Reproducing sEHR-CE: Language modelling of structured EHR data for efficient and generalizable patient cohort expansion

Jayashree Ganesan

{ganesan8}@illinois.edu

Group ID: 84

Paper ID: 22

Presentation link: https://youtu.be/UQDOobDZtHw

Code link: srijayashree/CS598: Deep Learning (github.com)

1 Introduction

EHR is a comprehensive record of patient's health information. Typical EHR contains clinical texts, medical history, diagnoses, prognosis, medication prescriptions, lab results and other relevant clinical information. Machine learning algorithms can be trained to predict patient outcomes, such as readmissions, length of stay, and mortality risk. These predictions can be used to identify high-risk patients and target interventions to improve their outcomes.

sEHR-CE (language modelling of **s**tructured **EHR** data for patient **C**ohort **E**xpansion) is a novel framework based on transformers to enable the integrated analysis of heterogeneous clinical datasets without relying on any manual curation and mapping. The model does this using textual descriptors of concepts and representing individuals' EHR as sections of text. Pre-trained language models are fine-tuned to predict disease phenotypes more accurately than non-text and single terminology approaches.

MIMIC III data with clinical text, ICD9 diagnostic code of the patient are used by sEHR-CE algorithm to identify individuals without a diagnosis who share clinical characteristics with patients.

2 Scope of reproducibility

"sEHR-CE framework based on transformers to enable integrated phenotyping and analyses of heterogeneous clinical datasets without relying on curated maps. Clinical terminologies are unified using textual descriptors of concepts and represent individuals' EHR as sections of text. Then it is fine-tuned pre-trained language models to predict disease phenotypes more accurately than non-text and single terminology approaches."

- sEHR-CE uses the pre-trained language model PubMedBERT as the encoder of the tokenized input sequences of clinical term descriptions.
- Fine tuning Masked Language Modeling task using clinical text for diabetes and heart-failure condition.
- sEHR-CE uses the fine-tuned encoder and a fully connected linear layer as the decoder.

2.1 Addressed claims from the original paper

- The paper makes 2 claims as the primary contributions.
    - The presentation of a cohort expansion method that provides phenotype predictions outperforming non-text and single terminology approaches.
    - An in-depth qualitative evaluation demonstrating that positively predicted controls share similar clinical representations with cases, providing a high degree of evidence that these may be previously undiagnosed or misdiagnosed individuals.
- I have addressed the first claim in the paper as part of the project.

3 Methodology

3.1 Model descriptions

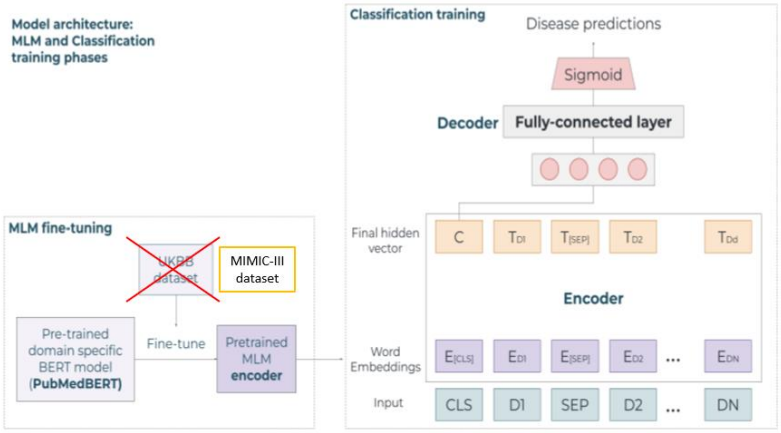
The model used is the same one as described in the paper.
The tokenized input sequence of an individual p, $X^{(p)}=(x_1^{(p)},....,x_n^{(p)})$ forms input to the encoder. The tokenized sequences are obtained by passing the text sequences from the health records through a Wordpiece Tokenizer.

sEHR-CE discussed in the paper uses pre-trained language model PubMedBERT as the encoder. I used the open source implementation of 'transformers' from huggingface as the BERT model for the project.

Since the input used differs from the general scientific text on which PubMedBERT was trained, the model was fine-tuned on the masked-language modeling (MLM) task by masking words (e.g. descriptions) at random following the original BERT paper [Devlin et al., 2019]. In the paper, the fine-tuning was done using the UK BioBank patient data. Since I did not have access to the above data and was using the MIMIC-III patient data, this was used to fine-tune the PubMedBERT model. The parameters for this task were essentially the same as used in the paper. I used 4000 patient entries, with early stopping for 5 epochs with a batch size of 32 and a learning rate of 4×10^{-5} using gradient descent with an AdamW optimizer. The output dimension of the encoder was 768.

The input tokens are first encoded using the above fine-tuned model and the hidden vector of the [CLS] token is passed to the decoder which is a fully connected linear layer. The output is passed through a sigmoid function to generate probabilities for each phenotype. In our test, we used 2 phenotypes namely heart failure and diabetes; so, we obtained 2 sets of probabilities as the model output.

The above model description is shown as a diagram from the paper, with the change for the project highlighted below.



3.2 Data descriptions

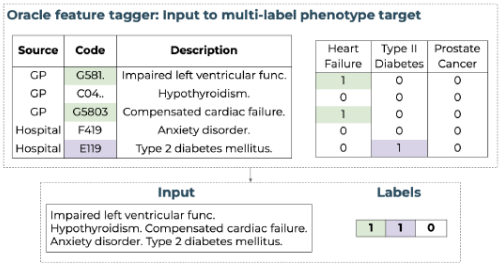
The work in the original paper is based on the UK Biobank (UKBB) [Sudlow et al., 2015] data, which is a large-scale research study of around 500k individuals between the ages of 40 and 54. It includes data, both taken at recruitment and during primary and secondary care encounters (GP and hospital).
I was not able to get access to the above data due to time constraints and hence decided to use the MIMIC-III dataset for the project. Like the original paper, the ICD9 diagnosis codes for each patient (from DIAGNOSES_ICD.csv) and the text descriptions for each of the diagnoses (from D_ICD_DIAGNOSES.csv) form the basis of the data used in the model.

In the initial phase, from the above raw data, there are 46517 patients with 58925 hospital visits having a total of 634709 diagnoses. In the original paper, only patients with greater than 5 records present in their clinical history are included to allow for sufficient information for any predictions. Following the same concept, I removed all the patients with fewer than 5 records; this reduced the final data to 39760 patients with 52074 hospital visits having a total of 612776 diagnoses.

With the above data, for each patient, I concatenate the descriptions of the ICD9 codes from the diagnoses to form the text input for each patient. The description has both short and long forms for each diagnosis; I used the short description in the project to limit the length of the input sequences. The usage of PubMedBERT restricts the length of input sequences we can use. To avoid excluding relevant clinical information by truncating the input sequences, similar to the paper, I break up long patient histories into multiple input sequences of smaller length.

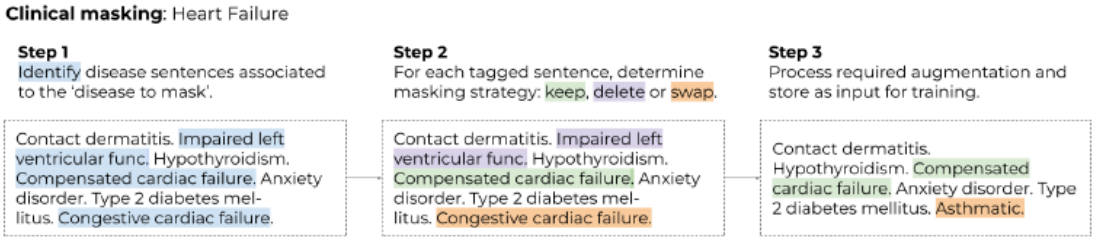
I used heart failure and diabetes as the diseases to be trained and evaluated. Based on the diagnosis for each patient and the ICD9 long title description, I set 0 or 1 labels for the patient based on the absence or presence of the disease. This forms the label vector that is used in the model for training, validation and testing.

An example input sequence after concatenation and label generation from the paper and a similar comparison from the project are shown below.



	A	B	C	D	E	F
	SUBJECT_I	HADM_ID	SEQ_NUM	SHORT_CONCAT	HF	Diabetes
	3	145834	0	Prim cardiomyopathy NEC, Subendo infarct, initial, CHF NOS, Acute kidney failure NOS, Septicemia NOS, Protein-cal malnutr NOS, Car	1	0
	4	185777	0	Cirrhosis of liver NOS, Inf mcrg rstn pncllins, Human immuno virus dis, Pneumocystosis, Cachexia, Alkalosis, Bacteremia, Mth sus Stph	0	0
	6	107064	0	Hyperpotassemia, Anemia NOS, History of tobacco use, Hyp kid NOS w cr kid V, Surg comp-peri vasc syst, Dis phosphorus metabol	0	0
	8	159514	0	Single lb in-hosp w cs, Need prphyl vc vrl hepat, NB obsrv suspct infect, Routine circumcision, Fetal/neonatal jaund NOS, NB transitor	0	0
	9	150750	0	CHF NOS, Hypertension NOS, Acute kidney failure NOS, Intracerebral hemorrhage, Food/vomit pneumonitis	1	0
	10	184167	0	NB obsrv suspct infect, Single lb in-hosp w/o cs, Neonat jaund preterm del, Preterm NEC 1250-1499g, 29-30 comp wks gestation	0	0
	12	112213	0	Hypertension NOS, Surg compl-heart, Hemorrhage complic proc, Cardiac arrest, Mal neo pancreas head, Cholelith w cholecys NEC, P	0	0
	13	143045	0	Crnry athrscl natve vssl, Intermed coronary synd, Pure hypercholesterolem, Hypertension NOS, DMII wo cmp nt st uncntr	0	1

Data augmentation (clinical masking)
I followed the same masking strategy as used in the paper, which was also used in Devlin et al. [2019], Wei and Zou [2019]. For a particular disease being tested, the corresponding descriptions in the data are removed with 80% probability, retained with 10% probability and replaced with a different word with 10% probability. An example of the masking described as shown in the paper is displayed below. The masking operation is done on the training and validation datasets and the terms are removed completely in the final test data set.



To see the effect of the masking operation, I applied the masks to heart failure cases, while it is not applied to the diabetes ones.

3.3 Hyperparameters

I used the same hyperparameters as used in the paper. The number of epochs was 3, the learning rate was 10^{-5}, and a warm-up proportion of 0.25.

3.4 Implementation

The paper is from a private company named BenevolentAI and as such, their code is not available in public. The code for the project is written based on the descriptions in the paper.
Code Repo https://github.com/srijayashree/CS598

3.5 Computational requirements

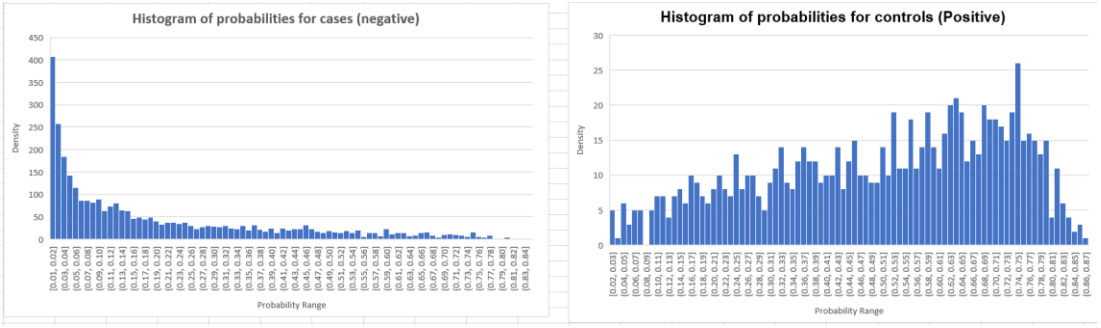
I was able to do the initial data preparation and augmentation using my personal laptop itself. However, for the actual model training and evaluation, my laptop was not sufficient, and I had to move to using Google Colab.
Even with using Colab, I was having issues with insufficient computation resources (25GB RAM and 225GB disk) with the amount of data I was using. Finally, I had to reduce the data for training, validation and evaluation to about 4000 entries each from 80000 to be able to complete the project.

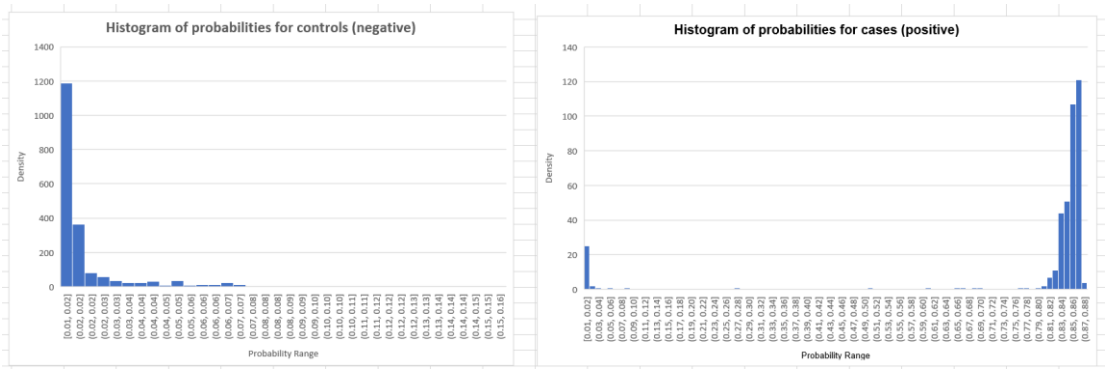
4 Results

I was able to get results that support the claims made in section 2.1.
The performance of the model is characterized in the paper using recall and AUC scores on the test set. The table below shows the scores obtained in the paper and in the project.

	Heart failure		Diabetes	
	Recall	AUC	Recall	AUC
Original paper	0.85	0.69	0.74	0.74
Project	0.55	0.85	0.0	0.53

The phenotype prediction performance is shown by plotting the histogram of the predicted probabilities for each of the diseases in both the control (negative) and case (positive) patients in the test set. In the plots, the x-axis is the probability as predicted by the model with values ranging from the min to max predicted probability. The y-axis is the number of patients having the corresponding probability. Heart failure is shown first and then diabetes.





From the above plots, we can see that the model is able to show higher probabilities for the cases (patients with positive labels) in both the masked heart failure and unmasked diabetes scenarios. While the probabilities for the diabetes scenario are clearly demarcated between high and low values (since the training data was not masked), they are more uniformly distributed in the heart failure scenario with a clear skew towards higher probabilities.

Similarly, for the control patients (ones with negative labels), the probabilities are distributed as expected with a clear skew towards low probabilities, thus showing it is doing the correct phenotyping.

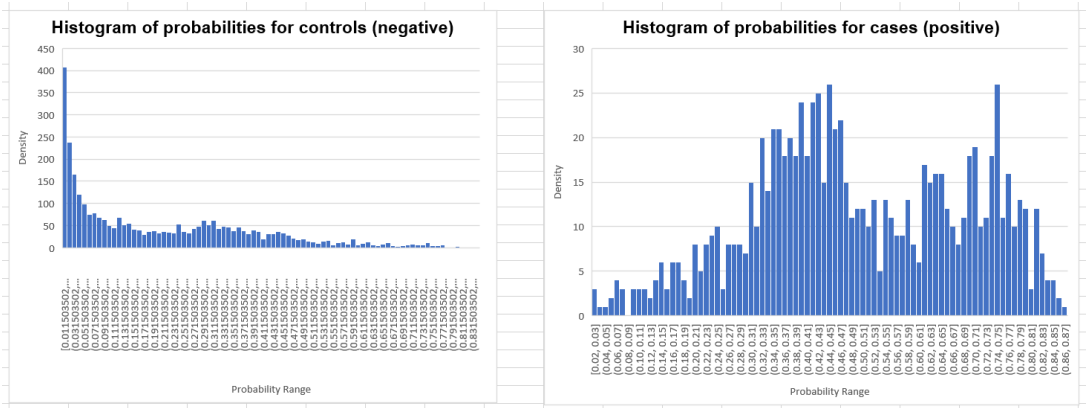
In addition, a few of the heart failure control patients (negative patients) do have a high probability from the model which might indicate they are at risk of having similar outcomes. This is the 2nd claim that was made in the paper.

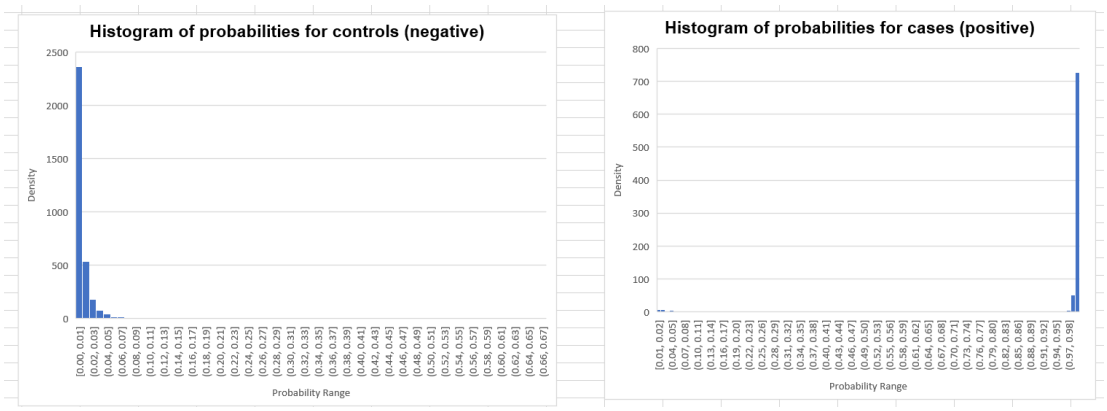
4.1 Additional results not present in the original paper

In addition to the tests above, where the disease related terms are completely removed in the test set, I also ran the same experiments on test sets where the disease terms have been masked with the same probabilities as the training set. The results from these tests are shown below.

	Heart failure		Diabetes	
	Recall	AUC	Recall	AUC
Test with same mask as training	0.29	0.76	0.94	0.98

The histogram of the probabilities for these test cases are shown below; first for heart disease and next for diabetes.





While for heart failure, the plots look similar to the earlier ones; for diabetes, the probabilities are more pronounced towards 0 and 1. This makes sense since the terms are not masked in the training data set nor are they removed in the test data set.

5 Discussion

While the histogram plots support the claims from the paper qualitatively, the recall scores are not as good as reported in the paper. Listed below are some reasons why this could be occurring.

1. I did not have access to the UK BioBank data on which the initial research was done. In place of the UKBB data, I used the MIMIC-III data which might not have the same characteristics as the UKBB data. In addition, the paper was relying on the fact that the UKBB data has both GP (primary physician) and hospital data to reduce bias towards acute events that usually present in hospitals. I'm not sure about the distribution of patient data in the MIMIC database.
2. Due to computational resources, I was not able to use all the data available for training purposes. During initial trial and error, as I increased the amount of data used for training, I was observing an improvement in the recall and AUC scores leading me to believe that the model performance can get better with larger training data.

5.1 What was difficult

The non-availability of both the code and data used in the original research made the reproduction of results difficult.

5.2 Recommendations for reproducibility

As mentioned in the beginning of this section, reproducibility could be improved by using the original UKBB data used for the research. Also, having higher computational resources allowing for using larger training data sets would help reproducibility.

6 Communication with original authors

I did not communicate with the original authors.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics, 2019.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank: An open accessnresource for identifying the causes of a wide range of complex diseases of middle and old age. PLOS Medicine, 12(3):1–10, 2015.

Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6382–6388. Association for Computational Linguistics, 2019.

Kyunghoon Hur, Jiyoung Lee, Jungwoo Oh, Wesley Price, Younghak Kim, and Edward Choi. Unifying heterogeneous electronic health records systems via text-based code embedding. In Proceedings of the Conference on Health, Inference, and Learning, volume 174 of PMLR, pages 183–203, 2022.

https://github.com/huggingface/transformers