Ped-BERT: Early Detection of Diseases for Pediatric Care

Cornelia Ilin[1, †] and Daniel Phaneuf[2]

[1]University of California, Berkeley
[2]University of Wisconsin, Madison
[‡]Corresponding author: cornelia.ilin@berkeley.edu

Abstract

Artificial intelligence (AI)-based diagnosis systems are particularly relevant in pediatrics given the well-documented impact of early-life health conditions on later-life outcomes. Yet, early identification of diseases for this age group has so far remained uncharacterized, likely because access to relevant health data is severely limited. Thanks to a confidential data use agreement with the California Department of Health Care Access and Information, we are able to develop Ped-BERT: A state-of-the-art deep learning model that accurately predicts the likelihood of 100+ conditions in a pediatric patient's next medical visit. We link mother-specific pre- and postnatal period health information to pediatric patient hospital discharge and emergency room visits. Our data set comprises 513.9K mother-baby pairs and contains medical diagnosis codes as well as temporal and spatial pediatric patient characteristics, such as age and residency zip code at the time of visit. Following the popular bidirectional encoder representations from the transformers (BERT) approach, we pretrain Ped-BERT via the masked language modeling objective to learn embedding features for the diagnosis codes contained in our data. We then continue to fine-tune our model to accurately predict diagnosis outcomes for a pediatric patient's next visit, given the history of previous visits and, optionally, the mother's pre- and postnatal health information. We achieve an area under the receiver operator curve (ROC AUC) of 0.923 and an average precision score (APS) of 0.403. Further, we assess the prediction accuracy of Ped-BERT in identifying a few rare genetic diseases. We also examine its fairness by determining whether prediction errors are evenly distributed across various subgroups of mother-baby demographics and health characteristics, or if certain subgroups exhibit a higher susceptibility to prediction errors.

Introduction

Early identification of diseases is vital for better treatment options, longer survival rates, improved long-term outcomes, and lower hospital utilization costs. In recent years, breakthrough progress in this area was made by leveraging electronic health records (EHR) and advanced deep learning (DL) architectures, such as convolutional neural networks (CNN, e.g., Nguyen et al. (Deepr)[1]), recurrent neural networks (RNN, e.g., Choi et al. (Doctor AI)[2]), long short-term memory networks (LSTM, e.g., Pham et al. (DeepCare)[3]), and an even more powerful architecture called bidirectional encoder representation from transformers (BERT). For instance, Li et al.[4] introduce BEHRT, a BERT-inspired model applied to EHR, capable of predicting the likelihood of more than 300 conditions in one's future medical visit; Shang et al.[5] propose G-BERT, a model that combines the power of graph neural networks (GNN) and BERT for diagnosis prediction and medication recommendation; Rasmy et al.[6] introduce Med-BERT, also a BERT model, to provide pretrained contextualized embeddings run on large-scale structured EHR.

To the best of our knowledge, most advances in this literature (a) rely on EHR representative of the **adult population**;[7,4] (b) need to specify the patient age distribution;[1,8,9,2,10,11,6,5] (c) use models that focus on predicting a limited set of health outcomes;[3,8] (d) focus on improving disease risk assessment performance by accounting only for the timing irregularity between clinical events (e.g., age at the time of visit);[1,2,4] (e) do not report prediction performance on rare diseases, or (f) do not use in-utero health information for diagnosis prediction.

However, computer-aided early detection of diseases holds particular significance in the field of pediatrics. Timely diagnosis and intervention are crucial for enhancing the long-term well-being of children, as highlighted in various studies.[12,13,14] Consequently, we have developed Ped-BERT, an architecture inspired by BERT.[15] Our model accurately predicts over 100 potential diagnoses that a child might face during their upcoming medical appointment. It could serve as a valuable tool for aiding pediatricians in their clinical decision-making processes. Ped-BERT leverages a rich dataset encompassing hospital discharge records and emergency room information for pediatrics, including the patient's age and the residential zip code or county at the time of the visit. Additionally, it

can optionally integrate maternal health data from both pre- and postnatal periods. To the best of our knowledge, our prediction framework, leveraging data that matches mother and baby pairs longitudinally is the first of its kind. Furthermore, this dataset empowers us to explore the model's capability to predict rare genetic diseases and to assess its overall fairness, including an examination of whether prediction errors are evenly distributed across different demographics of mother-baby pairs.

To summarize, we contribute to the literature as follows: first, we use a novel data set that links medical records of mother-baby pairs between 1991-2017 in California; second, we develop Ped-BERT, a DL architecture for early detection of diseases in pediatric patients seeking care in inpatient or emergency settings; third, we leverage both temporal and spatial patient characteristics, such as age and geographical location at the time of visit; fourth, we also report the model's performance in predicting rare genetic diseases, and fifth, we evaluate Ped-BERT's performance with fairness in mind.

Data

This study relies on data from the California Department of Health Care Access and Information (HCAI[16]). Through a confidential data use agreement, we access the universe of births between 1991 and 2012 (Birth data), patient discharge data (PDD), and emergency department visits (EDD) through 2017 from nearly $7,000$ California licensed healthcare facilities.[17] We use this data to pre-train and fine-tune Ped-BERT.

Birth data

We observe over 12M birth records registered in California, including maternal antepartum and postpartum hospital records for the nine months before delivery and one-year post-delivery (Figure 1a, top panel). We filter the data to retain only mother-baby pairs (birth IDs) for which the discharge records link to birth certificate data and the baby's social security number (SSN), if the SSN was assigned either at birth or within their first year of life. After filtering, our birth data

includes 763,895 mother-baby pairs whose medical records can be tracked over time by linkage with the PDD and EDD data via the SSN (Figure 1b, top panel). Among all variables present in the birth data, we retain information on the baby's gender, race, and residency zip code and county at birth. We also include information on the mother's race and education, the month prenatal care began, the number of prenatal visits, and the number of times the mother visited a healthcare facility in an emergency or inpatient setting nine months before and twelve months after birth.

Patient Discharge and Emergency Department Visits

The PDD and EDD datasets consist of over 59M inpatient discharges between 1991 and 2017 and over 81M in emergency visits between 2005 and 2017, respectively (Figure 1a, middle and bottom panels). If the emergency encounter resulted in a same-hospital admission, the inpatient record reflects the emergency encounter, and no separate emergency department visit is recorded.

We subset these data to include only those records for which the patient's SSN has a match in the Birth data (Figure 1b, middle and bottom panels). To improve our machine learning task, we further filter this data to select only those patients whose medical history includes at least three emergency or inpatient stays. After this last filtering, we have nearly 1M inpatient and 2.5M emergency discharge records for 513,963 mother-baby pairs. (Figure 1c, middle, bottom, and top panels). From the PDD and EDD data, we retain information on patient demographic characteristics (including residence zip code and county at the time of visit) and up to three disease codes as listed by the healthcare provider during the encounter. The disease codes in our data are classified using the 9th and 10th revisions of the International Statistical Classification of Diseases and Related Health Problems (ICD-9 and ICD-10, respectively). For ease of analysis and interpretability, we convert ICD-10 to ICD-9 codes using the AtlasCUMC dataset[18,19] and choose to operate at the two-digit sub-chapter level.

Via a random split, we use 70% and 30% of these $513,963$ mother-baby pairs, respectively, for fine-tuning Ped-BERT and for assessing prediction performance in the downstream task of predicting the next medical diagnosis. In the following, we refer to these two data sets simply as 'fine-tuning training set' and 'fine-tuning test set'.

Ped-BERT Pre-training Data

For the pre-training of Ped-BERT, it is important to highlight that our goal is to utilize patient records without matches in the fine-tuning data but with available SSN information that enables us to establish connections across time. This distinction is crucial because the data used for pre-training Ped-BERT should not align with our final prediction task to prevent data leakage.

We begin with the raw dataset comprising over 59M inpatient discharges (PDD data) and over 81M emergency visits (EDD data) (Figure 1a, middle and bottom panels). From this extensive dataset, we retain records of patients with valid SSN. Following this filtering process, we are left with nearly 3.8M inpatient stays and 16.2M emergency visits, corresponding to nearly 5.5M patient IDs (Figure 1d). Subsequently, we exclude all patients whose SSN match the $513,963$ birth IDs described in the previous subsection because we will use this data for fine-tuning Ped-BERT (Figure 1e). Finally, to improve our machine learning task, we further refine the data to include only patients with a minimum of three medical encounters. This step leaves us with approximately 2M inpatient discharges and 10M in emergency room visits, totaling 1,855,013 unique patients for pre-training Ped-BERT (Figure 1f).

Via a random split, we use 80% and 20% of these data, respectively, for pre-training Ped-BERT and testing prediction performance. In the following, we refer to these two data sets simply as 'pre-training training set' and 'pre-training test set'.

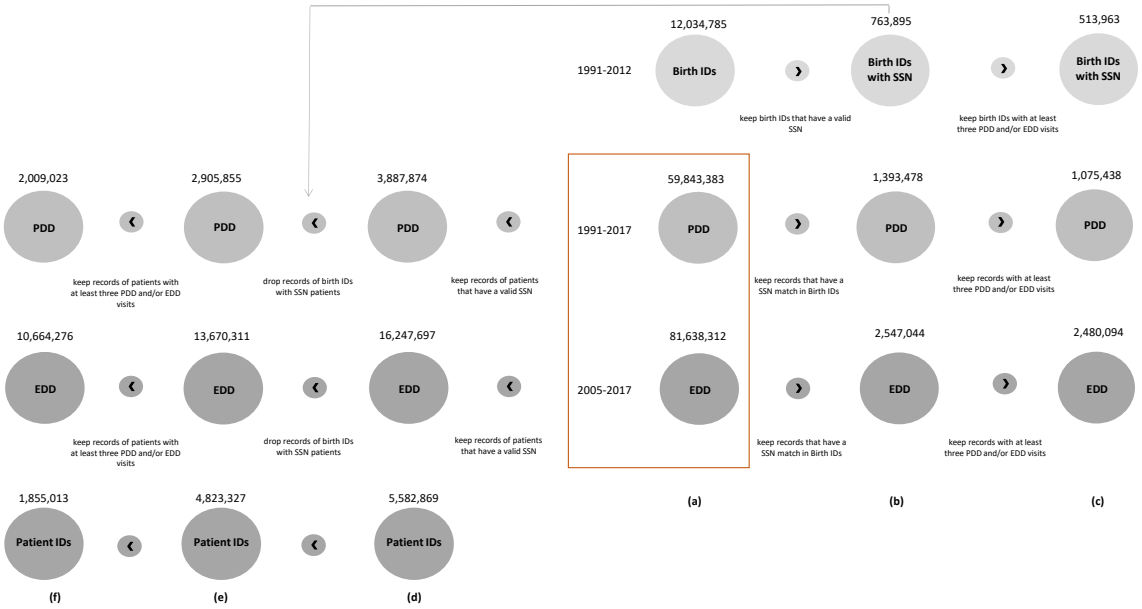

Figure 1: **Filtering, linking, and summary of our data. (a-b)** From the initial set of 12M birth IDs, 59.8M patient discharge data records (PDD), and 81.6M emergency department data records (EDD), we only retain those that can be linked via SSN at birth or in the first year of life: 764K, 1.4M and 2.5M, respectively. **(c)** We further filter by number of inpatient/emergency encounters, only retaining records for patients with at least three medical encounters. This final set consists of approximately 3.5M hospital visits (PDD and EDD combined) between 1991-2017 for $513,963$ mother-baby pairs. This data is used for fine-tuning Ped-BERT. **(d)** From the initial set of 59.8M PDD records and 81.6M EDD records, we only retrain those that can be linked via SSN at some point in life: 2.9M and 13.6M, respectively. **(e-f)** We further drop the records of patients whose SSN has a match in the $513,963$ mother-baby pairs data or have less than three inpatient/emergency encounters. This final set consists of around 2M and 10M records in the PDD and EDD data, respectively, corresponding to $1,855,013$ unique patient IDs. We use this data for pre-training Ped-BERT.

Patient Medical History

For our fine-tuning task of predicting diagnosis in the upcoming medical visit, we rely on patient health information, starting nine months before birth, until data censoring. Let $\mathbf{P}$ represent our sample of patients, and $\mathbf{T}$ represent a set of sorted time stamps. In our data, each patient $p \in \{1, 2, ..., P\}$, is described by a set of birth attributes, $p.A_b = \{A_1, A_2, \ldots, A_n\}$ recorded in the prenatal period and/or at the time of birth. Each pa-

tient is also characterized by a set of inpatient/emergency encounter attributes, $p.A_e = \{(A_1, A_2, \ldots, A_n|1), (A_1, A_2, \ldots, A_n|2), ..., (A_1, A_2, \ldots, A_n|T)\}$ recorded at time $t \in \{1, 2, ...T\}$ of encounter with the medical provider. The attributes in $p.A_b$ cover the baby's gender and race, mother's race and education, pregnancy month prenatal care began, the number of prenatal visits, mother inpatient/emergency visits nine months before and twelve months after birth, and residency zip code/county at birth. Similarly, the attributes in $p.A_e$ are sequences of patient disease codes, patient age, and patient residency zip code/county at the time of visit. Figure 2a illustrates, in tabular form, the medical history of a hypothetical patient with birth attributes (data column 2) $p.A_b$ = {female, hisp, hisp, < high school, 2, 9, 1, 3, 94002} and medical encounter attributes (data columns 3-7) $p.A_e$ = { ([D1, D2], 0, 94002 | visit=1), ([D1], 4, 94002 | visit=2), ..., ([D1], 7, 91000 | visit = 5)}. The diagnosis codes assigned by medical personnel are represented as D1, D2,... etc.

Descriptive statistics of the data utilized for fine-tuning Ped-BERT are presented in Figure 2. The distribution of the baby/patient's race is approximately even between males and females; both the baby/patient's and mother's race are predominantly white or Hispanic/other; most mothers have attained an educational level below high school or have completed college; prenatal care typically starts within 1-3 months of conception, with most mothers receiving 10-12 prenatal care visits; a majority of mothers in our data did not require inpatient or emergency room services in the prepartum and postpartum period (see Figure 2b for additional details). Lastly, mother-baby pairs in our data are well-distributed across California (Figure 2c).

Figure 2: **Patient medical history and descriptive statistics.** **(a)** Example, in tabular form, of a patient's medical history documenting data collected in the in-utero period or at the time of birth, and during the first five inpatient/emergency visits. **(b-c)** Summary statistics for mother-baby/patient demographics and health-related outcomes belonging to the 513,963 mother-baby pairs used for fine-tuning of Ped-BERT. **Abbreviations**: F = Female, M = Male, AS_PI = Assian_Pacific Islander, Bl = Black, Hisp_Oth = Hispanic_Other, NAm_EA = Native, Am_Eskimo_Aleut, Wh = White, <HS = less than High-school, grad = graduate education, b. = before, a. = after, unkn = unknown.

Methods

This study aims to introduce Ped-BERT, a BERT transformer-encoder-based architecture.[20,15] Ped-BERT consists of a bidirectional training procedure and masked language modeling approach (MLM), which enable the model to learn the probability distribution of different diagnosis outcomes in a pediatric patient's next inpatient or emergency visit. We describe our methodology below.

Models

We decompose our prediction task into two components. In the first step, we pre-train our Ped-BERT model using each patient's health attributes data, $p.A_e$, and BERT's MLM approach. The objective here is to learn good disease representations. Afterward, via the second step, we fine-tune Ped-BERT's parameters in a supervised fashion via the downstream task of predicting the diagnosis in the next medical visit.

Ped-BERT Pre-training

The pre-training stage is concerned with learning good disease embeddings. Concretely, Ped-BERT pretrains bidirectional diagnosis representations from medical histories by jointly conditioning both left and right diseases in a pediatric patient's medical history. This approach has been shown to outperform other deep learning architectures, such as CNN, RNN, and LSTM,[1,2,7,3] or left-to-right attention as presented in the original transformer architecture.[20] In addition, Ped-BERT is pre-trained using the MLM approach, whose objective is to randomly replace a fraction of the diagnosis codes with mask tokens [MASK] and task the model with predicting these hidden disease codes instead.

This stage relies on the unlabeled pre-training data split into 'pre-training training set' and 'pre-training test set', and for simplicity, Figure 3a illustrates the pretraining task of Ped-BERT using as an example the hypothetical patient introduced earlier (see Figure 2a). First, the model is given the patient's health history in the following format: **[CLS]** D_1 D_2 **[SEP]** D_1 **[SEP]** D_1 D_2 **[SEP]** D_1 **[SEP]** D_1 **[SEP]**. Here, [CLS] is a token denoting the beginning of the patient's medical

history, and the [SEP] token is added to indicate the end of a medical visit. Both tokens, [CLS] and [SEP], are added to aid with the subsequent diagnosis prediction task. The D tokens represent up to three medical diagnoses at the time of visit (see Figure 3a - Patient Diagnosis History)

Second, the data undergoes pre-processing for the MLM task, involving the random selection of 15 percent of the disease tokens for masking (see Figure 3a - Masking). The selection/masking process follows the original BERT model.[15]

Third, a trainable input embedding matrix is created. We first identify the unique diagnosis codes in the masked training data, map them to integer values, and then encode each patient's diagnosis history using this mapping. Since our disease sequences have different lengths, we use zero padding as a placeholder for adjusting sequence length. We continue by encoding information on visit position, patient's age and geographical location to give our model a sense of the timing, age, and location of events. While age embeddings have been used before (e.g., BEHRT[4]), the geographical location is unique to Ped-BERT. We hypothesize that one's location could be an essential determinant of health outcomes due to the environmental impacts of the quality of local resources, such as clean air and safe water, for example. These resources are prerequisites for health, and poor attributes can be particularly detrimental to vulnerable populations such as the very young. We pre-train Ped-BERT using the 'pre-training training set' with different input embedding specifications. We define our baseline specification as the sum of diagnosis and positional embeddings. We then assess for any MLM prediction performance improvement by adding age and location embeddings (see Figure 3a - Embeddings).

Finally, the output of the input embeddings sublayer is sent to multi-head attention and feed-forward network sublayers (see Figure 3a - transformer-encoder stack). The multi-head attention sublayer is followed by post-layer dropout and normalization. The output is passed to the fully connected feedforward network sublayer and followed by post-layer normalization. This last layer produces the logits for each token in the diagnosis vocabulary. The predicted masked token is extracted from these logits using a Softmax activation function, which provides a probability distribution over each diagnosis token in the vocabulary (see Figure 3a - MLM Predictions). We keep the 'pre-training test set' for final model evaluation (see Figure 3a - Evaluation).

Ped-BERT Fine-tuning for Diagnosis Prediction

A complete training procedure of Ped-BERT includes fine-tuning the model for specific downstream tasks using labeled data. Our main task in the fine-tuning stage is to predict the probability distribution over a set of diagnosis codes in a pediatric patient's next inpatient or emergency room visit. Figure 3b shows the workflow for applying the pre-trained Ped-BERT to this predictive task.

We start from the labeled fine-tuning data split into 'fine-tuning training set' and 'fine-tuning test set'. For each patient p and each data partition, we randomly choose a visit index v ($2 \leq v < T$) to split their health attributes data, $p.A_e$ into input-output pairs. The input is denoted by $X_{p.A_e} = \{(A_{\text{disease codes}}, A_{\text{age}}, A_{\text{zip}}|1), ..., (A_{\text{disease codes}}, A_{\text{age}}, A_{\text{zip}}|v)\}$ and the output by $y_{p.A_e}$, which is a multi-hot vector of length 105 (corresponding to the total number of disease codes in Ped-BERT's vocabulary) equal to 1 for diagnosis codes that exist in the next visit, $A_{\text{disease codes}}|v+1$. We tokenize and encode the diagnosis history of each patient, and feed the data into Ped-BERT for embeddings extraction (based on the output of the last layer of the transformer-encoder block, see Figure 3b - Preprocessing). We then use the 'fine-tuning training set' to fine-tune all Ped-BERT's learned parameters by fitting and optimizing a multiclass logistic regression model for subsequent diagnosis prediction (see Figure 3b - Learning). We keep the 'fine-tuning test set' until the very end for the final model evaluation (see Figure 3b - Evaluation).

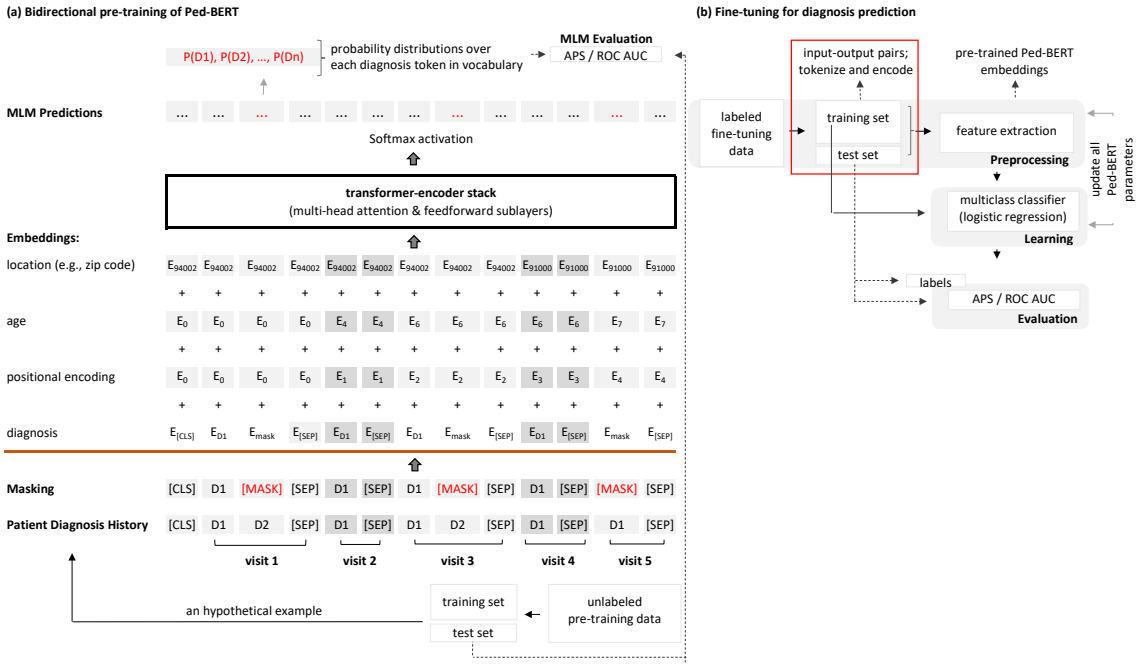

Figure 3: **Ped-BERT architecture. (a)** The pre-training task is explained using as an example the hypothetical patient introduced in Figure 2a: Ped-BERT sees the medical history and masks some of the diagnosis codes before sending them to embedding, multi-head attention, and feed-forward network sublayers. The task here is to predict the [MASK] disease codes. **(b)** In the fine-tuning task, the pre-trained Ped-BERT model parameters are fine-tuned using a logistic model with the objective of predicting the probability distribution over given diagnosis codes in a pediatric patient's next inpatient or emergency room visit. **(a-b)**The fine-tuning and pre-training steps are evaluated using the APS and ROC AUC scores.

Prediction Performance Evaluation

We evaluate the performance of both the pre-trained and fine-tuned Ped-BERT model for disease prediction using two key metrics: the Average Precision Score (APS) and the Area Under the Receiver Operating Curve (ROC AUC). Note that the APS summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight (see scikit-learn[21] for implementation details).

During the pre-training phase, we calculate these metrics by comparing the model's predictions to the actual ground-truth data associated with the [MASK] token for all patients within the "pre-

training test set." In the fine-tuning stage, we represent the model's predictions for each patient p as $y^*_{p.A_e}$ and gauge the model's performance by assessing the agreement between these predictions $(y^*_{p.A_e})$ and the actual values $(y_{p.A_e})$. This assessment is conducted by computing the APS and ROC AUC on the "fine-tuning test set" individually for each patient, and subsequently calculating the averages across all patients for all diagnosis codes, as well as the averages across all patients for each specific diagnosis code.

Results

We present results from Ped-BERT's pre-training stage and then evaluate Ped-BERT's fine-tuned ability to predict the diagnosis in the subsequent medical encounter for all disease codes, and separately, for rare genetic conditions. We conclude by discussing the results of a few fairness tasks and how Ped-BERT could guide medical practitioners.

Ped-BERT Pre-training Evaluation

The optimal architecture of Ped-BERT has the following specifications: the input diagnosis embedding matrix is of size 120 x 128, with the first dimension representing the length of the diagnosis vocabulary (115 unique two-digit diagnosis codes + OOV + $[MASK]$ + $[CLS]$ + $[SEP]$ + padding token) and the second dimension representing the embedding size; the patient history is restricted to a maximum length of 40 tokens; the encoder is a stack of 6 identical layers; inside each of these identical layers there is a multi-head attention sublayer containing 12 heads and a feedforward network sublayer containing 128 hidden units; dropout regularization rate is set to 0.1; pre-training is for 15 epochs using the Adam optimizer with a learning rate of $3e-5$ and a decay of 0.01.

Ped-BERT is pre-trained using different specifications for the input embedding matrix. As mentioned in the Methods section, we define our baseline embeddings specification as the sum of diagnosis embeddings and positional encodings. We then augment this baseline by adding age embeddings (+ age), zip embeddings (+ zip), county embeddings (+ cnty), age + zip embeddings (+ age + zip), and age + county embeddings (+ age + cnty). Figure 4a presents a couple of

interesting findings derived from the 'pre-training test set': adding age embeddings slightly improves the APS score relative to baseline [0.52 vs. 0.51]; adding county embeddings to the baseline + age specification results in negligible APS differences [APS: 0.521 vs. 0.52]; adding additional embeddings (such as age and/or county) to the baseline specification results in negligible differences in terms of ROC AUC. We also assess specifications with the patient's zip code instead of the county given as additional embeddings and find that the model performance is below the base specification in terms of both APS and ROC AUC (results not presented in Figure 4a). In summary, our results suggest that, in the context of pediatric patients, augmenting a pre-trained model with information on the patient's age at the time of medical encounter has a modest positive impact on model performance, while the addition of patient's county of residence at the time of the visit does not improve the results. For more information on the distributional details regarding the data used to pre-train Ped-BERT see Supplementary Figure S1).

We proceed to evaluate the quality of our pre-trained embeddings through both intrinsic and extrinsic methods. Intrinsic assessment involves examining the embeddings' quality through visual inspection and reporting cosine similarity among disease embeddings. For the extrinsic evaluation, we examine the embeddings' effectiveness in predicting patient gender distribution for specific disease codes.

To visually inspect Ped-BERT's embeddings, we reduce the embedding space to 2D using t-SNE (see scikit-learn[22] for implementation details). Figure 4b shows the reduced embeddings for the baseline + age input embeddings specification. The visualization reveals that similar diseases (such as those related to injury and poisoning, diseases of the respiratory system, and birth conditions) cluster together. Furthermore, diseases known to frequently co-occur (such as neoplasms, diseases of the blood, and blood-forming organs) are also grouped closely. Upon closer examination of these 2D disease embedding clusters, a remarkable association with the International Classification of Disease Codes (ICD codes) becomes evident. Notably, this finding is interesting because we did not explicitly provide this information to Ped-BERT during the pre-training phase. Subsequently, we proceed to report the cosine similarity between disease codes using Ped-BERT's learned embeddings. Upon aggregation at the chapter level, we observe a range of similarity values, with the minimum

and maximum values being -0.318 and 1, respectively; the values at the 25, 50, and 95 percentiles, are 0.093, 0.229, and 0.586, respectively (additional details are available in supplementary Figure S2).

Finally, we conduct an extrinsic evaluation of Ped-BERT's embeddings by assessing their performance in predicting the gender distribution of patients with congenital anomalies and tuberculosis. This evaluation is prompted by the increasing body of evidence highlighting sex-specific disparities in the prevalence of congenital anomalies and tuberculosis, with research studies demonstrating higher prevalence rates among pediatric males.[23,24] As shown in Supplementary Figure S3, Ped-BERT consistently predicts a higher prevalence of these two diseases among males when evaluated on the 'pre-training test set', with a Fisher's exact test value equal to 0.0862 ($p < 0.1$).

In summary, our current intrinsic and extrinsic evaluation results indicate that Ped-BERT has developed a substantial understanding of the contextual relationships between diseases.

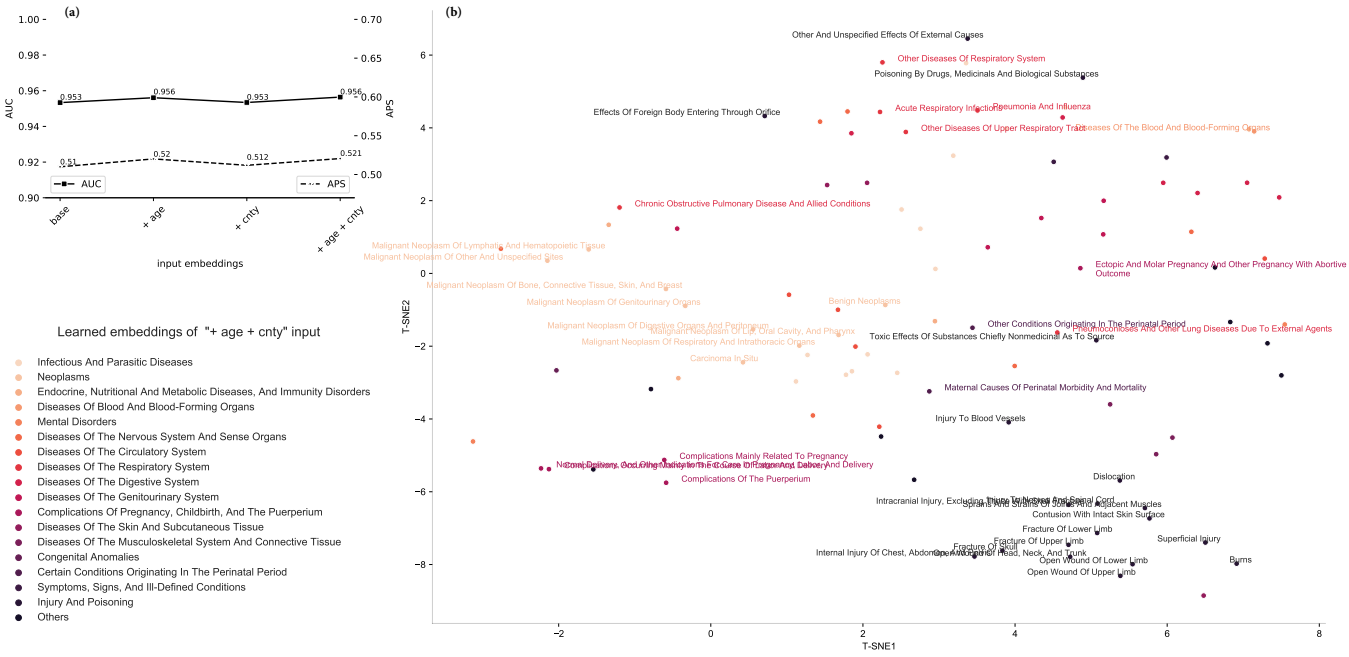

Figure 4: **Evaluation of Ped-BERT's MLM task**. (**a**) The average precision score (APS, right y-axis) and the area under the receiver operating curve (ROC AUC, left y-axis) were computed as sample averages for the following embedding specifications: base (which is the sum of diagnosis embeddings and positional encodings), base + age, base + county, and base + age + county embeddings. These metrics represent comparisons between the ground truth (unmasked tokens) and the MLM-predicted diagnosis (masked tokens) in the test data. (**b**) Intrinsic evaluation of the MLM embeddings via visual inspection for the base + age input embeddings specification. We reduce the dimension of the embedding matrix from 120×128 to 120×2 using t-SNE to create a 2D visualization of all 115 two-digit diagnosis codes in our vocabulary. Colors represent diagnosis chapters.

Ped-BERT Fine-tuning for Diagnosis Prediction

A complete training procedure of Ped-BERT includes fine-tuning. Ped-BERT is not designed for any specific task in the pre-training step but instead trained as a general disease model for pediatric patients. In the fine-tuning stage, we generalize Ped-BERT to predict the medical diagnosis in the subsequent inpatient or emergency pediatric visit. Specifically, we update the pre-trained model parameters for our specific downstream task using regular supervised logistic learning on labeled data by adding on top of the pre-trained Ped-BERT a feedforward layer with 64 hidden units and

an output layer containing a softmax activation function. The model is trained for 100 epochs using the Adam optimizer with a learning rate of $3e-4$ and early stopping. In Figure 5a (black lines), we focus on reporting results related to the base embeddings specification and its corresponding augmentations with age (+ age) and age + county (+ age + cnty). We find no differences in ROC AUC (continous black lines) and very small differences in APS (dashed black lines) across the three pre-trained embedding specifications (e.g., APS base: 0.392, base + age: 0.397, base + age + cnty: 0.399). For comparison, BEHRT's[4] performance in the downstream task of predicting the subsequent diagnosis codes for the adult population, shows a one-point difference in APS and an inisgnificant difference in ROC AUC between the base and base + age embeddings.

In Figure 5b, we report the ROC AUC for each diagnosis code in our 'fine-tuning test set' as derived from the base + age embeddings specification; we highlight the top five (blue colors) and the least five performances (red colors) in terms of AUC scores. Our results indicate that Ped-BERT exhibits high predictive performance for certain conditions, including maternal causes of perinatal morbidity and mortality (AUC = 0.983), malignant neoplasm of genitourinary organs (AUC = 0.950), congenital anomalies (AUC = 0.938), ischemic heart disease (AUC = 0.918), malignant neoplasm of bone (AUC = 0.906), and organic psychotic conditions (AUC = 0.901). On the other hand, it demonstrates lower prediction performance for conditions like injury of nerves of spinal cord (AUC = 0.619), malignant neoplasm of respiratory and intrathoracic organs (AUC = 0.615), toxic effects of substances (AUC = 0.614), persons with potential health hazards related to personal and family history (AUC = 0.531), and other spirochetal diseases (AUC = 0.410).

In Figure 5c, our focus centers on assessing the suitability of Ped-BERT for detecting rare genetic diseases for pediatric patients. To achieve this, we compute and report the ROC AUC scores for various genetic diseases, including other Diseases of the biliary tract (AUC = 0.645), other metabolic and immunity disorders (AUC = 0.598), diseases of white blood cells (AUC = 0.649), cerebral degenerations manifesting in childhood (AUC = 0.656), congenital anomalies of eyes (AUC = 0.895), and diseases of the capillaries (AUC = 0.588). These results indicate varying levels of prediction performance for these rare diseases, ranging from decent to suboptimal. For more details, please refer to Supplementary Table S2, which provides additional information on the

number of patients with these rare diseases in both the 'fine-tuning training set' and 'fine-tuning test set'.

The Role of Ped-BERT Pre-training and Birth Attributes Data

To further explore the efficacy of Ped-BERT's pre-training and the role of birth attributes data, we conduct two additional investigations. First, we compare the performance of a disease prediction model using randomly initialized base embeddings against the three models employing pre-trained Ped-BERT embeddings. In Figure 5a (black lines), we observe a significant enhancement in APS and a modest improvement in ROC AUC performance when comparing the model with randomly initialized base embeddings to the one with pre-trained base Ped-BERT embeddings (APS: 0.372 vs. 0.392, ROC AUC: 0.915 vs. 0.92). Second, we extend the analysis by incorporating birth attributes data ($p.A_b$) into both the randomly initialized and pre-trained Ped-BERT embedding models to assess potential improvements in disease prediction. In Figure 5a (maroon lines), notable distinctions are only evident for the model with randomly initialized base embeddings (APS: 0.372 vs. 0.392, ROC AUC: 0.915 vs. 0.922) and the model with pre-trained base Ped-BERT embeddings (APS: 0.392 vs. 0.403, ROC AUC: 0.92 vs. 0.926). These results suggest that a pre-trained Ped-BERT model with base embeddings is a good subsitute for a model with randomly initialize embeddings that also require birth attributes data for better prediction performance.

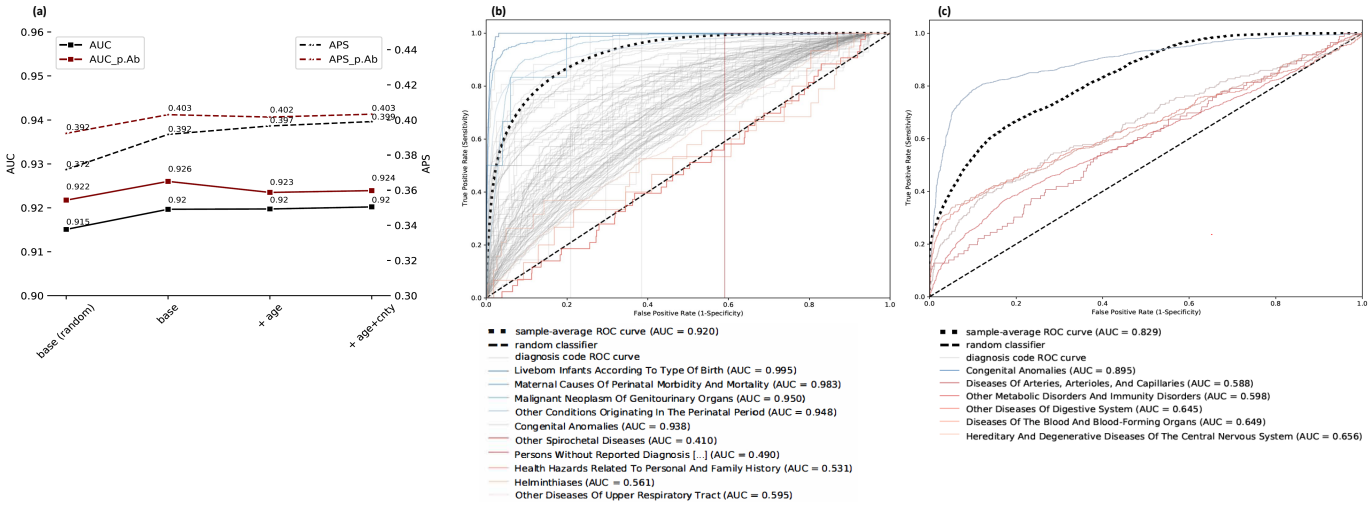

Figure 5: **Evaluation of the Disease Prediction task.** We consider the following embedding specifications: black lines = randomly initialized embeddings (base (random)), pre-trained base Ped-BERT embeddings, pre-trained base + age Ped-BERT embeddings, and pre-trained base + age + county Ped-BERT embeddings; maroon lines = augment the four models with features from the birth attributes data (*p.Ab*). (**a**) The APS (right y-axis) and ROC AUC (left y-axis) computed for each embedding specification scenario outlined above (black lines and maroon lines)). These metrics represent comparisons between the ground truth and the predicted diagnosis for each patient in the output partition of the input-output pairs of our test dataset. (**b**) True Positive Rates and False Positive Rates curves averaged across all patients for each diagnosis in the data (grey lines), for the top five (blue lines) and least five (red lines) diagnosis codes based on AUC scores, and average across all patients for all diagnosis codes in the data (dot-dashed black lines); the long-dashed line denotes a random classifier. (**c**) similar to (**b**) but for selected rare genetic diseases for top one (blue lines) and least five (red lines).

Fairness Tasks

We are interested in determining whether next-visit diagnosis prediction errors are uniform across subgroups in our data. Figure 5 already gives us some insights into the model's APS and ROC AUC performance (overall and by disease code), but it is desirable to understand how well it performs for different subgroups. For example, Figure 2 identifies groups of mother-baby/patient demographics and health-related outcomes belonging to the pairs used in this analysis. Our data also contains information on the mother's country of birth, which is rarely available to research and unique to our study. As such, in this section, we aim to assess the fine-tuned Ped-BERT's

prediction performance with fairness in mind and use the 'fine-tuning test set' and the pre-trained baseline + age embeddings for this task.

We find minimal differences in ROC AUC performance across groups of patient gender and race, mother race and education, month prenatal care began, the number of prenatal visits, and the number of times the mother visited a healthcare facility overnight or in an emergency setting (see Figure 6, top and middle panels). Next, we create bins for the mother's country at her own birth, for similar patient ages, for zip codes/counties belonging to the same geographical region,[25] and for similar PM 2.5 pollution values at the time of birth.[26] We find that Ped-BERT is more susceptible to prediction errors depending on the mother country of origin at her own birth, for patients in the age subgroups 3-17 and greater than 17 ((AUC: 0.933 and 0.901) than those in the 0-2 subgroup (AUC: 0.871), and for patients that have been born in a zip code with unhealthy pollution (AUC: 0.887) as opposed to moderate or good pollution (AUC: 0.907 and 0.914, respectively)(see Figure 6, bottom panel).

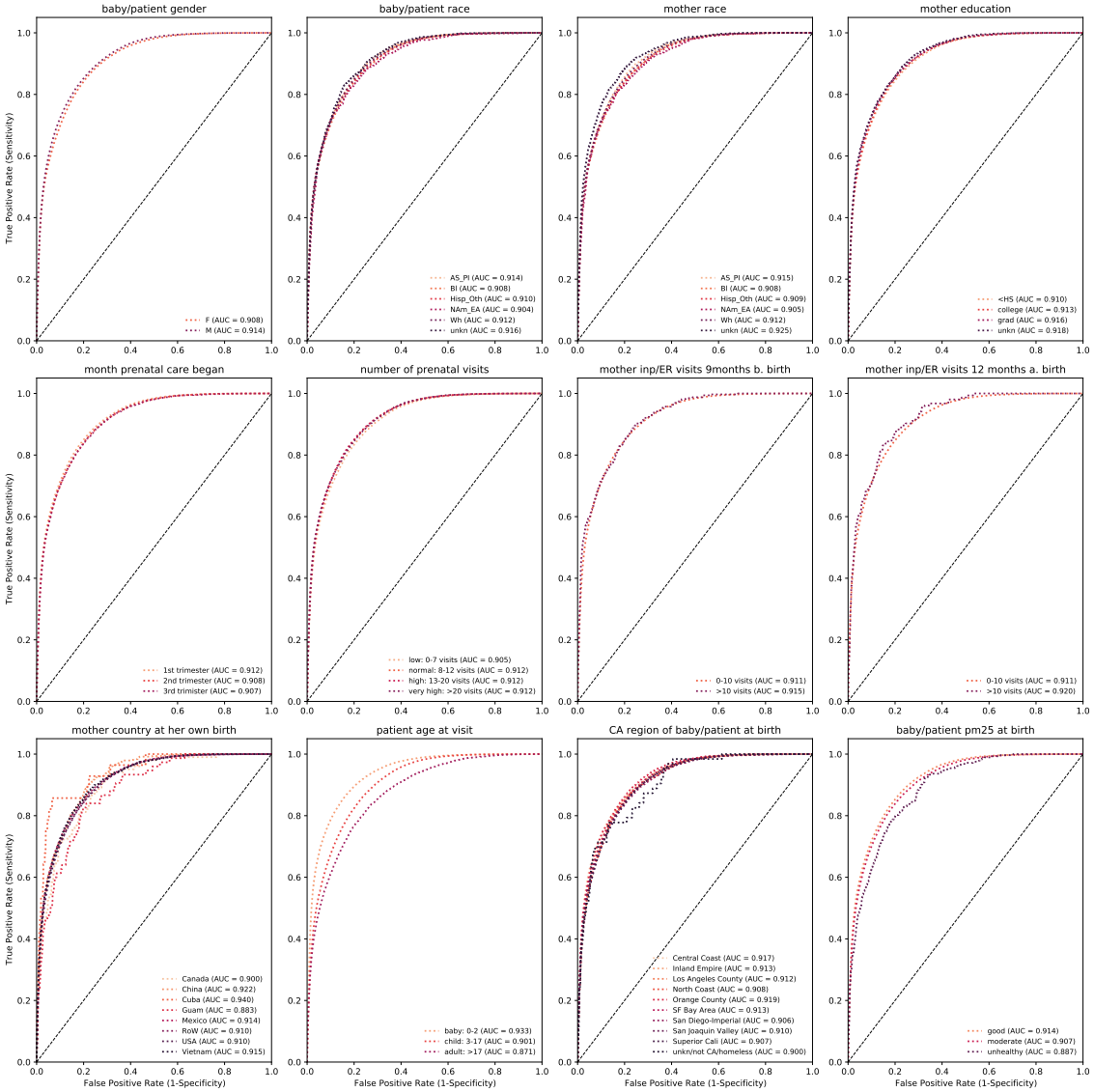

Figure 6: **Fairness tasks**: Using results from the fine-tuning stage, we compare evaluation results across different subgroups (e.g., baby/patient gender and race; mother race and education; month prenatal care began). The evaluation results rely on the fine-tuned model with base + age embeddings applied to the test sample. True Positive Rates (Sensitivity) vs. False Positive Rates (1-Specificity) are shown as red dots lines. A long-dashed line denotes a random classifier.

Model Application

Our results suggest that Ped-BERT may provide useful information to clinical providers. We imagine the approach can be utilized in two ways. First, a medical provider can use the fine-tuned Ped-BERT model with base + age embeddings to augment clinical decision-making with machine learning. Using our model's predictions would reduce uncertainty over the most likely conditions. The table below provides a prediction example for a patient randomly chosen from our 'fine-tuning test set'. The model is presented with the patient's previous two-digit ICD9 health history (along with age information at the time of visit - not presented here for simplicity). The model outputs a probability distribution over all diagnosis codes in Ped-BERT's fine-tuning vocabulary. Listed below are the top five predicted diseases.

User: Medical provider

```
Previous two-digit ICD9 health history
--------------------------------------
[CLS] 46 78 [SEP] 48 [SEP] 49 48 [SEP] 46 [SEP] 07 [SEP]

TOP five diagnosis predictions at the next medical encounter
------------------------------------------------------------
ICD9 code:      49      48       46      78      38
probability:   0.255   0.160    0.153   0.098   0.033

two-digit ICD9 code descriptions
--------------------------------
07: Infectious And Parasitic Diseases
38: Diseases Of The Ear And Mastoid Process
46: Acute Respiratory Infections
48: Pneumonia And Influenza
49: Chronic Obstructive Pulmonary Disease And Allied Conditions
78: Symptoms
```

A second way that a medical professional could use our approach would be to fine-tune the pre-trained Ped-BERT on their own corpus of medical records and make predictions for new patients.

Discussion

This research aims to improve the early detection of diseases in pediatric patients using a unique healthcare database and the latest developments in bidirectional encoder representations from trans-

formers (BERT). The data used in our analysis consists of vital statistics and birth information, as well as hospital discharge data and emergency room visits in California between 1991 and 2017 for 513,963 mother-baby pairs. A BERT-based model called Ped-BERT is trained using a masked language model (MLM) approach and is able to accurately predict the likelihood of over 100 conditions in a child's next medical visit. The study also evaluates Ped-BERT's prediction performance for rare genetic disorders, and for fairness by assessing whether prediction errors are uniformly distributed across different mother-baby demographics and health characteristics subgroups. The model has the potential to assist clinical providers in making machine learning-augmented decisions about pediatric healthcare.

The pre-training stage of Ped-BERT involves learning good representations of diseases by testing different combinations of input embeddings to represent a patient's health history. The baseline specification is the sum of diagnosis embeddings and positional encodings. Age and zop/county embeddings augment this baseline in our performance improvement tests. We find that adding age embeddings improves the APS score relative to baseline, and further expanding with county embeddings results in negligible APS differences relative to the baseline + age specification. We use intrinsic and extrinsic methods to evaluate the embedding quality further. Intrinsically, we find that the model has learned to cluster together diseases that belong to the same ICD chapter or are known to co-occur. Extrinsically, we find that the disease embeddings generated by Ped-BERT correctly predict the male-skewed gender distributions for congenital anomalies and tuberculosis.

The fine-tuning stage of the Ped-BERT model involves adapting the model for the specific downstream task of predicting the diagnosis in the subsequent inpatient or emergency pediatric visit. The results averaged across all patients and disease codes show insignificant differences in ROC AUC and minor differences in APS across the baseline, baseline + age, and baseline + age + cnty fine-tuned embedding specifications. The ROC AUC sample averages were also computed across all patients and (a) each diagnosis code, highlighting the top five and least five performances, and (b) six rare genetic diseases.

Finally, we assess the fine-tuned Ped-BERT for fairness, as models that perform poorly on certain subgroups can lead to unequal outcomes and perpetuate biases. In this case, Ped-BERT

performs generally well, with some differences based on the mother's country at her own birth, the patient's age, and patient's pollution exposure at birth.

We propose several possible directions for future research based on the architecture and properties of Ped-BERT. Its ability to encode diagnosis codes, age, and geographical location into a fixed-length vector representation can make it useful for many tasks. For example, one can focus on fine-tuning Ped-BERT for early detection of rare genetic pediatric conditions. It is also worth noting here that the specific dataset used can significantly impact the model's performance. For example, Ped-BERT was pre-trained on a dataset of medical records from California between 1991-2017, so it may not perform as well on tasks that involve other states in the US or other countries. Another possibility is increasing the training size for older patients and those living in less environmentally friendly areas. The rationale here is that a more diverse training set will expose the model to a broader range of ages and geographical locations by making the location embeddings more powerful for learning good disease representations while helping the model generalize better to new tasks.

Data Availability

Data used in this study can be divided into three categories - Health, Geospatial, and PM2.5 pollution. They are restricted access or publicly available at different locations.

We collected health data from The California Department of Health Care Access and Information (HCAI[16]), which provides confidential patient-level data sets to researchers eligible through the Information Practices Act (CA Civil Code Section 1798 et seq.). The geospatial data comes from the Census Bureau and includes 2010 ZCTA shapefiles,[27] 2010 county shapefiles,[28] 2010 ZCTA to county codes,[29] ZCTA to zip codes crosswalks, as well as the 2020 geographical division of California's 58 counties into ten regions.[25] The PM 2.5 pollution data was collected and made available by the Atmospheric Composition Analysis Group[30] of the Washington University of St. Luis. It provides information on concentrations of ambient fine particulate matter across North America, which combines data from chemical transport modeling, satellite remote sensing, and ground-based monitoring.

References

[1] Nguyen, P., Tran, T., Wickramasinghe, N. & Venkatesh, S. \mathtt {Deepr}: a convolutional net for medical records. IEEE journal of biomedical and health informatics **21**, 22–30 (2016).

[2] Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F. & Sun, J. Doctor ai: Predicting clinical events via recurrent neural networks. In Machine learning for healthcare conference, 301–318 (PMLR, 2016).

[3] Pham, T., Tran, T., Phung, D. & Venkatesh, S. Deepcare: A deep dynamic memory model for predictive medicine. In Pacific-Asia conference on knowledge discovery and data mining, 30–41 (Springer, 2016).

[4] Li, Y. et al. Behrt: transformer for electronic health records. Scientific reports **10**, 1–12 (2020).

[5] Shang, J., Ma, T., Xiao, C. & Sun, J. Pre-training of graph augmented transformers for medication recommendation. arXiv preprint arXiv:1906.00346 (2019).

[6] Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine **4**, 1–13 (2021).

[7] Choi, E. et al. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. Advances in neural information processing systems **29** (2016).

[8] Liang, Z., Zhang, G., Huang, J. X. & Hu, Q. V. Deep learning for healthcare decision making with emrs. In 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 556–559 (IEEE, 2014).

[9] Wickramasinghe, N. A convolutional net for medical records. (Engineering in Medicine and Biology Society, 2017).

[10] Lauritsen, S. M. et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nature communications **11**, 1–11 (2020).

[11] Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. NPJ digital medicine **1**, 1–10 (2018).

[12] Osmond, C. & Barker, D. Fetal, infant, and childhood growth are predictors of coronary heart disease, diabetes, and hypertension in adult men and women. Environmental health perspectives **108**, 545–553 (2000).

[13] Monteiro, P. O. A. & Victora, C. G. Rapid growth in infancy and childhood and obesity in later life–a systematic review. Obesity reviews **6**, 143–154 (2005).

[14] Yoshida-Montezuma, Y. et al. The association between late preterm birth and cardiometabolic conditions across the life course: A systematic review and meta-analysis. Paediatric and Perinatal Epidemiology **36**, 264–275 (2022).

[15] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

[16] California Department of Health Care Access and Information (HCAI) (website accessed on September 2022). URL https://hcai.ca.gov/data-and-reports/research-data-request-information/.

[17] California Department of Health Care Access and Information (website accessed on September 2022). URL https://data.chhs.ca.gov/dataset/licensed-healthcare-facility-listing/resource/eff78ca9-5595-4c3a-880d-3488f129329c?inner_span=True.

[18] AtlasCUMC (website accessed on September 2022). URL https://github.com/AtlasCUMC/ICD10-ICD9-codes-conversion.

[19] Centers for Medicare Medicaid Services (website accessed on September 2022). URL https://www.cms.gov/Medicare/Coding/ICD10/2018-ICD-10-CM-and-GEMs.

[20] Vaswani, A. et al. Attention is all you need. Advances in neural information processing systems **30** (2017).

[21] Scikit-learn (website accessed on December 2022). URL https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html/.

[22] t-SNE (website accessed on December 2022). URL https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html.

[23] Center for Disease Prevention and Control (website accessed on December 2022). URL https://www.cdc.gov/tb/statistics/reports/2020/table21.htm.

[24] Black, A. J., Lu, D. Y., Yefet, L. S. & Baird, R. Sex differences in surgically correctable congenital anomalies: a systematic review. Journal of Pediatric Surgery **55**, 811–820 (2020).

[25] California Census, 2020 Office (website accessed on December 2022). URL https://census.ca.gov/regions/.

[26] Environmental Protection Agency (EPA): the National Ambient Air Quality Standards for Particle Pollution (website accessed on December 2022). URL https://www.epa.gov/sites/default/files/2016-04/documents/2012_aqi_factsheet.pdf.

[27] Census Bureau: Geometry of ZCTA codes in California (2010 boundaries) (website accessed on February 2020). URL https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=ZIP+Code+Tabulation+Areas.

[28] Census Bureau: Geometry of County codes in California (2010 boundaries) (website accessed on February 2020). URL https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=Counties+%28and+equivalent%29.

[29] Census Bureau: ZCTA to county codes in California (2010 boundaries) (website accessed on February 2020). URL https://www.census.gov/programs-surveys/geography/technical-documentation/records-layout/2010-zcta-record-layout.html.

[30] Meng, J. et al. Estimated long-term (1981–2016) concentrations of ambient fine particulate matter across north america from chemical transport modeling, satellite remote sensing, and ground-based measurements. Environmental science & technology **53**, 5071–5079 (2019).

[31] California Department of Health Care Access and Information (HCAI): California Inpatient Data Reporting Manual (website accessed on September 2022). URL https://hcai.ca.gov/data-and-reports/submit-data/patient-data/inpatient-reporting/.

Role of funding source: The authors declare no funding source.

Declaration of interests: The authors declare no competing interests.

Code availability: The underlying code for this study is publicly available at `https://github.com/corneliailin/CA_hospitals`.

Appendices

A Data Acquisition and Processing

A.1 Health Data

Our data is distributed by the California Department of Health Care Access and Information (HCAI[16]). We requested the following files for research purposes:

Linked Birth Files (Birth data): a research database created to study delivery and birth outcomes. It includes maternal antepartum and postpartum hospital records for the nine months before delivery and one-year post-delivery. In addition, the linked file contains birth records and all infant readmissions occurring within the first year of life. The file contains all infants born in a given year, including births that happened in a California hospital that reported to HCAI, births that occurred in a California hospital that did not report to HCAI, and births that occurred outside California. It includes all infants and mothers, irrespective of whether they were linked to a birth record. The linked pairs of birth/delivery records have information associated with a mother/baby pair from the baby's discharge data record, the mother's discharge data record, and the birth certificate data. Linked birth files are available beginning with the 1991 calendar year reporting period (HCAI[16]).

The Patient Discharge Dataset (PDD): consists of a record for each inpatient discharge from a California-licensed hospital. Licensed hospitals include general acute care, acute psychiatric, chemical dependency recovery, and psychiatric health facilities. These datasets are available starting in 1983 (HCAI[16]). For more information on the data and reporting requirements, see the California Inpatient Data Reporting Manual.[31]

The Emergency Department Dataset (EDD): includes information from hospitals licensed to provide emergency medical services. The EDD encounters include those patients who had face-to-face contact with the provider. If the patient left without being seen, the patient would not have had a face-to-face encounter with a provider, and therefore the EDD encounter would not be reported. These data sets are available beginning January 2005 (HCAI[16]).

Our study's primary variable of interest is the primary, secondary, and tertiary ICD 9 or ICD10 diagnosis codes at the time of visits. We accessed it along with other relevant metadata, such as mother-baby demographics and mother health-related outcomes nine months before and 12 months after birth.

A.2 Geospatial Data

The geospatial data was constructed and made available by the Census Bureau. For California, the relevant 2010 ZCTA and county-specific shapefiles,[27, 28] and the 2010 ZCTA to county codes[29] were identified and mapped to our health data for visualization and analysis purposes. For the Fairness analysis, we extracted the California - Census 2020 geographical division of counties into regions. Table S1 contains a summary of the counties used for each region.

Table S1: **Geographical division of California's counties** - Details of counties included in each California region to support the Fairness analysis presented in Figure 6.

Region	County
Central Coast	Monterey, San Benito, San Luis Obispo, Santa Barbara, Santa Cruz, Ventura
Inland Empire	Riverside, San Bernardino
Los Angeles County	Los Angeles
North Coast	Del Norte, Humboldt, Lake, Mendocino, Napa, Sonoma, Trinity
Orange County	Orange
SF Bay Area	Alameda, Contra Costa, Marin, San Francisco, San Mateo, Santa Clara, Solano
San Diego - Imperial	Imperial, San Diego
San Joaquin Valley	Alpine, Amador, Calaveras, Madera, Mariposa, Merced, Mono, San Joaquin, Stanislaus, Tuolumne, Fresno, Inyo, Kern, Kings, Tulare
Superior Cali	Butte, Colusa, El Dorado, Glenn, Lassen, Modoc, Nevada, Placer, Plumas, Sacramento, Shasta, Sierra, Siskiyou, Sutter, Tehama, Yolo, Yuba

A.3 PM2.5 Pollution Data

PM2.5 pollution data comes from the Atmospheric Composition Analysis Group[30] of the Washington University of St. Luis. According to the source, this data is the estimated concentrations of ambient fine particulate matter across North America, which combines information from chemical transport modeling, satellite remote sensing, and ground-based monitoring. The estimates included information from updated historical emissions inventories and meteorological data, fine resolution satellite-based estimates of PM2.5, and ground-based measurements of PM2.5, PM10, and total sus-

pended particles (TSP) measurements. We extracted information at the monthly level for each zip code in our data to construct groups according to the EPA definition for healthy (0.0–12.0$\mu g/m^3$), moderate (12.1–35$\mu g/m^3$), and unhealthy ($> 35\mu g/m^3$) PM2.5 pollution exposure at birth.

B Supplementary Figures

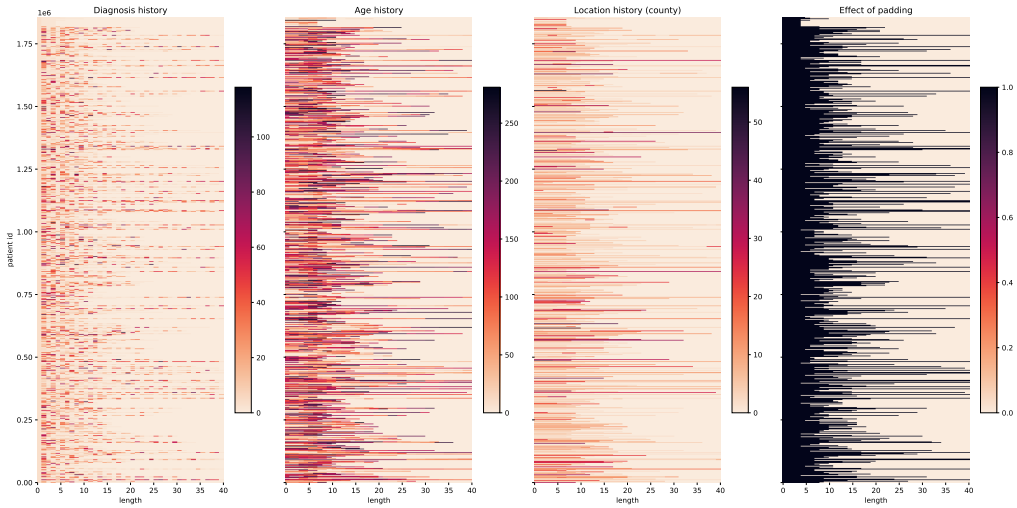

Figure S1: **Summary statistics of encoded input for pre-training Ped-BERT**. The x-axis represents the length of a given patient history, which we optimally set to 40 periods. Each tick on the y-axis represents a diagnosis, age, location history, and padding summary for a given patient ID in the pre-training data. Heatmap values and colors represent: for **(a-c)**, the encoded disease codes, age, and location history; for **(b)**: the effect of zero padding since not all patients have a history length equal to 40.

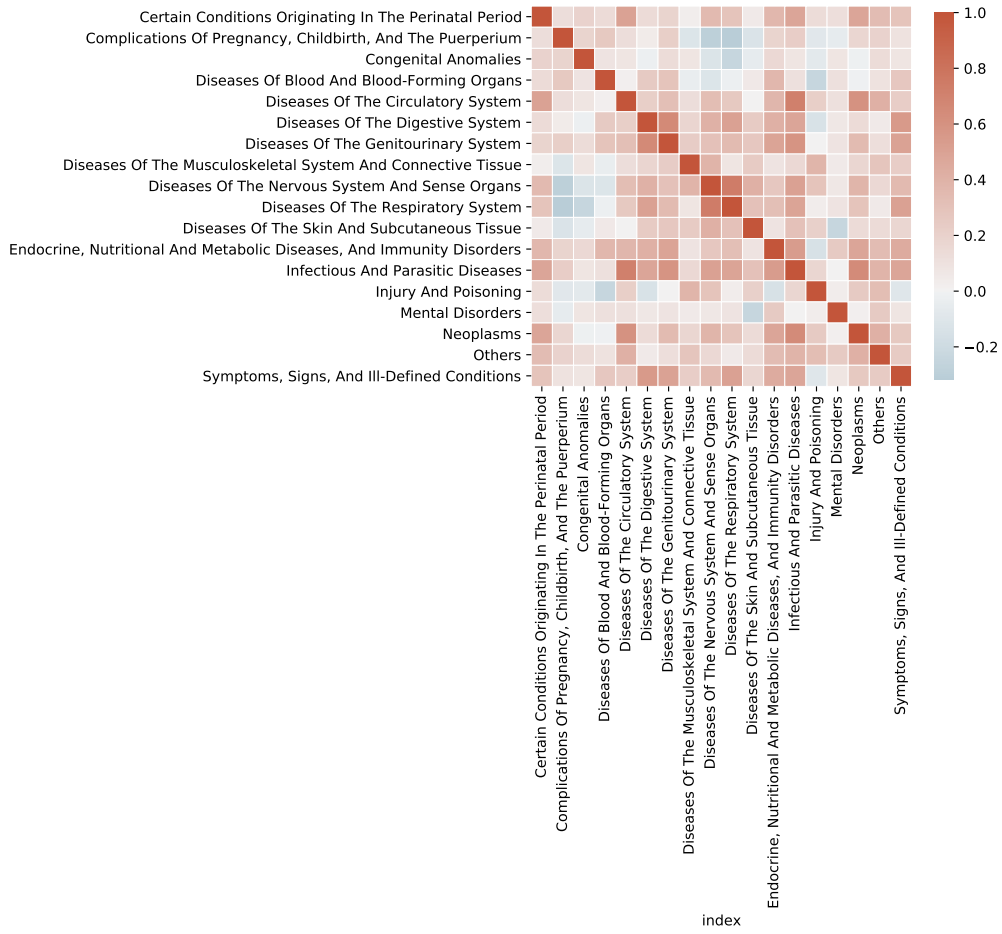

Figure S2: **Intrinsic evaluation of embeddings**. Learned embeddings are extracted from the pre-training stage for the base + age input embedding specification. The heatmap represents the cosine similarity for all the diagnosis codes in our data aggregated at the chapter level. Negative values (blue shades) reflect opposite similarities, and positive values (red shades) represent close similarities.

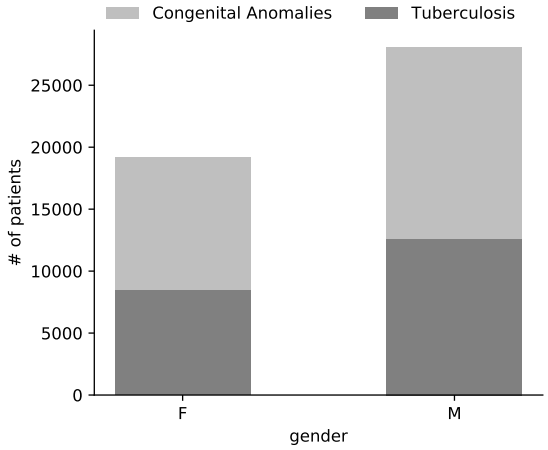

Figure S3: **Extrinsic evaluation of embeddings**. Assess the performance of the pre-trained PedBERT model in predicting the patient gender distribution for congenital anomalies (light gray) and tuberculosis (dark gray). The results presented here rely on the base + age embeddings specification using the 'pre-training test set'. Abbreviations: F = Female, M = Male.

Table S2: **Rare Genetic Diseases Prediction.** Number of patients in the 'fine-tuning training set' and 'fine-tuning test set' for selected rare genetic diseases at the two-digit ICD9 code level.

two-digit ICD9 diag code	# of patients in training data	# of patients in test data
27	3620	1541
28	1186	480
33	447	183
44	178	82
57	799	333
74	3816	1641