How Long Is Enough? Exploring the Optimal Intervals of Long-Range Clinical Note Language Modeling

Samuel Cahyawijaya[1]*, **Bryan Wilie**[1*], **Holy Lovenia**[1*], **MingQian Zhong**[2,3,4], **Huan Zhong**[2,3], **Nancy Y. Ip**[2,3,4], **Pascale Fung**[1]

[1]Center for Artificial Intelligence Research (CAiRE), Department of Electronic and Computer Engineering,
The Hong Kong University of Science and Technology, Hong Kong, China
{scahyawijaya, bwilie, hlovenia, pascale}@ust.hk

[2]Division of Life Science, State Key Laboratory of Molecular Neuroscience, Molecular Neuroscience Center,
The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China
{mzhongac,dorothyzhong,boip}@ust.hk

[3]Hong Kong Center for Neurodegenerative Diseases, Hong Kong Science Park, Hong Kong, China
{mzhongac,dorothyzhong,boip}@ust.hk

Abstract

Large pre-trained language models (LMs) have been widely adopted in biomedical and clinical domains, introducing many powerful LMs such as bio-lm and BioELECTRA. However, the applicability of these methods to real clinical use cases is hindered, due to the limitation of pre-trained LMs in processing long textual data with thousands of words, which is a common length for a clinical note. In this work, we explore long-range adaptation from such LMs with Longformer, allowing the LMs to capture longer clinical notes context. We conduct experiments on three n2c2 challenges datasets and a longitudinal clinical dataset from Hong Kong Hospital Authority electronic health record (EHR) system to show the effectiveness and generalizability of this concept, achieving 10% F1-score improvement. Based on our experiments, we conclude that capturing a longer clinical note interval is beneficial to the model performance, but there are different cut-off intervals to achieve the optimal performance for different target variables. Our code is available at https://github.com/HLTCHKUST/long-biomedical-model.

1 Introduction

Clinical note is one of the most abundant data available in EHR systems, which records most of the patient interaction with the hospital services, such as consultation with doctors, procedure note, laboratory report, discharge summary, etc.[1] Despite retaining rich clinical information, clinical notes are highly unstructured and composed of non-standardized information, which curbs the potential practicality of such information. Large pre-trained LMs, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2020), GPT-2 (Radford et al., 2019), etc., have been shown to work well in extracting crucial information from clinical notes by utilizing transfer learning and attention mechanism (Ji et al., 2021; Alsentzer et al., 2019; Lewis et al., 2020). The adaptation of these models to biomedical and clinical domain emphasizes this success, establishing many new state-of-the-art performances on multiple biomedical and clinical benchmarks (Peng et al., 2019; Gu et al., 2021; Zhang et al., 2022).

While the attention mechanism embedded in the pre-trained models enables them to achieve great performance, it is to be noted that it also causes a quadratic growth in computation cost with respect to input sequence length (Tay et al., 2022; Wang et al., 2020; Cahyawijaya et al., 2022). This makes efficiently processing long documents with pre-trained LMs difficult, especially in clinical note modeling, in which a single clinical note tends to consist of hundreds or even thousands of words (Uzuner et al., 2008; Uzuner, 2009; Stubbs et al., 2015; Gehrmann et al., 2018; Johnson et al., 2019; Stubbs et al., 2019). Current approaches to this problem commonly involve truncation, chunking, or windowing of the long input sequence, preventing the models from acquiring an entire medical record information provided by a whole clinical note. Considering that clinical note modeling requires capturing and understanding the underlying long-term dependencies in the clinical notes, this certainly puts a limit on their predictive capability.

[*] These authors contributed equally.

[1] https://www.healthit.gov/isa/uscdi-data-class/clinical-notes

Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI), pages 160 - 172
December 7, 2022 ©2022 Association for Computational Linguistics

For this reason, to maximize the models' capability without sacrificing a part of the input clinical notes, we explore the application of long-range adaptation through linear attention mechanism (Dai et al., 2019; Beltagy et al., 2020; Wang et al., 2020; Choromanski et al., 2021), which reduces the computation cost of attention from quadratic to linear in regards to input sequence length.

In this work, we focus on assessing the benefit of capturing longer clinical notes on large pre-trained LMs to n2c2 (National Clinical NLP Challenges)[2] clinical tasks by adapting a linear attention mechanism, i.e., Longformer (Beltagy et al., 2020). Furthermore, to test the generality of this approach, we evaluate it on a longitudinal clinical note corpus from Hong Kong Hospital Authority EHR system, which covers records from 43 hospitals in Hong Kong. Lastly, we hypothesize that modeling longer interval of clinical notes improves the prediction quality of the models on any clinical task. To prove our hypothesis, we conduct our experiment using different context-length, allowing the model to access various intervals of clinical notes. Our result suggests that a longer interval of clinical notes increases the prediction quality of the models in most cases, but there is a limit of context length required depending on the target variable.

Our contributions in this work can be summarized in three-fold:

- We assess the effectiveness of capturing longer interval of clinical notes on biomedical and clinical large pre-trained LMs on three n2c2 challenges which increase the performance by ∼10% F1-score,
- We evaluate the generalization of this approach using longitudinal clinical note data gathered in Hong Kong Hospital Authority EHR system on two clinical tasks, i.e., disease risk and mortality risk predictions, which improve the performance by ∼5-10 F1-score,
- We observe that each target variable has a different optimal clinical notes cut-off interval and we conclude that the optimal cut-off interval for mortality risk prediction is ∼2-3 months, while for disease risk prediction, it requires 3.5 years or even longer interval to achieve the optimal performance.

2 Related Works

Clinical Note Modeling Clinical notes have been utilized for various applications in healthcare. Text mining methods for analyzing pharmacovigilance signals using clinical notes have been explored and yield promising results (Haerian et al., 2012; LePendu et al., 2012, 2013). Clinical notes with other EHR data are also employed for estimating the readmission time and mortality risk of the next patient encounter (Hammoudeh et al., 2018; Rajkomar et al., 2018). Clinical note data is also effective for analyzing disease comorbidity, such as mental illness (Wu et al., 2013), autoimmune diseases (Escudié et al., 2017), and obesity (Pantalone et al., 2017). Predicting disease risk using clinical note data has also been explored (Miotto et al., 2016; Choi et al., 2018; Liu et al., 2019, 2018; Koleck et al., 2019). Despite all the efforts in clinical note modeling, to the best of our knowledge, how clinical note interval impacts the performance of pre-trained LMs has never been studied.

Biomedical and Clinical Pre-trained LMs Self-supervised pre-training LMs employing transformer-based architectures (Vaswani et al., 2017), such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2020), and ELECTRA (Clark et al., 2019), have thrived in various general domain NLP benchmarks (Wang et al., 2018; Rajpurkar et al., 2016; Ladhak et al., 2020; Lai et al., 2017; Wilie et al., 2020; Cahyawijaya et al., 2021; Park et al., 2021). To extend the understanding of these LMs to the linguistic properties in biomedical and clinical domain, a generation of LMs exploiting biomedical and clinical corpora emerges.

In 2019, Alsentzer et al. (2019) introduce BioBERT, an extended version of BERT pre-trained on large-scale biomedical data (i.e., PubMed abstracts and PMC full-text articles) which surpasses off-the-shelf BERT in three fundamental downstream tasks in biomedical domain. Due to the linguistic differences exhibited by non-clinical biomedical texts and clinical texts, Alsentzer et al. (2019) introduce ClinicalBERT by fine-tuning BERT and BioBERT on the MIMIC-III corpus, and improve the performance over five clinical NLP tasks.

Unlike prior works, PubMedBERT (Gu et al., 2020) performs biomedical pre-training from scratch, which offers larger performance gains over various biomedical downstream tasks in the

[2] https://n2c2.dbmi.hms.harvard.edu/

BLURB benchmark. Similarly, bio-lm (Lewis et al., 2020) employs recent pre-training advances, utilizes various biomedical and clinical corpora for pre-training, and achieves the highest performance on 9 biomedical and clinical NLP tasks. In 2021, BioELECTRA (Kanakarajan et al., 2021), a general domain ELECTRA (Clark et al., 2019) pre-trained on biomedical corpora, sets the new state-of-the-art performance for all datasets in the BLURB benchmark and 4 datasets in the BLUE benchmark (Peng et al., 2019).

Long Sequence Language Modeling Recent progress in language modeling is dominated by transformer-based models which shows a remarkable results on numerous tasks. Nevertheless, these models have limited capability to process long-range clinical notes data due to its quadratic attention complexity. Various approaches have been introduced to reduce this complexity problem, such as recurrence approach (Dai et al., 2019; Rae et al., 2020), sparse and local attention patterns (Kitaev et al., 2020; Qiu et al., 2020; Child et al., 2019; Zaheer et al., 2020; Beltagy et al., 2020), low-rank approximation (Wang et al., 2020; Winata et al., 2020), and kernel methods (Katharopoulos et al., 2020; Choromanski et al., 2021). Adaptation from existing pre-trained models to some of these methods have also been explored and show the potential for knowledge transfer (Beltagy et al., 2020; Choromanski et al., 2021). In this work, we utilize Longformer (Beltagy et al., 2020) to enable the model to capture long-range clinical note information.

3 Methodology

3.1 Problem Definition

Clinical notes are narrative patient data relevant to the context identified by note types[3]. There are multiple types of clinical notes, e.g., discharge summary, consultation note, progress note, lab report, etc. In general, a single clinical note consists of a text narrative and additional metadata defining the clinical note, e.g., note identifier, recording timestamp of the note, etc. In n2c2 challenges, a single clinical note is presented in a textual format with the metadata written on top of the text narrative, while a longitudinal clinical note is presented as a concatenation of several clinical notes with a separator text placed between two clinical notes. This clinical note is usually long, ranging from several hundreds to thousands words, while most existing biomedical and clinical pre-trained LMs can only capture up to 512 subwords, which is insufficient to capture the whole content of most clinical notes.

3.2 Long-Range Clinical Note LMs

We increase the capacity of LMs to process longer clinical notes by adapting Longformer (Beltagy et al., 2020) to the existing biomedical and clinical pre-trained LMs. Longformer enables linear attention mechanism by dividing single quadratic all-to-all attention into two attention steps, i.e., sliding-window and global attentions. Sliding-window attention allows each token to attend to neighboring tokens, while global attention allows some, usually a few, tokens to attend to all tokens, hence has a better computation complexity compared to the quadratic attention mechanism. It is to be noted that when extending an original transformer-based model into a Longformer, some new parameters are introduced, i.e., the new positional embeddings, the sliding-window projection parameters, and global attention projection parameters. For the positional embeddings, following (Beltagy et al., 2020), we copy the weights of the pre-trained positional embeddings to initialize the new positional embeddings. For the sliding-window and global attention parameters, we initialize both projection parameters with the pre-trained projection parameters.

4 Long-Range Clinical Note LMs on n2c2 Challenges

We assess the effectiveness of long-range clinical note LMs on US-based clinical note datasets from three n2c2 challenges. Additionally, we also evaluate six different pre-trained LMs without long-range adaptation to benchmark the performance of the biomedical and clinical LMs.

4.1 Dataset

We use three clinical datasets concentrating on classifying diverse clinical problems from n2c2. These datasets are: 1) n2c2 2006 smoking challenge, focusing on predicting smoking status of patients based on their discharge summary; 2) n2c2 2008 obesity challenge, focusing on recognizing obesity and its comorbidities of patients through their discharge summary; and 3) n2c2 2018 cohort selection challenge, focusing on determining if a patient meets selection criteria of certain clinical trials co-

[3] https://www.healthit.gov/isa/uscdi-data-class/clinical-notes

Dataset	|Train|	|Test|	Word count				Longitudinal?	#Label	#Class
			Median	Q3	95%	Max			
2006 Smoking	398	104	677	1096	1775	3023	No	5	1
2008 Obesity (Textual)	730	507	1084	1425	2094	4280	No	16	4
2008 Obesity (Intuitive)	730	507	1084	1425	2094	4280	No	16	4
2018 Cohort Selection	202	86	2550	3235	4578	7070	Yes	13	1

Table 1: The overall statistics of the n2c2 datasets used in our experiment.

horts through longitudinal clinical notes. We utilize BigBIO framework (Fries et al., 2022)[4] to load the n2c2 datasets. We provide overview of these datasets in Table 1.

n2c2 2006 Smoking Challenge We utilize the smoking prediction subtask from n2c2 2006 challenge (Uzuner et al., 2008), where each data instance consists of a de-identified discharge summary annotated by two pulmonologists with smoking status. This smoking status can be either past smoker" (when it is explicitly stated that the patient is a past smoker or that the patient used to smoke but has stopped for at least a year), "current smoker" (when it is explicitly stated that the patient is a current smoker or that the patient has smoked within the past year), "smoker" (when there is not enough temporal information to classify whether a patient is a "past smoker" or "current smoker"), "non-smoker" (when a patient's discharge summary indicates an absence of smoking habit), or "unknown" (when there is no mention of smoking).

n2c2 2008 Obesity Challenge The n2c2 2008 obesity challenge (Uzuner, 2009) consists of 1027 pairs of de-identified discharge summaries and 16 disease labels. The disease labels include obesity and its 15 comorbidities, e.g., asthma, atherosclerotic cardiovascular disease (CAD), congestive heart failure (CHF), depression, diabetes mellitus (DM), gallstones/cholecystectomy, gastroesophageal reflux disease (GERD), gout, hypercholesterolemia, hypertension (HTN), hypertriglyceridemia, obstructive sleep apnea (OSA), osteoarthritis (OA), peripheral vascular disease (PVD), and venous insufficiency.

The annotation for each discharge summary is done by providing each disease label with either "present", "absent", "questionable", or "unmentioned". The dataset has two types of annotations, i.e., textual judgement (only based on related explicit statements) and intuitive judgement (based on everything written in the discharge summaries). We use both annotations in our experiments and report the evaluation scores for each annotation.

n2c2 2018 Cohort Selection Challenge The 2018 Shared Task 1: Clinical Trial Cohort Selection (Stubbs et al., 2019) reuses 288 patient records from the 2014 n2c2 shared task dataset (Stubbs et al., 2015) and reframes it as a cohort selection task, which requires an automatic evaluation of whether a patient fits or does not fit in certain cohorts according to their longitudinal de-identified clinical notes, ranging between 2-5 clinical notes.

The cohorts or selection criteria used in the dataset as labels are: DRUG-ABUSE (current or past usage of drugs), ALCOHOL-ABUSE (current alcohol intake over weekly recommended limit), ENGLISH (English-speaking patient), MAKES-DECISIONS (patients required to make their own medical decisions), ABDOMINAL (history of related surgery), MAJOR-DIABETES (major diabetes-related complication), ADVANCED-CAD (advanced cardiovascular disease), MI-6MOS (myocardial infarction in the past 6 months), KETO-1YR (diagnosis of ketoacidosis in the past year), DIETSUPP-2MOS (dietary supplement intake in the past 2 months, excluding vitamin D), ASP-FOR-MI (usage of aspirin to prevent MI), HBA1C (any hemoglobin A1c value between 6.5% and 9.5%), and CREATININE (serum creatinine above the upper limit of normal). Two annotators with medical expertise classify each label of a patient's set of clinical notes as either "met" or "not met".

4.2 Models

In this experiment, we compare several pre-trained LMs, covering two variants of BERT model representing general domain LMs, i.e., uncased BERT[5] and cased BERT[6], two variants of biomedical do-

[4] https://github.com/bigscience-workshop/biomedical

[5] https://huggingface.co/bert-base-uncased

[6] https://huggingface.co/bert-base-cased

	2006 Smoking		**2008 Obesity (Text.)**		**2008 Obesity (Intui.)**		**2018 Cohort Selection**	
	micro-f1	**macro-f1**	**micro-f1**	**macro-f1**	**micro-f1**	**macro-f1**	**micro-f1**	**macro-f1**
				Baseline				
Top-5 scorer	88.00%	69.00%	97.04%	77.18%	95.58%	63.44%	90.30%	
Top-10 scorer	86.00%	58.00%	96.39%	61.40%	95.08%	62.87%	87.70%	
				Pre-trained Language Model				
BERT-cased	61.63%	31.79%	82.47%	38.73%	81.69%	51.71%	72.80%	48.45%
BERT-uncased	65.63%	41.12%	85.73%	40.83%	83.46%	53.28%	74.86%	51.32%
clinicalBERT	56.59%	39.34%	85.64%	40.64%	85.20%	54.88%	72.83%	49.99%
PubMedBERT	69.38%	41.65%	**88.98%**	46.27%	**87.11%**	**56.47%**	74.78%	49.94%
bio-lm	**71.44%**	**49.43%**	86.57%	43.15%	84.92%	54.73%	**75.03%**	**52.18%**
BioELECTRA	70.72%	48.26%	86.71%	**48.26%**	85.31%	55.00%	74.32%	49.10%
				Long-range Pre-trained Language Model				
bio-lm (1024)	82.12%	55.72%	92.52%	50.36%	90.36%	59.13%	77.03%	53.94%
bio-lm (2048)	**86.01%**	62.30%	96.44%	55.99%	94.76%	62.61%	76.76%	52.93%
bio-lm (4096)	84.52%	57.76%	**97.11%**	55.68%	**95.48%**	63.19%	79.42%	57.85%
bio-lm (8192)	84.66%	59.49%	97.07%	55.08%	**95.48%**	**63.20%**	81.43%	**61.95%**
BioELECTRA (1024)	82.98%	63.35%	93.54%	54.47%	90.40%	59.12%	74.95%	51.69%
BioELECTRA (2048)	82.84%	61.09%	96.03%	56.08%	91.69%	60.21%	77.59%	54.39%
BioELECTRA (4096)	80.40%	57.22%	95.81%	56.06%	92.88%	61.12%	79.10%	56.38%
BioELECTRA (8192)	85.21%	**64.32%**	96.20%	**59.59%**	92.78%	61.09%	**81.63%**	58.44%

Table 2: Evaluation results of our experiments on the n2c2 datasets. Top-5 and Top-10 scorers are retrieved from the submission benchmark of corresponding challenge. The number inside the bracket denotes the length of context that can be captured by the model. **Bold** and underline denotes the first and second best scores within a group.

main LMs, i.e. PubMedBERT (Gu et al., 2021)[7] and BioELECTRA (Kanakarajan et al., 2021)[8], one variant of clinical domain LM, i.e., ClinicalBERT (Alsentzer et al., 2019)[9], and one variant of mixed biomedical and clinical domains LM, i.e., bio-lm (Lewis et al., 2020)[10].

To enable longer context clinical note modeling, we adapt Longformer (Beltagy et al., 2020) with the initialization strategy specified in §3.2. We conduct experiments with four different context lengths, i.e., $\{1024, 2048, 4096, 8192\}$ on two pre-trained LMs variants, i.e., BioELECTRA and bio-lm.

4.3 Training and Evaluation

Following BERT, RoBERTa, and bio-lm experiments, we tune the learning rate for all BERT and RoBERTa models from [1e-5, 2e-5, 3e-5]. While for the BioELECTRA model, following ELECTRA (Clark et al., 2019) and BioELECTRA (Kanakarajan et al., 2021), we tune the learning rate from [5e-5, 1e-4, 2e-4]. In all experiments, we use a batch size of 8, and a linear learning rate decay. For the n2c2 2006 and n2c2 2008 tasks, we train the models for 50 epochs, while for the n2c2 2018 task, we train the models for 80 epochs. For the evaluation, we incorporate the official evaluation metrics defined for each challenge. All of them report micro-F1 and macro-F1 scores.

4.4 Result and Analysis

As shown in Table 2, in general, domain-specific LMs yield higher performance compared to general domain LMs, except for ClinicalBERT which performs on a par with the general domain BERT models. PubMedBERT, bio-lm, and BioELECTRA produce comparable evaluation performances across all tasks, with ∼2-5% higher F1-score compared to the general domain BERT and ClinicalBERT. Nevertheless, the scores are much lower compared to the Top-10 scorer on the challenge benchmark since the models can only capture partial information of the clinical note data.

By increasing the context length of the model, the performance rises significantly. Comparing with the original pre-trained versions of the models, the best performing long-range pre-trained LM improves the evaluation performance by ∼10% F1-score in all datasets. As shown in Figure 1, models with longer context length tend to perform better, but the performance gain is limited to the length of

[7] https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract

[8] https://huggingface.co/kamalkraj/bioelectra-base-discriminator-pubmed

[9] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

[10] https://huggingface.co/EMBO/bio-lm

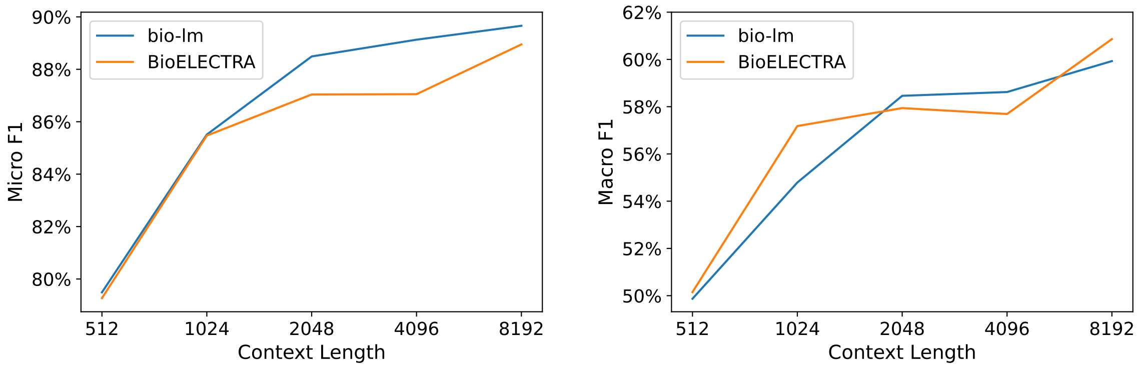

Figure 1: Effect of capturing longer clinical notes context to the evaluation performance, i.e., on micro-F1 **(Left)** and macro-F1 **(Right)**, averaged over the context length across the evaluated n2c2 tasks.

the clinical notes in the dataset. For instance, on the n2c2 2006 dataset, the performance improvement of both bio-lm and BioELECTRA models are steeper from context length 512 to 1024 rather than from context length 1024 to 2048, 2048 to 4096, and 4096 to 8192. This is because a huge portion of the notes in the datasets can be sufficiently captured within 1024 subwords. In contrast, the performance improvement on the n2c2 2018 dataset is more linear per context length step since most of the length of the clinical notes is much longer than the other two datasets. Every step of extending the context length provides more information to the model, which is likely to improve the model performance considerably.

On the n2c2 2006 and 2008 challenges, our best performing models mostly achieve a comparable score to the Top-10 or Top-5 scorer of the corresponding challenge benchmark. This is a remarkable feat since our models neither utilize any ensemble method, incorporate any clinical expert, nor exploit external data–common practices used by the top scorers in the challenge benchmarks.

5 Long-Range Clinical Note LMs on Hong Kong Longitudinal Dataset

We assess the generalization and effectiveness of long-range clinical notes LMs on Hong Kong longitudinal clinical note data. We construct a longitudinal dataset with two target variables, i.e., disease risk and mortality risk, and evaluate long-range LMs on the dataset. In addition, we add a baseline model, which takes high-level features extracted from the corresponding tabular data provided by the EHR system as the input, to assess the effectiveness of clinical note modeling.

Split	# Patients	# Seen patient records	# Unseen patient records
Train	278,253	2,027,561	-
Valid	3,621	3,177	-
Test	17,903	15,541	2,362
Total	299,777	2,046,279	2,362

Table 3: The overall statistics of our Hong Kong longitudinal dataset. **# Seen patient records** and **# Unseen patient records** indicate the number of records on the *seen* and *unseen* test set respectively.

5.1 Dataset Construction

We construct a longitudinal clinical note dataset for disease risk and mortality risk predictions from anonymized cancer cohort patient records gathered in the Hong Kong Hospital Authority EHR system covering 43 hospitals in Hong Kong. The patient records span across the year 2000 and 2018. We exclude all patients having less than two clinical notes and gather a total of ∼300,000 patients. To construct labelled data for the supervised learning, from patient P_i with T clinical records, we build $T-1$ labelled autoregressive data $\mathcal{D}^{Pi} = \{\{C_k^{P_i}\}_{k=1}^{t}, Y_{t+1}^{P_i}\}_{t=1}^{T-1}$, where $\{C_k^{P_i}\}_{k=1}^{t}$ denotes t prior clinical notes of the patient P_i, and $Y_{t+1}^{P_i}$ denotes the target criterion retrieved from the $t+1^{th}$ clinical record of the patient P_i. We collect over ∼2M labelled clinical notes from all patients with two targets: disease risk and mortality risk.

We take the last two health records from all patient records in the year 2018 as the validation and test sets. To assess the generalization to new patient data, we omit some patient data from the training set and only used the last labelled record of those patients as the *unseen* test set. The remaining test data becomes the *seen* test set. The dataset statistics

Test set	Models	Diagnosis				Mortality	
		Top-1	Top-3	Top-5	F1	F1	AUC
Seen	**EHR-FFN**	64.3%	75.7%	80.3%	40.6%	49.5%	78.1%
	BioELECTRA (512)	76.2%	88.6%	91.8%	51.6%	61.5%	**92.0%**
	BioELECTRA (2048)	79.8%	91.5%	94.3%	54.2%	**65.3%**	91.9%
	BioELECTRA (8192)	**81.3%**	**92.9%**	**95.5%**	**55.7%**	64.9%	91.8%
Unseen	**EHR-FFN**	17.8%	32.9%	43.1%	9.5%	49.6%	73.9%
	BioELECTRA (512)	63.4%	78.6%	83.7%	43.1%	52.2%	84.8%
	BioELECTRA (2048)	66.3%	81.2%	85.9%	45.1%	52.0%	85.8%
	BioELECTRA (8192)	**69.1%**	**84.0%**	**88.2%**	**46.8%**	**52.3%**	**88.1%**

Table 4: Evaluation results of our experiments on the *seen* patient test set and the *unseen* patient test set. **Bold** and underline denotes the first and the second best score on each test set, respectively.

is shown in Table 3. For the disease risk estimation, we take the final disease diagnosis on the next clinical record as the label. For cancer diseases, we group the diagnosis based on the cancer site categorization from the Hong Kong Cancer Registry[11], while for other diseases, we take the first three digits of the ICD-10 codes. In total, there are 79 classes for disease risk estimation. For the mortality label, we retrieve the mortality status from the discharge code from the next clinical record of the corresponding patient. The label distribution of the dataset is shown in Appendix A.

5.2 Models

We experiment using Longformer with three variants of sequence length, i.e., $\{512, 2048, 8192\}$. We initialized all models with the same pre-trained BioELECTRA (Kanakarajan et al., 2021) checkpoint as in §4.3. To assess the effectiveness of clinical note modeling, we employ another baseline using a 4-layer feedforward model (∼5M parameters), which takes an input of 3,942 dimension high-level features from the EHR database (EHR-FFN). Similar to DeepPatient (Miotto et al., 2016), we extract high-level features from the diagnoses, medications, procedures, and laboratory test records by counting the occurrence of each feature type. In addition, we also add other features such as length of stay, the indicator for emergency unit admission, age group, etc. The details of EHR-FFN and the extracted features are shown in Appendix B.

5.3 Training and Evaluation

We train all of the models with an initial learning rate of 5e-5, batch size of 48, and a linear learning rate decay. We train the model for 3 epochs and test the model with the best validation score. For evaluating the diagnosis label, we incorporate the F1-score along with the Top-1, Top-3,and Top-5 accuracy scores. For the mortality label, we incorporate F1-score and AUC. The evaluation is conducted on two different test sets: (i) the *seen* patient test set and (ii) the *unseen* patient test set.

5.4 Results and Analysis

Effect of Clinical Note Modeling We show our experiment results for the *seen* and the *unseen* test sets in Table 4. All BioELECTRA models yield higher results than the EHR-FFN for both test sets, showing the effectiveness of clinical note modeling for disease risk and mortality risk predictions using EHR data. From the comparison between different clinical notes interval of the BioELECTRA model, we found that modeling longer clinical note interval will likely increase the performance on both tasks. This behavior aligns with the results reported in §2. Nevertheless, this behavior does not apply to the mortality risk prediction on the *seen* test set. We describe this phenomenon further in §5.4.

Generalization to New Patient Data We observe that there is a huge gap of performance for the baseline EHR-FFN model, especially in the diagnosis predictions of *seen* and *unseen* test set (∼40 p.p.). In this case, utilizing clinical note modeling closes the performance gap on the *seen* and *unseen* test sets to be much narrower (∼10 p.p.) on either label, especially for the BioELECTRA model with longer context length. This suggests that longer clinical notes interval not only improves the performance of the model on the similar patient record distribution, but also improves the performance on

[11] https://www3.ha.org.hk/cancereg/allages.asp

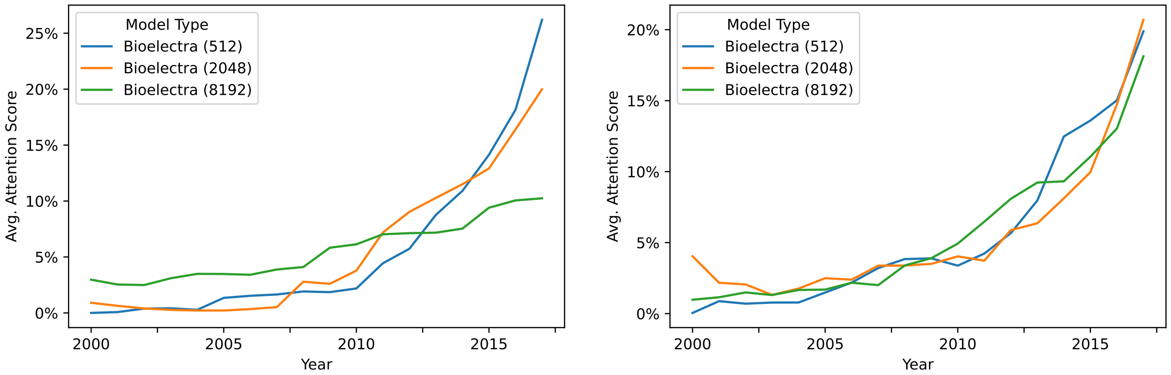

Figure 2: **(Left)** and **(Right)** show the clinical notes **time importance** of the disease risk prediction and mortality risk prediction, respectively.

the out-of-distribution patient records.

Optimal Cut-off Interval for Disease Risk and Mortality Risk Prediction We measure the number of clinical notes that can be processed by the models to analyze the optimal cut-off interval. Using the length statistics on our dataset, we find that our BioELECTRA (512), BioELECTRA (2048), and BioELECTRA (8192) models can encode 4, 17, 66 clinical notes on average, which correspond to the average clinical note intervals of 2-3 months, ∼1 year, and 3.5 years, respectively. As shown in Table 4, for the disease risk prediction label, the utilization of longer clinical notes intervals always yields better performance, while the same trend is not observed for the mortality risk label. This evidence suggests that there are different optimal interval of clinical notes required to infer the correct prediction for different target labels.

To verify this phenomenon, we analyze the input fractions considered to be important by the models. Specifically, we retrieve 1,000 correctly-predicted samples with the highest confidence values from each of the models and collect the clinical note timestamps corresponding to the high-magnitude (>5% of the total input gradient magnitude) input gradient with respect to the output prediction by using saliency map (Simonyan et al., 2014; Yosinski et al., 2015; Wallace et al., 2019). The timestamps from all samples are then aggregated with yearly granularity. We denote the number of year occurrences divided by the total number of timestamps collected as **time importance** to show how likely the model attends to the clinical note from the corresponding year given the label prediction in 2018.

As shown in Figure 2, for the disease risk label, the slope of the **time importance** curves over the years become more flattened as the utilized clinical note interval widens, indicating that the **time importance** spreads more uniformly on longer clinical note intervals. Whereas for the mortality risk label, the **time importance** curve has a similar slope over different clinical notes intervals. This evidence supports that for modeling an accurate disease risk prediction, a long clinical note interval ($\geq$ 3.5 years) is required. While for mortality risk prediction, a shorter clinical note interval (∼2-3 months) is sufficient to reach optimal performance.

6 Conclusion

In this paper, we show the importance of capturing longer clinical notes for biomedical and clinical large pre-trained LMs on 6 clinical NLP tasks on the United States and Hong Kong clinical note data. Our result suggests that utilizing longer clinical notes can significantly increase the performance of LMs by ∼5-10% F1-score without the loss of generalization to the unseen data. We also observe that incorporating a longer interval of clinical notes does not always entail performance improvement and there is an optimal cut-off interval depending on the target variable. Based on our analysis, we conclude that an interval of ∼2-3 months is the optimal cut-off for mortality risk prediction, while 3.5 years or an even longer interval of clinical notes is required to achieve the optimal performance for disease risk prediction. Future work in long-range clinical note modeling would open up opportunities towards a general solution in clinical NLP.

7 Limitation

Although there are many linear attention mechanisms that have been proposed (Dai et al., 2019; Kitaev et al., 2020; Beltagy et al., 2020; Zaheer et al., 2020), the exploration of linear attention in our experiments is currently limited to Longformer (Beltagy et al., 2020). Furthermore, the constructed longitudinal clinical note dataset from the Hong Kong Hospital Authority EHR system cannot be made public due to the data-sharing policy. Lastly, due to the limited computational power, we only conduct the long-range clinical notes experiment for bio-lm and BioELECTRA for the n2c2 experiment and BioELECTRA for the Hong Kong longitudinal dataset. We conjecture that the performance of the long-range versions of other pre-trained models will follow similar trends to the result on existing biomedical and clinical benchmarks.

Acknowledgements

This work has been partially funded by School of Engineering PhD Fellowship Award, the Hong Kong University of Science and Technology and PF20-43679 Hong Kong PhD Fellowship Scheme, Research Grant Council, Hong Kong.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel Cahyawijaya, Tiezheng Yu, Zihan Liu, Xiaopu Zhou, Tze Wing Tiffany Mak, Yuk Yu Nancy Ip, and Pascale Fung. 2022. SNP2Vec: Scalable self-supervised pre-training for genome-wide association study. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 140–154, Dublin, Ireland. Association for Computational Linguistics.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Edward Choi, Cao Xiao, Walter F. Stewart, and Jimeng Sun. 2018. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. *arXiv:1810.09593 [cs, stat]*. ArXiv: 1810.09593.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. 2021. Rethinking attention with performers. In *International Conference on Learning Representations*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Jean-Baptiste Escudié, Bastien Rance, Georgia Malamut, Sherine Khater, Anita Burgun, Christophe Cellier, and Anne-Sophie Jannot. 2017. A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease. *BMC Med. Inform. Decis. Mak.*, 17(1).

Jason Alan Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Myungsun Kang, Ruisi Su, Wojciech Kusa, Samuel Cahyawijaya, et al. 2022. Bigbio: A framework for data-centric biomedical natural language processing. *arXiv preprint arXiv:2206.15076*.

Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, Eric T. Carlson, Joy T. Wu, Jonathan Welt, John Foote, Edward T. Moseley, David W. Grant, Patrick D. Tyler, and Leo A. Celi. 2018. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLOS ONE*, 13(2):e0192360.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

K Haerian, D Varn, S Vaidya, L Ena, H S Chase, and C Friedman. 2012. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clin. Pharmacol. Ther.*, 92(2):228–234.

Ahmad Hammoudeh, Ghazi Al-Naymat, Ibrahim Ghannam, and Nadim Obeid. 2018. Predicting hospital readmission among diabetics using deep learning. *Procedia Computer Science*, 141:484–489.

Shaoxiong Ji, Matti Hölttä, and Pekka Marttinen. 2021. Does the magic of bert apply to medical code assignment? a quantitative study. *Computers in Biology and Medicine*, 139:104998.

Alistair Johnson, Tom Pollard, and Roger Mark. 2019. Mimic-iii clinical database demo.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELECTRA:pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online. Association for Computational Linguistics.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Franccois Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv:2001.04451 [cs, stat]*. ArXiv: 2001.04451.

Theresa A Koleck, Caitlin Dreisbach, Philip E Bourne, and Suzanne Bakken. 2019. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association*, 26(4):364–379.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen Mckeown. 2020. Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

P LePendu, S V Iyer, A Bauer-Mehren, R Harpaz, J M Mortensen, T Podchiyska, T A Ferris, and N H Shah. 2013. Pharmacovigilance using clinical notes. *Clinical Pharmacology & Therapeutics*, 93(6):547–555.

Paea LePendu, Srini Iyer, Cedric Fairon, and Nigam Haresh Shah. 2012. Annotation analysis for testing drug safety signals using unstructured clinical notes. *Journal of Biomedical Semantics*, 3:S5 – S5.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Feifan Liu, Abhyuday Jagannatha, and Hong Yu. 2019. Towards drug safety surveillance and pharmacovigilance: Current progress in detecting medication and adverse drug events from electronic health records. *Drug Saf.*, 42(1):95–97.

Jingshu Liu, Zachariah Zhang, and Narges Razavian. 2018. Deep ehr: Chronic disease prediction using medical notes. *Journal of Machine Learning Research (JMLR)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. 2016. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.*, 6:26094.

Kevin M Pantalone, Todd M Hobbs, Kevin M Chagin, Sheldon X Kong, Brian J Wells, Michael W Kattan, Jonathan Bouchard, Brian Sakurada, Alex Milinovich, Wayne Weng, Janine Bauman, Anita D Misra-Hebert, Robert S Zimmerman, and Bartolome Burguera. 2017. Prevalence and recognition of obesity and its associated comorbidities: cross-sectional analysis of electronic health record data from a large us integrated health system. *BMJ Open*, 7(11).

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*.

Jiezhong Qiu, Hao Ma, Omer Levy, Wen-tau Yih, Sinong Wang, and Jie Tang. 2020. Blockwise self-attention for long document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2555–2565, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2020. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. 2018. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

Amber Stubbs, Michele Filannino, Ergin Soysal, Samuel Henry, and Özlem Uzuner. 2019. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *J. Am. Med. Inform. Assoc.*, 26(11):1163–1171.

Amber Stubbs, Christopher Kotfila, Hua Xu, and Özlem Uzuner. 2015. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task track 2. *J. Biomed. Inform.*, 58 Suppl:S67–S77.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *ACM Comput. Surv.* Just Accepted.

Özlem Uzuner. 2009. Recognizing Obesity and Comorbidities in Sparse Data. *Journal of the American Medical Informatics Association*, 16(4):561–570.

Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying Patient Smoking Status from Medical Discharge Records. *Journal of the American Medical Informatics Association*, 15(1):14–24.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. Allennlp interpret: A framework for explaining predictions of NLP models. *CoRR*, abs/1909.09251.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. Cite arxiv:2006.04768.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, X. Li, Zhi Yuan Lim, S. Soleman, R. Mahendra, Pascale Fung, Syafri Bahar, and A. Purwarianti. 2020. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*.

Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, and Pascale Fung. 2020. Lightweight and efficient end-to-end speech recognition using low-rank transformer. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6144–6148.

Li-Tzy Wu, Kenneth R Gersing, Marvin S Swartz, Bruce Burchett, Ting-Kai Li, and Dan G Blazer. 2013. Using electronic health records data to assess comorbidities of substance use and psychiatric diagnoses and treatment settings among adults. *J. Psychiatr. Res.*, 47(4):555–563.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. In *Deep Learning Workshop, International Conference on Machine Learning (ICML)*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. In *NeurIPS*.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

A Label Distribution

Our Hong Kong longitudinal clinical notes dataset is extracted from Hong Kong Hospital Authority EHR system which covers records from 43 hospitals in Hong Kong. For the diagnosis, to reduce the dimensionality, we group the diagnosis labels into 79 classes. For cancer diseases, we group the diagnosis based on the cancer site categorization from the Hong Kong Cancer Registry[12]. While for other diseases, we take the first three digits of the ICD-10 codes as the grouping. We show the label distribution of our Hong Kong longitudinal clinical notes dataset in Figure 3.

B Detail of EHR-FFN Model

We derive 3,942 features from the tabular data for each encounter. We derive these features from 4 data tables: diagnosis, procedure, prescription, and inpatient data. Specifically, we generate one-hot representations for each derived feature and concatenate all the one-hot representation into a single vector . The detail of each one-hot feature is shown in Table 5. We extract the feature vectors per patient encounter. To aggregate all the historical tabular feature vectors, we aggregate the vectors into a single feature vector by summing up all the vectors producing a single high-level feature vector per patient. To learn the high-level feature vector, we employ a feed forward network with 3 hidden layers with a total size of ∼5M parameters. The hyperparameters of the feed forward model is shown in Table 6.

Feature Name	Length	Description
Diagnosis Type	1699	Diagnosis type based on ICD-10 code
Procedure Type	127	Procedure type based on ICD-9 code
Prescription Type	1271	Type of presribed drug based on regional standard
Prescription BNF	73	Type of presribed drug based on BNF Therapeutic Classification
Emergency Indicator	1	Indicator for emergency unit admission
Length of Stay	5	Length of stay in the hospital
Age Group	5	Age of the patient during admission to the hospital
Ward Type	4	Type of hospital ward
Ward Sub-Care Type	6	Sub-type of hospital ward

Table 5: Details of the tabular features

[12] https://www3.ha.org.hk/cancereg/allages.asp

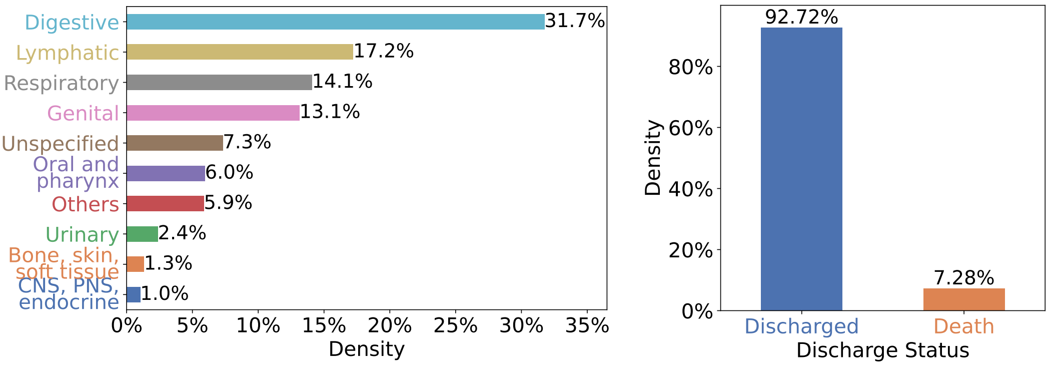

Figure 3: Label statistics of our dataset. **(Left)** shows the aggregated distribution of diagnosis based on the cancer ICD-10's site grouping[13]. `Unspecified` denotes all cancer diagnoses with unspecified site. `Others` denotes diseases other than cancer. **(Right)** shows the distribution of the discharge status (discharged/death) gathered from all inpatient records, which is used to define the mortality label.

Hyperparameter settings	**Value**
Tabular Encoder	
#hidden layers	3
hidden size	[1024, 512, 256]
input size	3942
layer activation	ReLU
drop out	0.1

Table 6: Details of the model hyperparameters