DK-BEHRT: Teaching Language Models International Classification of Disease (ICD) Codes using Known Disease Descriptions

Ulzee An[1,2], Simon A. Lee[2], Moonseong Jeong[1] Aditya Gorla[3] Jeffrey N. Chiang[2,4] Sriram Sankararaman[1, 2,5]

[1]Department of Computer Science, UCLA
[2]Department of Computational Medicine, UCLA
[3]Bioinformatics IDP, UCLA
[4]Department of Neurosurgery, UCLA
[4]Department of Human Genetics, UCLA
ulzee@ucla.edu

Abstract

The widespread digitization of healthcare and patient data has created new opportunities to explore machine learning techniques for improving patient care. The sheer scale of this data has particularly motivated the use of deep learning methods like BERT, which can learn robust representations of medical concepts from patient data without the need for direct supervision. Simultaneously, recent research has shown that language models (LMs) trained on scientific literature can capture strong domain-specific knowledge, including concepts highly relevant to healthcare. In this paper, we leverage two complementary sources of information—patient medical records and descriptive clinical text—to learn complex clinical concepts, such as diagnostic codes, more effectively. Although significant strides have been made in using language models with each data type individually, few studies have explored whether the domain expertise acquired from scientific text can provide a beneficial inductive bias when applied to learning from patient records. To address this gap, we propose the Domain Knowledge BEHRT (DK-BEHRT), a model that integrates disease description embeddings from domain-specific language models, like BioGPT, into the attention mechanisms of a BERT-based architecture. By incorporating these "knowledge" embeddings, we aim to enhance the model's ability to understand the clinical concept (e.g. ICD Codes) more effectively and predict clinical outcomes with higher accuracy. We validate this approach on the MIMIC-IV dataset and find that incorporating specialized embeddings consistently improves predictive accuracy for clinical outcomes compared to using generic embeddings or training the base model from scratch.

Introduction

The increasing availability of electronic health records (EHR) datasets has motivated research into whether complex health-related patterns can be learned directly from patient data using machine learning. While traditional methods have relied on manual feature extraction and domain expertise, deep learning has gained interest for its ability to learn from diverse information (such as diagnoses, medications, and clinical notes) without requiring complex model designs.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Among deep learning approaches, the work of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019) has offered a straightforward approach to uncovering discriminative patterns in vast data without strong supervision, relying on the transformer architecture (Vaswani 2017). Originally developed for generic natural language processing (NLP) tasks, BERT uses a Masked Language Modeling (MLM) objective, which involves masking random words in a sentence and training the model to predict them, helping the model learn general-purpose relationships between arbitrary observations. Given its success in NLP (Acheampong, Nunoo-Mensah, and Chen 2021; Ono and Lee 2024), BERT has been adapted for the biomedical domain through specialized versions (Li et al. 2020; Rasmy et al. 2021; Rupp, Peter, and Pattipaka 2023; Zhou et al. 2023; Antal et al. 2024). Through a similar training procedure which masks medical observations (such as one diagnosis in time) in a patient's medical history, the model learns a high-dimensional internal representation of observed features that have been shown to yield state of the art accuracies when predicting clinical outcomes (Hager et al. 2024; Li et al. 2024; Lee, Brokowski, and Chiang 2024).

Given its ability to learn from free text effectively, the transformer architecture has also been used extensively in a related domain: learning high-level insights from scientific research articles (Beltagy, Lo, and Cohan 2019; Gu et al. 2021), and specifically those pertaining to biological findings (Huang, Altosaar, and Ranganath 2019; Lee et al. 2020; Gu et al. 2021; Touvron et al. 2023; Achiam et al. 2023; Shin et al. 2020; Luo et al. 2022).

At first glance, the distinction between language models for patient records and biomedical knowledge can appear similar as their underlying transformer architectures and training process are similar (several BERTs and GPTs have been proposed for both types of data). However, a question still remains on the degree to which knowledge gained using state-of-art methods in patient records or published knowledge is largely overlapping or is highly disjoint.

To address these questions, our goal is to use clinical features from electronic health records (EHR) along with known descriptions of diseases to make ICD codes more contextual and relevant for predicting patient outcomes. We

propose training a model, similar to BERT, on EHR data, enhanced with "domain knowledge" from human-curated disease descriptions to better understand the learning of language models on the underlying disease latent space. By incorporating knowledge from these descriptions and biomedical language models, we aim to improve the model's ability to predict patient outcomes and potential future diagnoses.

The proposed methodology proceeds as follows: we first train the model using a strategy similar to MedBERT (Rasmy et al. 2021), with the primary objective of optimizing accuracy in predicting clinical outcomes. We then introduce a novel mechanism that enables the model to interpret relationships between diagnosis codes by incorporating connections derived from disease descriptions, represented through latent representations from an external language model. This approach provides a more contextualized understanding of each ICD code, leading to enhanced predictions of patient health.

Under our framework, we explore the usefulness of the latent representations of several LMs, ranging from those trained on generic text (OpenAI's text-embedding-3-large[1]) to those trained specifically on PubMed publications (BioGPT, (Luo et al. 2022)). We implement a benchmark procedure similar to (Harutyunyan et al. 2019; Wang et al. 2020) based on the MIMIC-IV dataset (Johnson et al. 2024, 2023), measuring the model's ability to predict mortality and lengths of stay in ICU visits after undergoing a finetuning procedure. In addition, we survey the models' ability to predict if a patient will receive a life-altering diagnosis for the first time in three broad categories: major depressive disorder (F320∼F329), cardiac complications (I250∼I259), and motor neuron disorders (G20∼G259). Through our benchmarks, we observe that the introduction of LM embeddings in BERT training is a net benefit, as well as finding that using embeddings from LMs trained on biomedical texts are exclusively preferable in comparison to ones trained on generic texts.

The clinical relevance of this work lies in its potential to bridge gaps in how complex medical concepts (e.g. ICD Codes, SNOMED Codes, CPT Codes, Medications, Procedures, etc.) like ICD codes are understood and used by language models for predicting patient outcomes. These emerging technologies have shown promise in many applications, but in healthcare, unique challenges arise that differ from those in natural language processing, requiring a more focused approach to training and understanding medical concepts. For example ICD codes often reflect complex, nuanced conditions, but without context, models trained solely on these codes may fail to capture the full clinical picture due to their non-trivial design (a series of alpha numeric characters with a non-trivial pattern). For instance, arbitrary or less explicit relationships between certain diagnoses may not be effectively learned, leading to gaps in the model's understanding of clinically significant connections. By integrating descriptions of diseases, we enhance the model's contextual understanding, making these codes more informative and relevant for outcome predictions. This approach ensures that the model can access latent relationships and broader patterns that are often missed when relying solely on ICD codes, thus improving its clinical applicability for language models to learn and predict outcomes for real-world patient care scenarios. Using a design similar to the one we propose can work towards building a generalist AI as current medical LM's struggle to grasp the nuances of clinical concepts as noted in the recent literature (Soroush et al. 2024; Lee and Lindsey 2024; Chen et al.).

We review a series of works related to our approach in the Related Works Section and explain our methodology in greater detail in Methods Section. We present our findings in the Results Section as well as sharing ablation experiments to determine the amount of contribution from the biomedical LM embeddings.

[1]https://openai.com/index/new-embedding-models-and-api-updates/

Related Works

Representation learning with electronic health records

Given the success of the BERT model (Devlin et al. 2019) in the field of natural language processing, many works have sought to explore their utility in the biomedical domain where various types of records can be found in text form. Early works such as BEHRT (Li et al. 2020) and MedBERT (Rasmy et al. 2021) demonstrated the potential of this approach. Recent works continue to expand the variety of features and covariates that are made available to the model depending on the dataset (ExBEHRT (Rupp, Peter, and Pattipaka 2023), IRENE (Zhou et al. 2023), M-BioBERTa (Antal et al. 2024)), which is expected to benefit the contextual embedding process. Works such as TransformEHR (Yang et al. 2023c), Gatortron (Yang et al. 2022), and CLMBR (Wornow et al. 2023) also propose that the use of transformer decoders (only forward-directional attention) can improve predictive accuracy on downstream tasks that generally try to forecast medical incidences in the future.

Language models for biomedical knowledge

Transformer architectures have also been explored for the purpose of creating language models (Brown et al. 2020) that reason with knowledge rooted in biomedical scientific research. The BERT framework has also been used in this context: early works demonstrated that BERT can extract reliable representations of scientific concepts from research articles (SciBERT (Beltagy, Lo, and Cohan 2019)), and that fitting them specifically on domain specific text directly benefits their capabilities on domain specific tasks (Alsentzer et al. 2019). Several variations have been proposed with varying sources of the research text such as ClinicalBERT (Huang, Altosaar, and Ranganath 2019), BioBERT (Lee et al. 2020), PubMedBERT (aka. BiomedNLP) (Gu et al. 2021), and BioMegatron (Shin et al. 2020).

Based on the recent success of conversational language models (Touvron et al. 2023; Achiam et al. 2023), recent works have pursued models that are decoder-based (forward-directional attention, unlike BERT), which are better suited for generating text. Notable works using this ap-

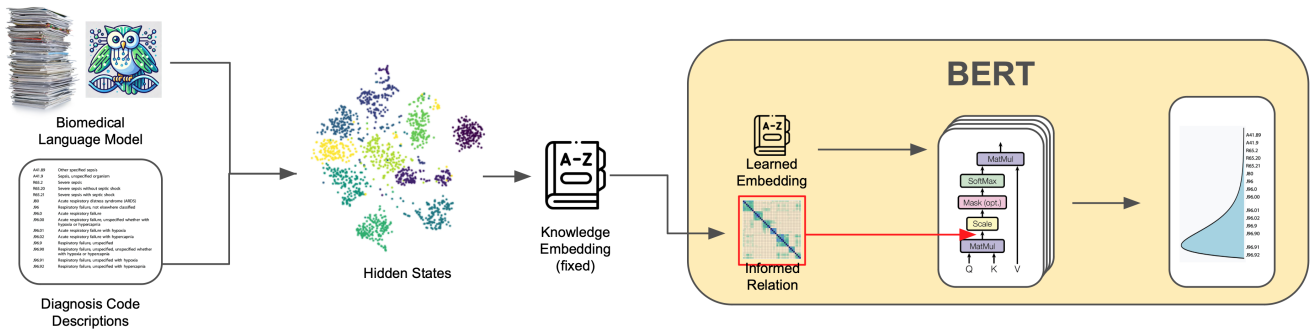

Figure 1: Proposed flowchart of DK-BEHRT.

proach include BioGPT (Luo et al. 2022). The model still learns embeddings for each token, allowing one to examine its internal representations during the generation process. The recent work by (Kane et al. 2023) finds that BioGPT produces the current state-of-art in embeddings of diagnosis codes (ICD 10) in terms predicting their semantic relations. Further works in this area demonstrate a wide array of capabilities such as identifying clinical concepts in freehand medical text (Gu et al. 2021; Vu, Nguyen, and Nguyen 2020), allowing interactive chatting on clinical topics (Varshney et al. 2023), or writing discharge notes (Ellershaw et al. 2024).

Predicting clinical outcomes with language models

While language models have excelled in both learning patient narratives in health records and biomedical knowledge form research articles, few works explore whether knowledge learned from either sources of information have any synergy.

A series of recent works explore whether language models trained on general text or biomedical scientific text are performant in predicting clinical outcomes out of the box given patient records. Examples of such works include (Gupta et al. 2022), MIMIC-IV-Ext (Hager, Jungmann, and Rueckert), (Yang et al. 2023a), MEME (Lee et al. 2024), CliBench (Ma et al. 2024b), and (Hager et al. 2024; Li et al. 2024; Lee et al. 2025; Lee, Lee, and Chiang 2024). Notably, (Hager et al. 2024) finds that conversational language models are currently not reliable enough for real-world use in clinical settings. The works do not consider whether the language model should learn from the patient records directly to reinforce their scientific knowledge, although if such direction would truly be beneficial is not clearly assessed.

To our knowledge, Gatortron (Yang et al. 2022) has been the most comprehensive effort to incorporate patient-level records (clinical notes) and aggregate knowledge (PubMed articles and Wikipedia) simultaneously in a language model. Our work differs in that we propose obtaining the domain knowledge from a model off the shelf, as opposed to fitting an entirely new language model from scratch which is a costly procedure. We also note that MIMIC-III was part of Gatortron's training data, complicating a fair comparison in our experimental setup.

Learning diagnosis codes with prior knowledge

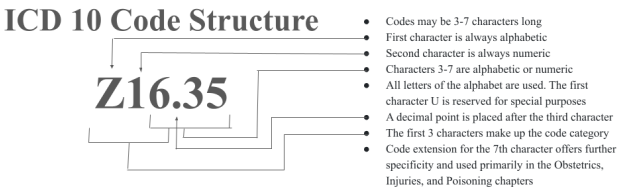

Figure 2: A visual of the deliberate and non-trivial design of the International Classification of Disease (ICD) code studied in our work.

Several challenges exist in allowing deep learning architectures to learn diagnosis codes effectively (Huang, Tsai, and Chen 2022). The main challenge is the large label space of conventional diagnosis coding systems (70k+ for ICD 10 and 360k+ for SNOMED); additional ambiguities exist when applying language models such as how diagnosis histories should serialized and tokenized. The illustration in Figure 2 explains the structure of such ICD codes. Existing works have proposed paired network to learn the semantic similarities (Yuan, Tan, and Huang 2022), enforce a generation procedure according to the code hierarchy (Yang et al. 2023b), use a featurization that accounts for parent codes (Wornow et al. 2023), or directly finetune an LM to reiterate hierarchical relations of the codes (MERA, (Ma et al. 2024a)). While the experimental setting of these works differ from ours, they each find that the semantic awareness in dealing with ICD codes has the potential of improving performance in downstream tasks. Our work also seeks to exploit the underlying semantic structure of ICD codes. A key distinction of our work in this regard is that we examine the semantic similarity of the codes according to an LM's understanding of its scientific domain, as opposed to adhering to the hierarchy determined by the convention.

Attention with inductive bias

Additional context may be passed to a transformer-based model via additional tokens or additional positional encod-

ings (modify the existing tokens); both approaches influence the attention of the transformer. For instance, HVAT (Shao et al. 2023) proposed an auxiliary embedding procedure with concept embeddings, and explored embeddings relating to Alzheimer's disease.

Instead, several prior works have proposed modifying the attention mechanism directly to partially bypass its quadratic cost (which increases with additional tokens) or the need for the model to disentangle encodings added to the embedding. For instance, the Graphormer model (Ying et al. 2021) proposes to close the gap between transformers and graph neural nets by adding relation matrices computed with prior node and edge information into the attention computation. In principle, the proposed mechanism is similar to one explored in our work, albeit our proposed usecase differs.

Past works have also applied the idea in the context of improving the estimation of edges in graph neural networks in conjunction with a knowledge graph (Choi et al. 2017; Ma et al. 2018; Ye et al. 2021; Gao et al. 2022; Ma et al. 2022). Our work is conceptually similar to the prior works' use of knowledge graphs, however the previously proposed pipelines are considerably more complex to configure than the standard BERT approach.

Methods

We propose leveraging knowledge aggregated by a biomedical language model to improve the prediction of health-related outcomes from medical records. To do this, we modify a standard BERT architecture and pipeline to utilize embeddings of ICD code descriptions as encoded by the LM (generally trained on research articles published in the biomedical domain) while it learns from a dataset of patient records. Our work focuses on the understanding of human diseases and their progression through ICD codes; to better observe the impact of our contribution, we deliberately exclude prescriptions and laboratory results from our study. An overall flowchart of our proposed approach is visualized in Figure 1.

Diagnosis code or "Knowledge" embeddings

We propose the use of a medical language model to obtain transferable representations of disease diagnoses in the form of diagnosis code embeddings (referred to as knowledge embeddings for brevity). We obtain the plain English description of all 71,704 possible ICD 10 codes listed by the Centers for Medicare Medicaid Services (CMS) [2]. We then process each diagnosis code description through a given a medical language model to obtain the embedding (using their default tokenizer). Current language models infer a hidden state of size of $768 \sim 4096$ (depending on the model) preceding the final layer for each token that is processed. Given a description text of length l_c in tokens for code c, we obtain the hidden states $\mathbf{h}_c = \{h_{c,1}, \ldots, h_{c,l_c}\}$ and take the average of the hidden states across the sequence dimension to obtain the embedding $e_c = \frac{1}{l_c}\sum_{i=1}^{l_c} h_{c,i}$. Our approach relies on these embeddings which are generated once for all relevant ICD codes and do not have to be recomputed in later stages. Furthermore, we do not cut off the codes to simplify the space of all ICD codes. (all codes observed in the datasets are used as-is regardless of their specificity, unlike prior works).

[2]https://www.cms.gov/medicare/coding-billing/icd-10-codes

Knowledge-based attention

For a given dataset consisting of D diagnosis codes, we obtain the relevant pre-computed embeddings $\mathbf{e} = \{e_1, \ldots, e_D\}$, which we refer to as knowledge embeddings. We explore the usage of knowledge embeddings in the attention operation of the standard Transformer layer to allow the model to relate diagnosis codes more easily without needing to learn them from scratch.

Specifically, given a sequence of diagnosis token indices over time $d_i = \{d_{i,1}, \ldots, d_{i,l_i}\}$ for the i-th patient history with length l_i, we obtain the corresponding sequence of knowledge embeddings $e(d_i) = \{e_{[d_{i,1}]}, \ldots, e_{[d_{i,l_i}]}\}$. Then, we compute the similarity matrix $S_i = f(e(d_i))^T f(e(d_i))$ where the operation $f(e(d_i))$ applies a learnable shared linear projection f over all embeddings $f(e(d_i)) = \{f(e_{[d_{i,1}]}), \ldots, f(e_{[d_{i,l_i}]})\}$.

Given the attention operation for the i-th sequence, we introduce the matrix S_i inside the softmax operation, a step that is responsible for calculating the relative importance of tokens over the sequence given any specific position of the sequence:

$$\text{Attn}(Q_i, K_i, V_i) = \text{softmax}(\frac{Q_i K_i^T}{\sqrt{d}} + S_i)V_i \qquad (1)$$

The operation inside the softmax consists of $Q_i \in R^{l_i \times d}$, $K_i \in R^{l_i \times d}$, given latent dimension size d. Intuitively, the presence of S_i allows the Transformer layer to start attending to diagnosis codes which are conceptually related immediately without having to learn them from the dataset. We implement the knowledge-based attention operation for all layers in the BERT model. While further enhancements can be made to the attention mechanism, we note a tradeoff between additional complexity and the module's compute cost.

DK-BEHRT Architecture

We begin by modifying a transformer architecture which is similar to MedBERT (Rasmy et al. 2021). We clarify that either the medical language model (for the prior knowledge) and the outcome prediction model may leverage BERT, and our contribution is mainly in the latter application. We configure the model with 4 transformer layers resulting in a 18 million parameter model (within the scope of the compute available for our work, we did not observe increased performance with additional layers). Specific to our approach, we replace all attention operations in the transformer layers with the knowledge-based attention unit.

As each patient's diagnosis history was passed to the model, both their sex and age at the point of diagnosis are also made available to the model as covariates. We convert the sex and age information simultaneously into a simple positional encoding, where a constant floating point value

Table 1: The number of samples which were parsed from the MIMIC-IV dataset (total over training, val, and test sets). The "hospital" set reflects the total number of samples for whom pretraining could be performed. The "icu-" and "hosp-" counts reflect the total number of samples considered in benchmarks. The number of cases for each benchmark is also reported.

group	unit	total	cases
hospital	patients	180,644	-
icu	patients	65,366	-
icu-mortality	visits	267,987	83,207
icu-lenofstay	visits	267,987	73,520
hosp-F320∼F329	visits	74,877	6,807
hosp-I250∼F259	visits	30,661	3,804
hosp-G20∼G259	visits	16,522	1,502

between $0.00 \sim 0.10$ for males is added to embeddings ($0.00 \sim -0.10$ for female) pertaining to the age of the patient proportionally in the range of $0 \sim 100$ at the time of each diagnosis. The absolute date-time of diagnosis was not considered as an additional feature since the information was not available in MIMIC (and potentially in other datasets that undergo anonymization). Finally, we implement a tokenizer which take series of ICD10 codes and converts them into one-hot indices.

For the purpose of masked language model training, we define a linear prediction head which is responsible for predicting the probability over all possible tokens for masked positions. During the finetuning stage, we define similar linear prediction head for binary classification.

Training and Finetuning

We adopt a two-stage training procedure similar to the methodology followed by (Rasmy et al. 2021; Rupp, Peter, and Pattipaka 2023). The initial stage involves pretraining MedBERT and DK-BEHRT on all medical histories of patients to enable the model to learn general representations. During this stage, the model is trained using masked token modeling (with masking probability of 50%), which helps the model understand the semantics of diagnosis codes across a diverse range of clinical narratives. The pretraining step has been found to be crucial in allowing such models to perform more effectively on specialized tasks. Following the pretraining phase, we perform a fine-tuning procedure where the pretrained model is further trained on a task-specific dataset. We allow all weights of the model to be updated given a small learning rate (5×10^{-5}). We share further characteristics of the training process in Appendix , as well as sharing code that can be used to replicated our results [3].

Datasets and Benchmarks

We explore diagnosis histories of patients in the MIMIC IV dataset (Johnson et al. 2023), a publicly available anonymized medical records dataset. Following (Harutyunyan et al. 2019), we used a standard procedure to define five downstream tasks for which all methods were evaluated. We first defined mortality and length of stay prediction tasks based on ICU visits. Mortality of a patient up to 1 year after their ICU visit was considered to be a case. The length of stay prediction was treated as a binary classification task for stays exceeding 72 hours.

In addition to the two ICU-based tasks, we determined the times at which patients were first diagnosed with three diseases of interest: major depressive disorder, chronic ischaemic heart disease, and extrapyramidal and movement disorders. The diseases corresponded broadly to ICD code blocks of F320∼F329, I250∼I259, and G20∼G259. We considered a series of visits leading to the target diagnosis within a range of 180 days as a case. Controls were considered as any sub-series of visits (starting from their first known) where the diagnosis was not observed in the given range. We measure the precision recall score (PR), receiver operating characteristic score (ROC), and F1 score for all tasks.

Starting with the pre-training stage, a train-test-validation split of 70%/15%/15% was generated across patients. Patients who were grouped in the training and validation sets were used to fit the model in the pre-training and finetuning stages, while patients in the test set were reserved exclusively for evaluation after the finetuning stage. We report in Table 1 the statistics of the dataset and the number of cases and controls identified for each task.

Model Optimizations

We implement the knowledge-based attention layer, the pretraining stage, and the finetuning stage using the Huggingface framework [4]. We rely on the framework's default optimizer (AdamW, (Loshchilov 2017)) and the default linear learning rate scheduler for all fitting stages. We allow the pre-training to continue for 100,000 steps with a learning rate of 5×10^{-4} after which the model approximately converged. During finetuning, we set the starting learning rate as 5×10^{-5} and checkpoint the model every 50 minibatches. The checkpoint with the lowest validation loss is recovered to assess the accuracy on the final test set. Finetuning was allowed to run for 10 epochs within which we observed all method were able to begin overfitting the dataset.

The entire pipeline for one configuration of the model could be run using a single 48GB A100 GPU within the span of 1 day. We plot the time elapsed over number of steps taken for both the base BERT architecture and the proposed BERT architecture in Figure 3 (for the mortality benchmark). As the two architectures are similar, the additional time elapsed can be fully attributed to the modified attention layer.

Results

Evaluated methods

We evaluated Logistic Regression and XGBoost as baselines to establish the relative difficulty of the benchmark tasks.

[3] https://anonymous.4open.science/r/icdbert-F270

[4] https://huggingface.co/docs/transformers/en/index

Table 2: The precision recall (PR), receiver operating characteristic (ROC), and F1 metrics in predicting clinical outcomes and future diagnoses in the MIMIC dataset. The highest scores are in bold, with second highest underlined (no highlight for ties up to 0.1 percent). 95%-CI shown in parentheses. DK-BEHRTs are prefixed as DKB, followed by the embedding source.

	Mortality			Length of Stay		
Model	PR	ROC	F1	PR	ROC	F1
LR	42.1 (0.7)	65.7 (0.6)	67.6 (0.4)	35.6 (0.7)	56.7 (0.7)	63.3 (0.5)
XGB	43.5 (0.5)	70.8 (0.3)	70.3 (0.3)	41.1 (0.8)	64.1 (0.6)	70.1 (0.5)
GatorTron	65.3 (0.6)	79.2 (0.4)	78.3 (0.4)	45.7 (0.9)	65.7 (0.6)	71.6 (0.3)
MedBERT	65.8 (0.8)	78.8 (0.6)	77.1 (0.5)	46.3 (1.0)	66.0 (0.8)	72.0 (0.4)
DKB TE3-L	63.7 (0.6)	78.4 (0.5)	75.7 (0.6)	46.6 (1.1)	66.0 (0.7)	71.7 (0.4)
DKB BioMega	66.1 (0.6)	79.3 (0.5)	77.5 (0.4)	**46.8** (1.1)	66.7 (0.7)	71.6 (0.4)
DKB PMBERT	66.1 (0.7)	79.4 (0.5)	**78.9** (0.5)	46.1 (0.9)	66.8 (0.8)	**72.2** (0.4)
DKB BioGPT	**68.0** (0.5)	**79.8** (0.5)	78.6 (0.4)	46.8 (1.0)	**67.3** (0.6)	71.9 (0.4)

	F320∼F329			I250∼ I259			G20∼ G259		
Model	PR	ROC	F1	PR	ROC	F1	PR	ROC	F1
LR	14.8 (2.2)	58.7 (1.7)	89.2 (0.8)	50.9 (4.1)	76.0 (2.2)	88.1 (0.7)	10.1 (1.4)	49.7 (3.6)	88.0 (1.2)
XGB	19.6 (3.2)	64.9 (1.7)	90.8 (0.6)	59.1 (2.5)	84.6 (1.2)	91.6 (0.8)	14.9 (3.1)	64.1 (4.6)	91.2 (0.9)
GatorTron	20.5 (2.2)	66.9 (1.7)	90.7 (0.6)	60.5 (2.3)	84.9 (1.1)	91.9 (0.6)	**18.4** (4.8)	63.2 (4.7)	91.4 (1.0)
MedBERT	21.1 (2.8)	68.0 (1.9)	90.8 (0.6)	60.5 (2.5)	84.5 (1.1)	91.8 (0.7)	13.7 (1.2)	65.7 (1.4)	90.8 (1.1)
DKB TE3-L	21.2 (2.6)	67.5 (1.8)	90.8 (0.6)	60.5 (2.6)	85.1 (1.0)	91.6 (0.8)	14.0 (2.8)	62.4 (4.6)	91.4 (1.0)
DKB BioMega	22.0 (2.7)	68.1 (2.0)	90.8 (0.6)	61.4 (2.1)	85.4 (0.9)	91.8 (0.8)	17.9 (4.6)	**67.1** (4.3)	91.4 (1.0)
DKB PMBERT	21.8 (3.2)	68.0 (1.7)	90.8 (0.6)	61.2 (2.2)	85.3 (0.9)	91.9 (0.8)	17.3 (5.5)	62.6 (5.2)	91.4 (1.0)
DKB BioGPT	**22.1** (2.5)	**68.5** (1.9)	90.8 (0.7)	**63.1** (1.9)	**86.3** (0.9)	**92.0** (0.7)	17.1 (4.4)	64.1 (4.8)	91.4 (1.0)

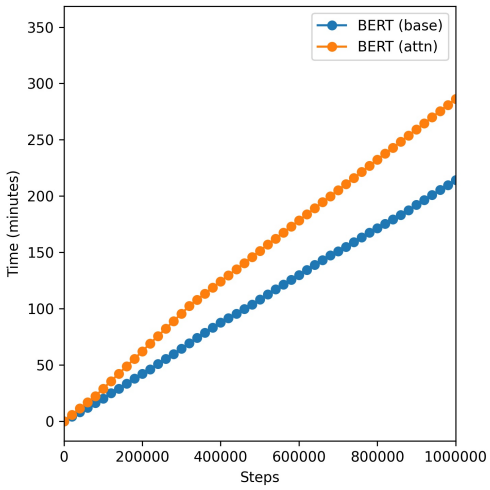

Figure 3: Time elapsed (minutes) in training a base BERT architecture (base) and the proposed DK-BEHRT architecture (BERT-attn). Training was done with a single 48GB A100 GPU

For both methods, we featurized all diagnoses received over past visits into a vector where each entry tallies the number of times a given diagnosis appeared, effectively collapsing the temporal aspect. To the feature vector, we appended the age of the patient at the latest visit in the visit history and the sex of the patient as covariates.

We explored several variations of DK-BEHRT. First, we determined the effectiveness of a relatively standard BERT model (akin to MedBERT (Rasmy et al. 2021)). We then explored the proposed architecture of DK-BEHRT with embeddings from four different language models: (1) **text-embedding-3-large** from OpenAI (TE3-L) which was not specifically trained on medical text, (2) **BioMegatron** ((Shin et al. 2020), MB BioMega) which was trained on a 6.1 billion-word dataset of PubMed abstracts & PMC articles, (3) **PubMedBERT** ((Gu et al. 2021) , MB PMBERT) which was trained on 14 million PubMed abstracts, and (4) **BioGPT** ((Luo et al. 2022), MB BioGPT) which was trained on 15 million PubMed abstracts.

Evaluation of Gatortron embeddings GatorTron was trained using a variety of sources of data: >90 billion words from University of Florida Health (>82 billion), Pubmed (6 billion), Wikipedia (2.5 billion), and MIMIC-III (0.5 billion) (Yang et al. 2022). As the version III and IV of MIMIC overlap in content and we could not rule out if GatorTron would have been trained on parts of our test split, we could not include it fairly and had to make a separate note about this model in the main results (Table 2).

Despite this drawback, we still evaluate GatorTron using our approach using our experimental setup and report its accuracy in benchmarks in Table 2. Although there are a few choices of the GatorTron model, we evaluate GatorTron-base, which is the smallest model (345 million parameters) due to compute limitations.

Table 3: Given the method with the highest number of top metrics (BioGPT and knowledge-based attention), we explore ablated versions DK-BEHRT. We replace the performant embeddings with entirely randomly generated embeddings (RandEmb), use the embeddings but finetune the downstream tasks with no pretraining (NoPretrain), and remove the shared projection layer for the embeddings (NoEmbProj). 95%-CI shown in parentheses.

	Mortality			Len of Stay		
Model	PR	ROC	F1	PR	ROC	F1
RandEmb	62.5 (1.0)	77.5 (1.1)	71.2 (0.5)	45.7 (1.0)	65.3 (0.6)	63.3 (0.5)
NoPretrain	52.1 (0.8)	72.7 (0.5)	71.9 (0.5)	38.5 (0.8)	61.3 (0.5)	70.6 (0.3)
NoEmbProj	66.3 (0.6)	78.7 (0.5)	78.9 (0.4)	47.3 (0.8)	67.1 (0.6)	72.6 (0.5)
DKB BioGPT	68.0 (0.5)	79.8 (0.5)	78.6 (0.4)	46.8 (1.0)	67.3 (0.6)	71.9 (0.4)

	F320∼F329			I250∼ I259			G20∼ G259		
Model	PR	ROC	F1	PR	ROC	F1	PR	ROC	F1
RandEmb	20.6 (1.6)	67.5 (2.0)	90.8 (1.1)	60.9 (2.7)	84.7 (1.1)	89.6 (2.0)	15.3 (1.2)	60.3 (1.1)	89.7 (0.8)
NoPretrain	19.3 (2.8)	67.5 (2.0)	90.8 (0.6)	60.4 (2.3)	83.9 (1.4)	91.9 (0.8)	15.1 (6.2)	60.8 (5.3)	91.43 (1.0)
NoEmbProj	21.4 (2.3)	67.3 (2.0)	90.8 (0.6)	60.6 (2.7)	84.4 (1.3)	91.4 (0.6)	18.3 (5.1)	66.3 (4.1)	91.4 (1.0)
DKB BioGPT	22.1 (2.5)	68.5 (1.9)	90.8 (0.7)	63.1 (1.9)	86.3 (0.9)	92.0 (0.7)	17.1 (4.4)	64.1 (4.8)	91.4 (1.0)

Main Results

We report in Table 2 the accuracies of the evaluated methods in predicting clinical outcomes and future diagnoses. We observe overall that DK-BEHRT improves prediction accuracy given downstream tasks. This could be seen first in that models with the modified attention obtained an higher accuracies on the tasks overall in comparison standard MedBERT (MB) implementation. Secondly, we note that all top scores are obtained with the help of embeddings from LMs trained on biomedical research text (BioMegatron, PubMedBERT, and BioGPT); no high scores are obtained with the help of text-embedding-3-large which is described as a general purpose embedding model (OpenAI's best embedding model at the time of writing). As BioGPT obtained the highest number of top scores, we also note that it can be considered a generative pre-trained relative of PubMedBERT, and was developed in the context of the earlier model (Luo et al. 2022).

We note that top score for specific disease could be obtained by using embeddings from models other than BioGPT. For instance, embeddings from BioMegatron appears to help notably more as opposed to PubMedBERT and BioGPT for motor neuron disease related diagnoses (G20∼G259). We hypothesize that each LM might hold biological information that varies slightly, to the extent that some models may enable a greater understanding of subsets of diseases where others are lacking. Finally, we note that using BERT, regardless of embeddings, remains a notably better option than the baseline approaches of using logistic regression and XGBoost for the given benchmarks. And while Gatortron is evaluated with some levels of skepticism due to potential data leaks, we observe that our knowledge embeddings still outperform Gatortron across most tasks and evaluation metrics.

Ablation Experiments

We performed ablation experiments to ascertain the degree of contribution resulting from the knowledge-based attention module. Given an experimental run which obtained the most number of high scores overall (BioGPT embeddings), we reran the same configuration with randomly generated embeddings for all diagnosis codes (RandEmb), skipped the pretraining stage (NoPretrain), and removed the embedding projection layer (NoEmbProj). The resulting accuracies obtained for the same benchmark is shown Table 3.

We observed that random embeddings (RandEmb) could not lead to higher downstream accuracies (simply relying on the small architectural change), and in some cases lead to lower accuracies than training a standard MedBERT (eg. Mortality task). This implied that the modified architecture indeed has the potential of influencing the learning process and that the quality of embeddings are of importance. When skipping pretraining (NoPretrain), which implies only finetuning from scratch, we observed that the downstream accuracies were strictly less ideal than keeping the pretraining step. The experiment highlighted the importance of pretraining, and that the LM embeddings alone were not enough to overcome learning the downstream tasks from scratch.

Finally, we highlight the results of skipping the shared linear projection layer (NoEmbProj), where we observed that scores were higher for a subset of benchmarks in comparison to including the layer. We first note that the projection mechanism is necessarily an important factor in the improved scores (and its modifications could lead to further improvements), as it solely affects the presence of the predetermined embeddings in attention operations. For the purpose of the study, we justified the use of the shared layer to allow the model to flexibly scale the embedding dimensions as its ideal scaling may differ depending on the source LM. Despite the flexibility, we note the possibility of the projection to dilute the embeddings and become susceptible to overfitting.

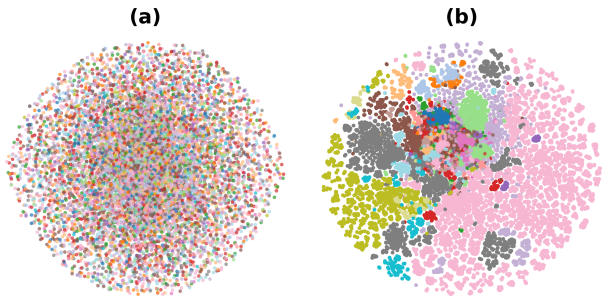

Figure 4: Comparison between tSNE of the diagnosis code embeddings (a) learned from scratch by a BERT model and (b) embeddings obtained from BioGPT of the same codes. Codes are highlighted according to their chapter (first digit).

Improved Latent Space

To qualitatively inspect of the space of embeddings learned from scratch in comparison to the knowledge embeddings introduced from BioGPT, we visualize the tSNE of both embeddings in Figure 4. We noticed that there does appear to be a noticeable difference in the 2D disease latent space where individual colors represent different disease categories (e.g Musculoskeletal system, Nervous system disease, etc.) or ICD chapters. From an explainability perspective, this provides large evidence that the knowledge embeddings due have an influence over what the language model is learning in its latent space.

Discussion

This study proposes a framework for leveraging knowledge embeddings from pre-trained biomedical language models (LMs) within the attention mechanism of a BERT model to enhance clinical outcome predictions from patient records. Our findings indicate that incorporating embeddings from domain-specific LMs (such as BioMegatron, PubMedBERT, and BioGPT) consistently improves the model's performance across multiple prediction tasks, including mortality, length of stay, and the onset of specific diseases. The results suggest that embeddings trained on biomedical data are more effective than general-purpose embeddings in capturing the complexities and nuances of medical data. The clinical implications of this work are that we can leverage well known biomedical facts and descriptions to help these AI systems learn complex clinical concepts more effectively.

Our ablation experiments reinforce the value of these embeddings, demonstrating that both pre-training on domain-specific data and the use of a knowledge-based attention mechanism are essential for achieving improved accuracy. Specifically, replacing domain-specific embeddings with random embeddings or removing pre-training significantly reduced model performance, underscoring the critical role of prior knowledge in the latent representation of disease codes. Additionally, removing the projection layer for embeddings highlighted the influence of this layer on embedding transferability, with evidence suggesting that it provides flexible scaling but may occasionally lead to overfitting.

Future Directions and limitations

While our approach demonstrates strong performance improvements, it opens up several avenues for further research. First, an expansion of this framework to incorporate additional clinical information, such as prescription and lab data from sources like the MIMIC dataset, could offer a more holistic view of a patient's medical profile and improve predictive capabilities. Furthermore, given the recent advances in generative transformer models, exploring a generative approach to pre-training may yield even better representations, albeit with greater computational requirements.

Our ablation studies revealed that the projection layer used in the knowledge-based attention module plays a significant role in modulating the effect of embeddings. Further research should investigate optimal projection mechanisms that minimize information loss while retaining relevant knowledge across multiple biomedical domains. Additionally, while our study examined embeddings from various biomedical LMs, an interesting direction would be to develop a hybrid model that selectively incorporates embeddings from multiple sources, choosing those that best capture the nuances of specific disease categories.

A limitation of this work lies in the variability of embedding performance across tasks. While BioGPT embeddings provided the highest predictive performance overall, embeddings from other LMs, such as BioMegatron, performed better for specific diseases, like motor neuron disorders. This suggests that each LM may contain unique information relevant to particular conditions. Future studies could explore methods for dynamically selecting embeddings based on the disease context.

Data and Code Availability Code to replicate our work can be found in an anonymized repository [5]. Our work explores the MIMIC-IV[6] dataset which can be accessed publicly after undergoing an approval process.

Institutional Review Board (IRB) Our work did not require IRB approval.

References

Acheampong, F. A.; Nunoo-Mensah, H.; and Chen, W. 2021. Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review*, 54(8): 5789–5829.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alsentzer, E.; Murphy, J. R.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; and McDermott, M. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

[5] https://anonymous.4open.science/r/icdbert-F270

[6] https://physionet.org/content/mimiciv/3.0/

Antal, M.; Marosi, M.; Nagy, T.; Juhász, G.; and Antal, P. 2024. M-BioBERTa: Modular RoBERTa-based Model for Biobank-scale Unified Representations.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620. Hong Kong, China: Association for Computational Linguistics.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Chen, C.; Yu, J.; Chen, S.; Liu, C.; Wan, Z.; Bitterman, D. S.; Wang, F.; and Shu, K. ???? ClinicalBench: CAN LLMS BEAT TRADITIONAL ML MODELS IN CLINICAL PREDICTION?

Choi, E.; Bahadori, M. T.; Song, L.; Stewart, W. F.; and Sun, J. 2017. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 787–795.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Ellershaw, S.; Tomlinson, C.; Burton, O. E.; Frost, T.; Hanrahan, J. G.; Khan, D. Z.; Horsfall, H. L.; Little, M.; Malgapo, E.; Starup-Hansen, J.; et al. 2024. Automated Generation of Hospital Discharge Summaries Using Clinical Guidelines and Large Language Models. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Gao, J.; Yang, C.; Heintz, J.; Barrows, S.; Albers, E.; Stapel, M.; Warfield, S.; Cross, A.; and Sun, J. 2022. MedML: fusing medical knowledge and machine learning models for early pediatric COVID-19 hospitalization and severity prediction. *Iscience*, 25(9).

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23.

Gupta, M.; Gallamoza, B.; Cutrona, N.; Dhakal, P.; Poulain, R.; and Beheshti, R. 2022. An extensive data processing pipeline for mimic-iv. In *Machine Learning for Health*, 311–325. PMLR.

Hager, P.; Jungmann, F.; Holland, R.; Bhagat, K.; Hubrecht, I.; Knauer, M.; Vielhauer, J.; Makowski, M.; Braren, R.; Kaissis, G.; et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 1–10.

Hager, P.; Jungmann, F.; and Rueckert, D. ???? MIMIC-IV-Ext Clinical Decision Making: A MIMIC-IV Derived Dataset for Evaluation of Large Language Models on the Task of Clinical Decision Making for Abdominal Pathologies.

Harutyunyan, H.; Khachatrian, H.; Kale, D. C.; Ver Steeg, G.; and Galstyan, A. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1): 96.

Huang, C.-W.; Tsai, S.-C.; and Chen, Y.-N. 2022. PLM-ICD: Automatic ICD Coding with Pretrained Language Models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, 10–20. Seattle, WA: Association for Computational Linguistics.

Huang, K.; Altosaar, J.; and Ranganath, R. 2019. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv:1904.05342*.

Johnson, A.; Bulgarelli, L.; Pollard, T.; Gow, B.; Moody, B.; Horng, S.; Celi, L. A.; and Mark, R. 2024. MIMIC-IV (version 3.0). *PhysioNet*.

Johnson, A. E.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T. J.; Hao, S.; Moody, B.; Gow, B.; et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1.

Kane, M. J.; King, C.; Esserman, D.; Latham, N. K.; Greene, E. J.; and Ganz, D. A. 2023. A compressed large language model embedding dataset of ICD 10 CM descriptions. *BMC bioinformatics*, 24(1): 482.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Lee, S. A.; Brokowski, T.; and Chiang, J. N. 2024. Enhancing Antibiotic Stewardship using a Natural Language Approach for Better Feature Representation. *arXiv preprint arXiv:2405.20419*.

Lee, S. A.; Halperin, H.; Halperin, Y.; Brokowski, T.; and Chiang, J. N. 2025. Using Foundation Models to Prescribe Patients Proper Antibiotics.

Lee, S. A.; Jain, S.; Chen, A.; Biswas, A.; Fang, J.; Rudas, A.; and Chiang, J. N. 2024. Multimodal clinical pseudo-notes for emergency department prediction tasks using multiple embedding model for ehr (meme).

Lee, S. A.; Lee, J.; and Chiang, J. N. 2024. FEET: A Framework for Evaluating Embedding Techniques. *arXiv preprint arXiv:2411.01322*.

Lee, S. A.; and Lindsey, T. 2024. Do Large Language Models understand Medical Codes? *arXiv preprint arXiv:2403.10822*.

Li, L.; Zhou, J.; Gao, Z.; Hua, W.; Fan, L.; Yu, H.; Hagen, L.; Zhang, Y.; Assimes, T. L.; Hemphill, L.; et al. 2024. A scoping review of using Large Language Models (LLMs) to investigate Electronic Health Records (EHRs). *arXiv preprint arXiv:2405.03066*.

Li, Y.; Rao, S.; Solares, J. R. A.; Hassaine, A.; Ramakrishnan, R.; Canoy, D.; Zhu, Y.; Rahimi, K.; and Salimi-Khorshidi, G. 2020. BEHRT: transformer for electronic health records. *Scientific reports*, 10(1): 7155.

Loshchilov, I. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; and Liu, T.-Y. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6): bbac409.

Ma, F.; You, Q.; Xiao, H.; Chitta, R.; Zhou, J.; and Gao, J. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM international conference on information and knowledge management*, 743–752.

Ma, M. D.; Xiao, Y.; Cuturrufo, A.; Wang, X.; and Wang, W. 2024a. Memorize and Rank: Enabling Large Language Models for Medical Event Prediction. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Ma, M. D.; Ye, C.; Yan, Y.; Wang, X.; Ping, P.; Chang, T.; and Wang, W. 2024b. CliBench: Multifaceted Evaluation of Large Language Models in Clinical Decisions on Diagnoses, Procedures, Lab Tests Orders and Prescriptions.

Ma, X.; Wang, Y.; Chu, X.; Ma, L.; Tang, W.; Zhao, J.; Yuan, Y.; and Wang, G. 2022. Patient health representation learning via correlational sparse prior of medical features. *IEEE Transactions on Knowledge and Data Engineering*, 35(11): 11769–11783.

Ono, K.; and Lee, S. A. 2024. Text Serialization and Their Relationship with the Conventional Paradigms of Tabular Machine Learning. *arXiv preprint arXiv:2406.13846*.

Rasmy, L.; Xiang, Y.; Xie, Z.; Tao, C.; and Zhi, D. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1): 86.

Rupp, M.; Peter, O.; and Pattipaka, T. 2023. Exbehrt: Extended transformer for electronic health records. In *International Workshop on Trustworthy Machine Learning for Healthcare*, 73–84. Springer.

Shao, Y.; Cheng, Y.; Nelson, S. J.; Kokkinos, P.; Zamrini, E. Y.; Ahmed, A.; and Zeng-Treitler, Q. 2023. Hybrid value-aware transformer architecture for joint learning from longitudinal and non-longitudinal clinical data. *Journal of personalized medicine*, 13(7): 1070.

Shin, H.-C.; Zhang, Y.; Bakhturina, E.; Puri, R.; Patwary, M.; Shoeybi, M.; and Mani, R. 2020. BioMegatron: larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4700–4706.

Soroush, A.; Glicksberg, B. S.; Zimlichman, E.; Barash, Y.; Freeman, R.; Charney, A. W.; Nadkarni, G. N.; and Klang, E. 2024. Large language models are poor medical coders—benchmarking of medical code querying. *NEJM AI*, 1(5): AIdbp2300040.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Varshney, D.; Zafar, A.; Behera, N. K.; and Ekbal, A. 2023. Knowledge grounded medical dialogue generation using augmented graphs. *Scientific Reports*, 13(1): 3310.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Vu, T.; Nguyen, D. Q.; and Nguyen, A. 2020. A Label Attention Model for ICD Coding from Clinical Text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 3335–3341. Main track.

Wang, S.; McDermott, M. B.; Chauhan, G.; Ghassemi, M.; Hughes, M. C.; and Naumann, T. 2020. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, 222–235.

Wornow, M.; Thapa, R.; Steinberg, E.; Fries, J.; and Shah, N. 2023. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36: 67125–67137.

Yang, X.; Chen, A.; PourNejatian, N.; Shin, H. C.; Smith, K. E.; Parisien, C.; Compas, C.; Martin, C.; Costa, A. B.; Flores, M. G.; et al. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1): 194.

Yang, Z.; Batra, S. S.; Stremmel, J.; and Halperin, E. 2023a. Surpassing GPT-4 Medical Coding with a Two-Stage Approach. *arXiv preprint arXiv:2311.13735*.

Yang, Z.; Kwon, S.; Yao, Z.; and Yu, H. 2023b. Multi-Label few-shot ICD coding as autoregressive generation with prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 5366–5374.

Yang, Z.; Mitra, A.; Liu, W.; Berlowitz, D.; and Yu, H. 2023c. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature communications*, 14(1): 7857.

Ye, M.; Cui, S.; Wang, Y.; Luo, J.; Xiao, C.; and Ma, F. 2021. Medpath: Augmenting health risk prediction via medical knowledge paths. In *Proceedings of the Web Conference 2021*, 1397–1409.

Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; and Liu, T.-Y. 2021. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34: 28877–28888.

Yuan, Z.; Tan, C.; and Huang, S. 2022. Code Synonyms Do Matter: Multiple Synonyms Matching Network for Automatic ICD Coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 808–814. Dublin, Ireland: Association for Computational Linguistics.

Zhou, H.-Y.; Yu, Y.; Wang, C.; Zhang, S.; Gao, Y.; Pan, J.; Shao, J.; Lu, G.; Zhang, K.; and Li, W. 2023. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nature biomedical engineering*, 7(6): 743–755.