	survival

Table 1: Description of the cohort definitions process for all 4 tasks. *Observation and Prediction window.

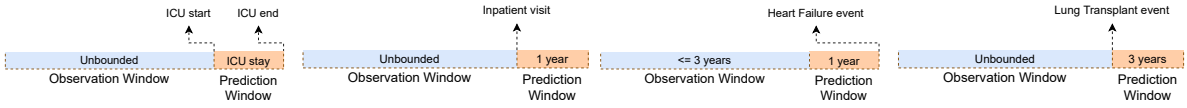
where $f(x)$ is the activation of an intermediate layer of D. This added objective forces G to generate representations that match the statistics of the ones produced by BERT instead of simply trying to maximize the discriminator loss. Lastly, the overall loss of the generator can be formulated as the sum of both of its losses $\mathcal{L}_G = \mathcal{L}_{G_{unsup}} + \mathcal{L}_{G_{FM}}$. This training process allows us to learn better representations to outperform the generator because of the information gained from out-of-cohort patients that our model would otherwise not have access to.

4. Data

We have evaluated our method on multiple medical predictive tasks with varying data scarcity: a) mortality prediction in the Intensive Care Unit (ICU) (Mortality-ICU), b) mortality prediction within one year after discharge (Mortality-Disch), c) heart failure prediction within the next year (Heart Failure), and d) survival prediction 3 years after a lung transplant (Lung Transplant). To further demonstrate the broad applicability of CEHR-GAN-BERT, we have used three different public EHR datasets with a number of in-cohort patients ranging from 224 to 29,736. Specifically, we have used a) MIMIC-IV (Johnson et al., 2021), a dataset from the Beth Israel Deaconess Medical Center containing hospitalization data as well as ICU stays to perform the Mortality-ICU task; b) Synthea (Walonoski et al., 2017), a synthetic patient population simulation platform that we have used to generate 70,000 synthetic adult and pediatric patients from every US states in respect of their population size for the Mortality-Disch task and to generate a larger cohort (120,000 patients) for the Lung Transplant task; and c) the All of Us Research Program EHR dataset (The All of Us Research Program, 2019), which is a real-world publicly-available EHR dataset of adult patients across the US for the Heart Failure task. For brevity, we present a detailed description of the cohort extractions and inclusion criteria in Appendix A. we provide a summary of the cohort definitions for all downstream tasks in Table 1 as well as a visual description of the different time windows in Figure 3 and some descriptive statistics in Table 2.

5. Results

To demonstrate the performance of our proposed method, we compare CEHR-GAN-BERT to multiple EHR prediction models, including BERT-based architectures, on four different tasks as mentioned in sections A.4 (Lung Transplant), A.1 (Mortality-ICU), A.2 (Mortality-Disch), and A.3 (All of Us Heart Failure). We also experimented with the same tasks by



(a) (b) (c) (d)

Figure 3: Division of the data between the prediction and observation window for the (a) Lung Transplant, (b) Mortality-ICU, (c) Mortality-Disch, and (d) Heart Failure tasks.

Table 2: Cohorts Statistics.

	Mortality-ICU	Mortality-Disch	All-of-Us Heart Failure	Lung Transplant
Patients # (Males / Females / ND)	21,665 (11,592 / 10,073 / 0)	10,930 (4,788 / 6,142 / 0)	29,736 (11,063 / 18,240 / 433)	224 (120 / 104)
Unlabeled data (Males / Females / ND)	54,960 (22,627 / 32,333 / 0)	39,970 (20,893 / 19,077 / 0)	26,687 (9,678 / 16,679 / 330)	851 (423 / 428)
Age at end of observation window	18-102 (69)	1-108 (41)	41-83 (60)	23-65 (49)
Number of visits per patient	2-48 (4.25)	1-119 (32.6)	1-102 (12.27)	8-92 (38.26)
Unique medical codes per patient	6-258 (76.39)	1-106 (32.7)	1-252 (32.12)	13-91 (39.52)
Positive rate	10.9	12.4	10.6	75.89

retaining varying percentages of the data to further evaluate the performance of our model on smaller data. Lastly, we investigate the impact of the Discriminator's unsupervised loss $L_{D_{unsup}}$ by varying its weight in multiple testing scenarios. All of the experiments have been realized using 5-fold cross-validation with a 70/10/20 percent on training, validation, and test split.

5.1. Prediction tasks

To evaluate the performance of our model on larger EHR datasets, we have compared our model to multiple state-of-the-art architectures for EHR prediction, namely Dipole (Ma et al., 2017), BEHRT (Li et al., 2020), and CEHR-BERT (Pang et al., 2021) as well as two non-temporal commonly used baselines: logistic regression and multi-layer perceptron (MLP). For both of the non-temporal methods, we have counted the occurrence of each medical code in the patient history resulting in an input vector of size V for every patient, where V is the number of medical codes in the vocabulary as used for the non-temporal baselines in Choi et al. (2016) and have yielded better results in our experiments that a regular one-hot encoding (10% increase on all metrics for the Lung Transplant task, for example). RareBERT (Prakash et al., 2021) could have been a great baseline, but the study does not provide the code. We provide more information and implementation details on the baselines in Appendix B.

Table 3 shows the AUROC, AUPRC, and F1 score for all the methods tested on all tasks and datasets. The Lung Transplant task has the smallest dataset and highlights the goal of this study the best. Dipole performed poorly on this task which is most likely due to a lack of data for a deep learning model to learn from, emphasizing the already proven benefits of transfer learning in BERT-based models for small datasets. Even though our method is principally tailored for few-shot learning in an EHR setting, CEHR-GAN-BERT outperforms all other baselines, including CEHR-BERT, for the Mortality-Disch task on the

Table 3: Comparison of our proposed method with 2 non-temporal baselines and 3 temporal models in multiple medical binary classification tasks. Mean ± standard deviation.

		Logistic Regression	MLP	Dipole	BEHRT	CEHR-BERT	CEHR-GAN-BERT
Mortality-ICU - MIMIC-IV	AUROC	74.61 ± 1.45	72.97 ± 1.63	92.24 ± 0.68	92.67 ± 0.19	**93.70 ± 0.35**	93.47 ± 0.47
	AUPRC	33.97 ± 2.16	28.85 ± 2.34	59.56 ± 1.20	60.54 ± 1.66	64.26 ± 2.03	**64.46 ± 1.22**
	F1	53.93 ± 1.96	48.62 ± 2.39	59.96 ± 1.77	55.76 ± 2.87	58.78 ± 2.59	**59.07 ± 1.18**
Mortality-Disch - Synthea	AUROC	91.89 ± 0.54	91.67 ± 0.69	95.42 ± 0.65	97.26 ± 0.28	97.67 ± 0.34	**97.99 ± 0.37**
	AUPRC	77.57 ± 2.27	73.95 ± 1.85	84.12 ± 1.65	85.43 ± 2.42	87.52 ± 1.74	**89.08 ± 0.41**
	F1	83.00 ± 1.33	83.05 ± 2.02	84.54 ± 1.29	84.26 ± 1.55	85.34 ± 0.42	**86.59 ± 0.57**
Heart Failure - All of Us	AUROC	66.39 ± 2.36	68.86 ± 2.69	86.57 ± 1.91	87.07 ± 1.76	87.34 ± 1.74	**87.65 ± 1.63**
	AUPRC	30.70 ± 1.79	20.22 ± 3.01	60.26 ± 2.32	61.92 ± 1.64	63.68 ± 2.82	**64.64 ± 1.67**
	F1	45.50 ± 2.91	36.75 ± 2.98	55.34 ± 1.86	54.81 ± 1.31	57.28 ± 2.10	**57.44 ± 1.10**
Lung Transplant - Synthea	AUROC	68.31 ± 7.26	64.84 ± 8.66	52.77 ± 4.85	70.47 ± 7.07	72.40 ± 7.54	**78.08 ± 5.81**
	AUPRC	34.29 ± 8.25	32.97 ± 8.32	34.94 ± 5.35	40.02 ± 9.75	44.83 ± 7.30	**49.89 ± 4.99**
	F1	43.09 ± 8.85	41.90 ± 8.41	24.91 ± 10.12	45.65 ± 5.07	46.15 ± 6.61	**54.28 ± 5.43**

Synthea dataset, especially for the imbalance-proof metrics like the AUPRC or F1 score, where we can notice an improvement of more than 1% on both of these metrics. We can observe a similar pattern on the Heart Failure task on the All of Us dataset even though CEHR-GAN-BERT achieves results closer to the second-best performing model on the F1 score. The slightly lower performance can be explained by the larger size of the dataset in comparison to the one used with Synthea, therefore reducing the importance of the unsupervised section of our architecture. Finally, as we expected due to the larger size and the lesser extent of demographic information, CEHR-BERT and CEHR-GAN-BERT achieved similar performances (with differences of less than 0.5% on all metrics) on the Mortality-ICU task on the MIMIC-IV dataset. Our architecture seems to provide better stability than the other BERT-based models with smaller standard deviations on average, which can lead to better performance guarantees in deployment. Overall, these results show that while CEHR-GAN-BERT is designed for scenarios with smaller-size datasets, it still achieves competitive results on regular datasets (our model's performance remains within the standard errors of competing methods).

5.2. Few-Shot Learning

Following the same experimental setup described before, we have randomly removed 99, 97.5, 95, 90, 75, and 50% of the training data (with the 70/10/20 split and 5-fold cross-validation process) and have kept the full 10 and 20% as validation and testing data. This process results in varying amounts of data available for the models to learn from, going as low as 80 labeled patients and 10 or fewer positive samples. As shown in Figure 4 for the Mortality-Disch task, while our architecture slightly outperforms other state-of-the-art EHR predictive models on larger amounts of data ($\leq$ 75% of missing data), CEHR-GAN-BERT's ability to learn from low amounts of annotated patients is best highlighted when only a small portion of the dataset is available ($\geq$ 90% missing data or less than 1000 patients here). Notably, our proposed method was able to outperform CEHR-BERT, the second best-performing method tested, by more than 20% on both the AUPRC and F1

scores, when only using 1% of the training data (slightly over 70 labeled patients). Similar results can be observed on the Mortality-ICU task with the MIMIC-IV dataset. Though not as drastically, CEHR-GAN-BERT outperforms CEHR-BERT with $\geq 90\%$ of missing data, especially for the F1 score which indicates a better model calibration. Notably, we can see an improvement of about 10% on all metrics when 99% of the training data is missing, where both BEHRT and CEHR-BERT achieved an F1 score very close to 0, while the proposed method achieved an F1 score of 14.4%. In this experiment, we have not included the non-temporal baselines (Logistic Regression and Multi-Layer Perceptron) and Dipole because their results throughout the experiments have turned out to be extremely volatile with standard deviations of over 15% on all reported metrics in some cases. We also excluded the Heart Failure task due to computational resource limitations on the All of Us Workbench (and the data cannot leave the platform). Still, we have observed similar patterns and results on the All of Us dataset as well. Additionally, we have not performed this experiment on the Lung Transplant task as it was already a few-shot learning task by nature.

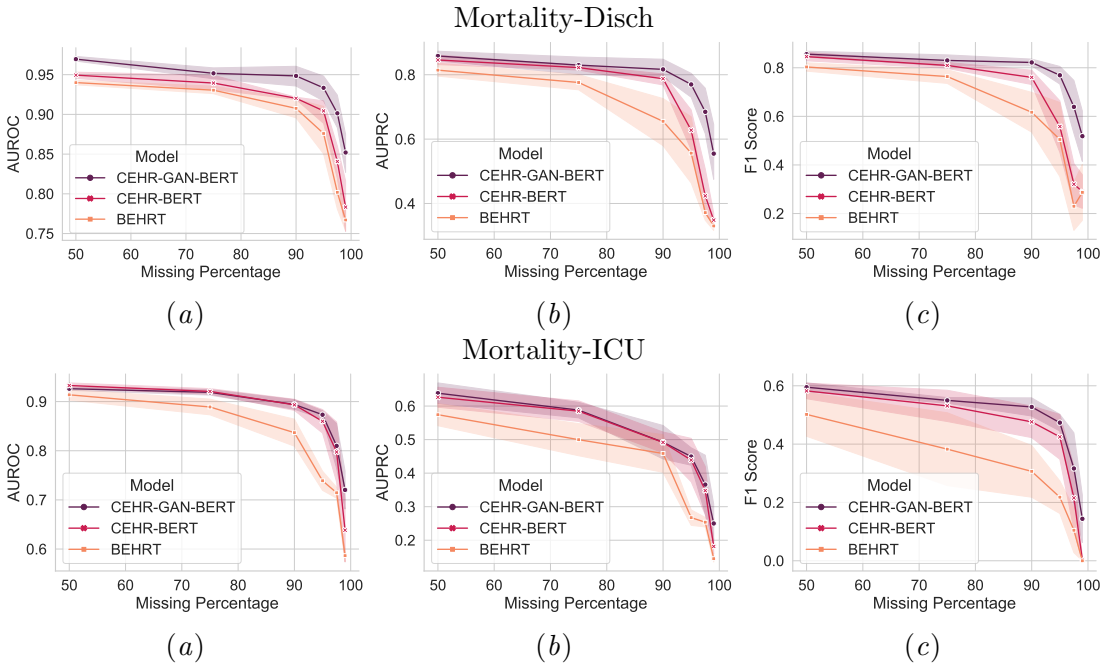


Figure 4: Comparison of (a) AUROC, (b) AUPRC, and (c) F1 score between CEHR-GAN-BERT (ours) and 2 BERT-based methods for EHR on few-shot learning for the Mortality Prediction after Discharge (Mortality-Disch) task with Synthea and the Mortality Prediction in ICU (Mortality-ICU) task with MIMIC-IV. The shaded areas correspond to the standard deviation.

5.3. Experiment on unsupervised loss weight

Throughout our experiments, we have noticed that the amount of patient data substantially impacts the discriminator and generator performances in the unsupervised tasks. When large amounts of data are available, the BERT network can learn a more detailed representation of each patient and thus help the discriminator better differentiate between the real and fake patient representations, heavily outperforming the generator. On the other hand, when only a small subset of the patients is used as input, it becomes much easier for the generator to mimic BERT's output as it is a lot less diverse. To counterbalance this phenomenon, we have introduced a weight α on $L_{D_{unsup}}$ and have also conducted experiments to evaluate CEHR-GAN-BERT's performances with different values for α (0, 0.25, 0.50, 0.75, and 1) in four scenarios: one with only 5% of the training data, one with only 10%, one with 50%, and one with the entirety of the training data. Similar to the experiments on few-shot learning, we have limited our experiments to the Mortality Prediction task of Synthea. In these scenarios, when $\alpha = 1$, no weight is applied to $L_{D_{unsup}}$ while having $\alpha = 0$ results in the regular CEHR-BERT study with the addition of the demographics and position embeddings. We present the different results of our experiment for all scenarios in Table 4. We show that the optimal value for α decreases as the size of the dataset increases with an optimal value of 0.175 when using 100% of the training data. In this scenario, a higher value for α would force D to mostly focus on the unsupervised task, not putting enough emphasis on the prediction task. On the other hand, a clear shift in α's optimal value can be noticed when the quantity of data available is drastically lower, where we have achieved the best results with a value of 0.9 and 0.95 when using only 10 and 5% of the data, respectively. Moreover, we note that adding the demographics and position embeddings to CEHR-BERT ($\alpha = 0$) increases the performances, with the greatest improvements seen when lower amounts of data were available (CEHR-BERT's results on few-shot learning can be seen in Figure 4). We also note that though in some scenarios a suboptimal α can achieve good results with a low amount of data available, it will hamper the model's stability, with higher standard deviations as we can see when only 5% of the training data was used and $\alpha \leq 0.25$.

5.4. Bias Evaluation

A concern about transformer-based methods trained on larger datasets is the degree to which training on the samples unrelated to the ultimate downstream task can lead to undesired biases toward certain individuals or groups. To study this concern, we have evaluated our model's performance across three sensitive attributes (gender, race, and ethnicity) on the Mortality-Disch task (Synthea), using the same setup as the results presented in Table 3. We report two metrics for each demographic subgroup: Equalized Opportunity and Equalized Odds (Table 5). Equalized Opportunity ensures that positive patients are equally likely to be predicted as positives regardless of demographics (true positive rate), while Equalized Odds also ensures that the false positive rates are equal.

Table 4: Experiments on the discriminator's unsupervised loss $L_{D_{unsup}}$ weight α on the Mortality-Disch Prediction task with various amount of training samples (100%, 50%, 10%, 5%, the value in parentheses is the number of patients used for training) and values for α (0, 0.25, 0.5, 0.75, 1).

		$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
100% of data (7651)	AUROC	97.68 ± 0.48	**97.76 ± 0.39**	95.96 ± 0.84	95.57 ± 1.11	95.51 ± 0.68
	AUPRC	**88.76 ± 0.6**	87.57 ± 1.49	84.52 ± 1.79	84.13 ± 2.05	84.07 ± 1.25
	F1	85.94 ± 0.97	**86.13 ± 0.63**	85.25 ± 0.97	84.99 ± 1.11	84.58 ± 1.11
50% of data (3825)	AUROC	**97.11 ± 0.37**	96.78 ± 0.47	95.81 ± 0.70	95.23 ± 0.74	94.80 ± 0.52
	AUPRC	85.19 ± 1.36	**85.31 ± 1.22**	83.87 ± 1.74	83.92 ± 0.58	81.37 ± 1.41
	F1	84.05 ± 2.01	**84.88 ± 0.61**	84.46 ± 1.36	84.53 ± 0.79	83.53 ± 1.31
10% of data (765)	AUROC	94.19 ± 0.87	**94.34 ± 1.82**	93.83 ± 2.1	93.28 ± 1.55	93.77 ± 0.66
	AUPRC	79.05 ± 1.89	79.97 ± 1.79	78.76 ± 1.16	79.30 ± 2.92	**80.80 ± 1.63**
	F1	79.65 ± 2.43	78.56 ± 1.96	79.87 ± 1.03	80.55 ± 1.9	**81.92 ± 2.02**
5% of data (383)	AUROC	92.42 ± 1.93	92.46 ± 1.33	92.47 ± 0.92	92.18 ± 1.37	**92.78 ± 0.68**
	AUPRC	72.47 ± 6.41	75.12 ± 5.42	76.26 ± 3.15	75.82 ± 2.73	**76.74 ± 2.71**
	F1	70.59 ± 6.26	71.88 ± 6.46	77.34 ± 1.43	**77.91 ± 3.01**	77.56 ± 2.86

Table 5: Bias evaluation on the Mortality-Disch task using the Synthea Dataset. Mean ± standard deviation.

		True Positive Rate	False Positive Rate
Gender	Male	89.26 ± 1.78	2.61 ± 0.34
	Female	87.90 ± 0.70	2.9 ± 0.98
Race	White	87.21 ± 1.55	2.78 ± 0.82
	Black	85.38 ± 4.95	1.81 ± 0.77
	Asian	86.65 ± 2.34	3.40 ± 0.92
	Other	91.88 ± 7.22	1.95 ± 1.84
Ethnicity	Hispanic	88.60 ± 1.55	2.75 ± 0.76
	Non-Hispanic	85.84 ± 1.25	2.46 ± 1.43

6. Discussion

While most modern machine learning applications in healthcare leverage large EHR datasets, data availability remains a common issue in the field. Electronic health records are often hard to gather and clean, which, followed by a restrictive cohort selection process can become a real challenge when applying deep learning techniques. In this study, we have proposed an architecture to address this issue. CEHR-GAN-BERT, with the addition of the adversarial learning setting, proved to be able to learn much more intricate patient representations

through the information extracted from out-of-cohort patients and is thus able to achieve good results on downstream predictive tasks with fewer data available than other baselines. We have demonstrated the importance of the utilization of both in-cohort and out-of-cohort patients in an adversarial setting to allow the model to learn good representations of the patients' medical records and enable a much more efficient learning outcome on downstream tasks with limited amounts of annotated data as highlighted by the experiments on α. The hyperparameter α also allows us to prevent the adversarial network from failing to converge by having either the discriminator or the generator outperform significantly the other and greatly stabilize the network in low data-availability scenarios.

As Table 5 shows, though we can observe some variations in the True and False Positive Rates amongst some minority populations, most of the disparities fall within a single standard deviation. Thus, both the Equalized Opportunity and Equalized Odds principles are satisfied in most cases. Nonetheless, we would like to acknowledge the complex nature of handling various types of biases and note that some subgroups do seem to be favored by the model. We expect using fairness mitigation techniques (Wan et al., 2021) could improve these results and future work would be studying and improving this aspect further.

CEHR-GAN-BERT's potential clinical applications can solve major issues in public health with its inherent capacity of learning intricate and accurate medical representations of patients in a few-shot learning setting with as low as 100 labeled patients. Applications such as rare disease prediction, prediction based on often missing measurements possibly requiring particular settings or intrusive methods, or survival prediction often rely on small datasets. Therefore, developing accurate predictive models can be pivotal in such clinical applications and we believe that CEHR-GAN-BERT opens new major opportunities in public health.

Limitations – Our study is limited in a few ways. First, our model's stability is depending on a good tuning of the unsupervised loss weight α, especially so when only a small amount of data is available. While we have confirmed a minimal impact of this on the robustness of our model, tuning the hyperparameter for each task can be costly. Second, we have not yet conducted thorough experiments on the influence of the amount of unlabeled data on the model and how it could affect the optimal value of α. Intuitively, we believe that a much larger amount of out-of-cohort patients could force the model to solely focus on the unsupervised part as many batches could be left without any labeled patient, which is why we have tried to keep a ratio between 1:1 and 10:1 between the in and out-of-cohort patients throughout our experiments. Lastly, the addition of demographic information could yield some biases towards some subpopulations, where some fairness mitigation techniques might be needed in some cases. In the future, we plan to investigate further how to achieve a competitive performance on both large and small datasets, establishing a clearer relationship between the weight of the unsupervised loss and the number of unlabeled patients. Following the recent advances in transformers and BERT-based models, we also plan on exploring different architecture possibilities for our transformer network such as Electra (Clark et al., 2020), XLNet (Yang et al., 2019b), or RoBERTa (Liu et al., 2019) to name a few. Additionally, we will investigate the benefits of using a transformer-based model for our generator.

7. Conclusion

In this study, we have presented CEHR-GAN-BERT, a semi-supervised transformer-based architecture for few-shot learning on electronic health records. Our adversarial architecture allows us to leverage knowledge from both in-cohort and out-of-cohort patients overcoming the ubiquitous challenge in applications of machine learning in healthcare of limited data availability. This method is capable of learning more detailed medical representations enabling new learning opportunities in few-shot learning scenarios such as rare disease prediction or procedure survival. We have evaluated our method on multiple tasks and datasets and achieved state-of-the-art results in few-shot learning on datasets with less than 200 patients while remaining on par on larger datasets. Additionally, we show how the model's performance can be stabilized on varying sizes of data.

Acknowledgments

The All of Us Research Program is supported by the National Institutes of Health, Office of the Director. In addition, the All of Us Research Program would not be possible without the partnership of its participants. Our study was also partially supported by the NIH awards: 3P20GM103446 and 5P20GM113125.

References

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2021.

Zhengping Che, Yu Cheng, Shuangfei Zhai, Zhaonan Sun, and Yan Liu. Boosting deep learning risk prediction with generative adversarial networks for electronic health records. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 787–792, 2017.

Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 08 2016. ISSN 1067-5027.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555, 2020. URL https://arxiv.org/abs/2003.10555.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online, July 2020. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics, Vol 1*, pages 4171–4186, 2019.

J. Fessler, Cédric Gouy-Pailler, M. Fischler, and Morgan Le guen. Machine learning in lung transplantation. *The Journal of Heart and Lung Transplantation*, 39:S385, 04 2020.

M. Gupta, H. T. Bunnell, T. T. Phan, and R. Beheshti. Concurrent Imputation and Prediction on EHR data using Bi-Directional GANs: Bi-GANs for EHR imputation and prediction. *ACM BCB*, 2021, Aug 2021.

Mehak Gupta, Raphael Poulain, Thao-Ly T. Phan, H. Timothy Bunnell, and Rahmatollah Beheshti. Flexible-window predictions on electronic health records. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12510–12516, Jun. 2022. doi: 10.1609/aaai.v36i11.21520. URL https://ojs.aaai.org/index.php/AAAI/article/view/21520.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: Transformer for electronic health records. *Scientific Reports*, 10(1):7155, 2020. ISSN 2045-2322.

Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-behrt: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1903–1911, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874.

Yiwen Meng, William Speier, Michael K. Ong, and Corey W. Arnold. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3121–3129, 2021.

Syed Asil Ali Naqvi, Karthik Tennankore, Amanda Vinson, Patrice C Roy, and Syed Sibte Raza Abidi. Predicting kidney graft survival using machine learning methods: Prediction model development and feature significance analysis study. *Journal of Medical Internet Research*, 23(8):e26843, 2021. ISSN 1438-8871.

Chao Pang, Xinzhuo Jiang, Krishna S. Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 239–260. PMLR, 04 Dec 2021.

Raphael Poulain, Mehak Gupta, Randi Foraker, and Rahmatollah Beheshti. Transformer-based multi-target regression on electronic health records for primordial prevention of cardiovascular disease. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 726–731, 2021.

P. K. S. Prakash, Srinivas Chilukuri, Nikhil Ranade, and Shankar Viswanathan. Rarebert: Transformer architecture for rare disease patient identification using administrative claims. In *AAAI*, 2021.

Sina Rashidian, Fusheng Wang, Richard Moffitt, Victor Garcia, Anurag Dutt, Wei Chang, Vishwam Pandya, Janos Hajagos, Mary Saltz, and Joel Saltz. Smooth-gan: Towards sharp and smooth synthetic ehr data generation. In Martin Michalowski and Robert Moskovitch, editors, *Artificial Intelligence in Medicine*, pages 37–48, Cham, 2020. Springer International Publishing.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1), 2021. ISSN 2398-6352.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the 30th Interna-*

tional Conference on Neural Information Processing Systems, NIPS'16, page 2234–2242, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Siao Sun, Fusheng Wang, Sina Rashidian, Tahsin Kurc, Kayley Abell-Hart, Janos Hajagos, Wei Zhu, Mary Saltz, and Joel Saltz. Generating longitudinal synthetic ehr data with recurrent autoencoders and generative adversarial networks. In El Kindi Rezig, Vijay Gadepally, Timothy Mattson, Michael Stonebraker, Tim Kraska, Fusheng Wang, Gang Luo, Jun Kong, and Alevtina Dubovitskaya, editors, *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, pages 153–165, Cham, 2021. Springer International Publishing. ISBN 978-3-030-93663-1.

The All of Us Research Program. The all of us research program. *New England Journal of Medicine*, 381(7):668–676, 2019. PMID: 31412182.

Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 08 2017. ISSN 1527-974X.

Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. Modeling techniques for machine learning fairness: A survey, 2021. URL https://arxiv.org/abs/2111.03015.

John Weldon, Tomas Ward, and Eoin Brophy. Generation of synthetic electronic health records using a federated gan, 2021.

Yinchong Yang, Zhiliang Wu, Volker Tresp, and Peter A. Fasching. Categorical ehr imputation with generative adversarial nets. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–10, 2019a.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019b. URL http://arxiv.org/abs/1906.08237.

Kezi Yu, Yunlong Wang, and Yong Cai. *Modelling Patient Sequences for Rare Disease Detection with Semi-supervised Generative Adversarial Nets*, page 141–150. Advanced Data Mining and Applications, 2020.

Appendix A. Cohort extraction

A.1. Mortality Prediction in the ICU

First, we have evaluated our model on a mortality in the ICU prediction task using the MIMIC-IV dataset. For this task, we aim to predict the mortality of patients who have been admitted to the ICU, that is, between the admission time and discharge time. Here, we have defined our anchor event (event in the medical record that we will use to define the different time windows for our data) to be an admission to the ICU. We then defined our observation window (time window that will be used as input to our model) as the period from the first

medical entry to the entry into the ICU, and our prediction window spans from the entry into the ICU to the discharge time. This process results in an unbounded (no time restriction) observation window and a prediction window (time period of the prediction) of variable time that represents the length of stay in the ICU. For our input data, we have limited our model to solely use the medical records from the hosp module (non-ICU) within the observation window, we then have extracted all the conditions, prescriptions, procedures, and gender of every patient to form our input data. Because MIMIC-IV combines the ethnicity and race information of the patients, we have removed one of the corresponding embedding layers from our model.

A.2. Mortality prediction after discharge

Additionally, we have conducted experiments on a Mortality prediction after discharge task, inspired by the experiments described in CEHR-BERT (Pang et al., 2021). This experiment, realized on the Synthea dataset, aims to predict mortality one full year after an inpatient visit where the patient was discharged (patients who have died during the visit are not included in the cohort) while taking into account the entirety of the EHR data before the inpatient visit, as opposed to CEHR-BERT where the only used the year before the visit as input data. We believe that taking into consideration the total history of each patient would be closer to the decision process of medical professionals. When patients have multiple inpatient visits, we have chosen the latest one in the history as the anchor event, which results in an unbounded observation window spanning from the beginning of the medical record of a patient to its latest in-patient visit, and a one-year prediction window representing the entire year following the patient's latest in-patient visit. Additionally, we have extracted the conditions, prescriptions, and procedures codes occurring in the observation window of each patient to serve as input for the input sequence as well as the demographic information (gender, race, and ethnicity).

A.3. Heart Failure Prediction

Similar to the task described in Choi et al. (2016), we have evaluated our method on a Heart Failure prediction task for patients between the age of 40 and 85 on the All of Us dataset. Specifically, we aim to predict the occurrence of heart failure as defined by one of the SNOMED codes shown in Table 6. In cases where a patient had multiple occurrences of a heart failure SNOMED code, we have defined the first visit of such an appearance as the anchor event for that specific patient. As opposed to the original study that used a shorted observation window, we have defined the observation window to be the time period between the fourth and the first year preceding the anchor event, and the prediction window to be the full year before the heart failure event. This results in an observation window of 3 years and a prediction window of 1 year. In other words, we try to predict heart failure within the next 365 days using at most the previous 3 years as input. For this task, we have included up to 10 control patients for every positive patient and have used the same time windows to ensure all patients have a similar medical history length. As described for the Mortality-Disch task, we have extracted the patients' demographics and all condition, prescription, and procedure codes throughout the observation window to serve as input for

our model.

Table 6: Heart Failure SNOMED codes used for the Heart Failure prediction task on the All of Us dataset as presented in Section A.3.

SNOMED Code	Description
84114007	Heart failure
42343007	Congestive heart failure
441530006	Chronic diastolic heart failure
441481004	Chronic systolic heart failure
88805009	Chronic congestive heart failure
153941000119100	Chronic combined systolic and diastolic heart failure
443253003	Acute on chronic systolic heart failure
153951000119103	Acute on chronic combined systolic and diastolic heart failure
194779001	Hypertensive heart and renal disease with (congestive) heart failure
5148006	Hypertensive heart disease with congestive heart failure
85232009	Left heart failure disorder

A.4. Survival Prediction after lung transplant

Finally, we have evaluated our model on a lung transplant survival prediction task using the Synthea dataset. As described in Naqvi et al. (2021), machine learning has proved to achieve promising results in predicting both short-term (1 year) and long-term (5 years) survival of kidney transplant patients, however, very little work has investigated survival analysis for lung transplantation. Based on Fessler et al. (2020); Naqvi et al. (2021), the goal of this task is to highlight CEHR-GAN-BERT's performances on a very reduced dataset, which was further developed in Section 5.2. For this task, we aim to predict survival within 3 years following a lung transplant using the entirety of the medical record preceding the lung transplant as input. Therefore, we define our anchor event as an occurrence of a lung transplant procedure (defined by the SNOMED code 88039007). This results in an unbounded observation window and a prediction window of 3 years. We define positive cases as patients who have not died during the three years following their lung transplant procedure, therefore, patients who have no visits after the end of the third year are not included in the cohort. Similar to the previous tasks, we use the gender, conditions, prescriptions, and procedures codes in addition to the race and ethnicity of each patient during the observations window as input features. This task is critical to demonstrate CEHR-GAN-BERT's potential impact on real-world clinical applications due to the importance of having models that are able to predict accurately lung transplant survival to better optimize the allocation of available lungs while overcoming the challenge of the small sample size inherent to the task.

Appendix B. Baselines description

For the Logistic Regression, we have used scikit-learn's implementation with L-2 normalization and Stochastic Average Gradient, and we have used our discriminator's network (2 hidden layers of size 288 with the Adam optimizer (Kingma and Ba, 2014) and a dropout

rate of 0.2) trained on the non-temporal input (without the adversarial part) for the MLP. Specifically, we have compared our results to the following temporal models for EHR:

Dipole - Dipole (Ma et al., 2017) uses bi-directional RNNs couple with three attention mechanisms to predict diabetes at the next visit. This method has been among the most used architectures for EHR predictive tasks for a few years before the applications of BERT-based models.

BEHRT - BEHRT (Li et al., 2020) is, to the best of our knowledge the first application of BERT to the EHR world, it leverages the original NLP architecture by considering each medical code as a word, the visits as sentences, and the medical record of a patient as a document as well as the pre-training and fine-tuning paradigm. This approach enabled new opportunities and has been the starting point of many BERT-based models for EHRs,

CEHR-BERT - CEHR-BERT (Pang et al., 2021) is one of the latest applications of BERT models in an EHR setting. The authors have introduced a new input representation process with the addition of the attention tokens in-between visits as well as the temporal embedding. While they have also introduced a second learning objective (visit-type prediction), in this study, we have solely used CEHR-BERT as the starting point of our architecture without the second objective. Though, it is reasonable to believe that the addition of the second learning objective will have a similar impact on our architecture than on CEHR-BERT.