Learning to Extract Physician-Determined Diagnosis from Clinical Narratives Using Dual Noisy Labels

Anonymous ACL submission

Abstract

Electronic health records (EHRs) offer valuable opportunities for retrospective studies across a wide range of diseases. Many retrospective studies require reliable diagnosis labels for each patient but EHR datasets often do not provide precise labels that reflect the physician-determined diagnosis. Therefore, existing NLP approaches in retrospective studies typically rely on diagnosis proxies such as ICD codes, which are often noisy and may not reflect the physician-determined diagnosis. In this work, we propose a dual noisy label learning approach based on a latent variable framework to train a deep learning model that extracts physician-determined diagnoses from discharge summaries by leveraging two complementary noisy supervision signals: ICD diagnosis codes and discharge diagnoses. We evaluate our methodology on Acute Kidney Injury (AKI) diagnosis as a challenging case using a gold-standard dataset created through expert adjudication. Our approach substantially outperforms commonly used proxies and our interpretability analysis shows that the model's attention aligns closely with expert clinical reasoning. This approach can be applied for extraction of arbitrary physician-determined diagnoses and allows generating reliable diagnosis labels from clinical narratives without requiring manual annotation.

1 Introduction

Electronic health records (EHRs) databases (Finnegan and Mountford, 2025) such as MIMIC (Johnson et al., 2023) have become valuable resources for retrospective clinical research. In retrospective studies, researchers analyze historical patient records to investigate disease incidence and treatment patterns (Agniel et al., 2018; Sauer et al., 2022; Khaled et al., 2025; Fan et al., 2025; Mitchell et al., 2025) or to develop machine learning models that predict disease onset, progression, and patient outcomes across large populations (Miotto et al., 2016; Yang et al., 2022; Alhumaidi et al., 2025).

Retrospective studies require reliable diagnosis labels for each patient encounter. However, extracting such diagnosis information from EHR data is surprisingly challenging, because EHR datasets typically do not provide precise labels that reflect the diagnosis determined by physicians. Another challenge is that the definition of "diagnosis" itself is often unclear. Researchers have used several strategies to extract diagnosis from EHR data, including formula-based definitions, International Classification of Diseases (ICD) codes, discharge diagnoses, and expert adjudication, but each of these approaches has important limitations.

Formula-based diagnosis based on lab charts exist only for a handful of diseases and are often difficult to implement in practice (Smith et al., 2019; Strauss et al., 2024). Expert adjudication involves multiple clinicians independently reviewing EHR records and determining the diagnosis through consensus. This process is commonly considered the gold standard in clinical studies. However, expert adjudication is extremely time-consuming and expensive, and therefore is typically used only in regulatory clinical trials (Liu et al., 2016).

Because expert adjudication cannot scale to large EHR datasets, researchers often rely on diagnostic proxies such as ICD diagnosis codes, which are assigned by medical coders primarily for administrative and billing purposes (Pendergrass and Crawford, 2019; Quan et al., 2005; Yonamine et al., 2024; Wei and Denny, 2015). However, ICD diagnosis codes are often biased toward overdiagnosing, a phenomenon known as upcoding, which leads to many false positives (Joiner et al., 2024; O'malley et al., 2005; Costich et al., 2022; Joos et al., 2019). A less commonly used proxy is the discharge diagnosis section of discharge summaries, produced by the medical team at the time of patient discharge, which lists primary and sec-

ondary diagnoses in free text. Discharge diagnoses tend to under-report diagnoses because physicians may omit secondary or less critical diagnoses, resulting in many false negatives (Greer et al., 2016).

We believe that the physician's opinion is closer to the ground truth diagnosis than any other method. This opinion is documented in the discharge summary (Wei and Denny, 2015). Therefore, our objective is to extract diagnoses from discharge summaries, which we refer to as *physician-determined diagnoses*. However, extracting diagnoses from discharge summary is challenging: in some cases the diagnosis is explicitly stated, while in other cases it must be inferred from the clinical narratives. As a result, determining physician-determined diagnoses often requires clinicians to review discharge summaries and adjudicate, which is time-consuming and expensive.

A straightforward approach to extract diagnosis information from discharge summaries would be to train a natural language processing (NLP) model on manually labeled discharge summaries. A growing body of literature (Douglas et al., 2025; Sammani et al., 2021; Shickel et al., 2017) has demonstrated the effectiveness of deep learning models for automated ICD coding, largely because ICD labels are readily available at scale. However, if we are interested in training models to extract physician-determined diagnosis, this introduces the labeling bottleneck: manual annotation by clinical experts is prohibitively expensive and does not scale.

To address this challenge, we propose a weakly supervised learning approach that leverages existing noisy diagnostic proxies as supervision signals. Specifically, we treat ICD-coded diagnoses and discharge diagnoses as complementary noisy labels. We begin by identifying a "gray zone" in which these two proxies disagree, as shown in Figure 1. To account for noise and disagreement between the two proxies, we model the physician-determined diagnosis as a latent variable and marginalize over it when deriving the training objective. This enables the model to learn from both signals without treating either as ground truth. We then fine-tune a large language model on discharge summaries using this latent variable approach.

In this study, we focused on MIMIC-IV dataset and evaluated our methodology on Acute Kidney Injury (AKI) as a representative example which is one of the most challenging diseases to diagnose because the diagnostic guidelines are difficult to implement in practice (Fatehi and Ch, 2017).

Our contributions are as follows:

1. We introduce a dual noisy label learning approach based on a latent variable framework that trains a language model to extract physician-determined diagnoses from clinical narratives by leveraging two complementary noisy diagnostic labels without requiring manual annotation.

2. We evaluated our approach on Acute Kidney Injury (AKI) as a representative example using a gold-standard dataset created through expert adjudication from MIMIC-IV. Our approach achieves substantially higher agreement with these labels than commonly used proxies such as ICD codes and discharge diagnoses.

3. We provide an interpretability analysis, showing that the model's attention aligns with expert clinical reasoning.

2 Related Work

In the literature, there are several ways researchers attempt to automatically obtain diagnosis labels from EHR data. A large line of work focuses on predicting proxy signals such as ICD codes directly from clinical text (Dong et al., 2022). In these approaches, models take clinical notes as input and are trained to predict the ICD codes. Researchers have explored different deep learning encoders to learn representations of discharge summaries including Recurrent neural networks (RNNs) (Vu et al., 2020), Convolutional neural networks (CNNs) (Mullenbach et al., 2018; Liu et al., 2021), and pretrained transformer-based encoders (Aden et al., 2024; Nguyen et al., 2023; Huang et al., 2022; Douglas et al., 2025). While these methods have achieved strong performance in predicting administrative codes, they ultimately learn to reproduce the proxy labels themselves rather than the underlying physician-determined diagnosis. Nevertheless, ICD codes are often used as diagnosis labels in downstream EHR research.

Another line of work attempts to infer diagnoses using rule-based phenotyping algorithms. These approaches use complex rule sets involving diagnosis codes, medications, and laboratory values to identify patients with specific diseases (McCarty et al., 2011). However, such rule-based approaches require substantial domain expertise and are difficult to scale across diseases (Yu et al., 2018).

More recent work has explored weakly supervised phenotyping methods that attempt to infer whether a patient has a particular disease using diagnostic proxies (Yu et al., 2018; Liao et al., 2019;

Ferté et al., 2021). Given EHR notes, these approaches typically treat signals such as ICD codes and mentions of the target disease in clinical notes as proxy labels. In addition, they construct disease-specific features by identifying medically related concepts for each disease. Using these proxy signals and concept-based features, they train regression models that predict each proxy from the remaining signals. The resulting models are then combined, often through majority voting, to estimate whether the patient truly has the disease.

In contrast, our method is not disease-specific and does not require manually identifying concepts related to each disease. Instead, we leverage advance deep learning models originally developed for ICD code prediction to learn representations directly from the narratives in discharge summaries. Rather than only training separate models to predict each proxy signal, we explicitly model the systematic biases of different proxies and their relationship through a latent-variable framework to infer the physician-determined diagnosis.

3 Methodology

3.1 Problem Formulation

Let S_i denote a discharge summary associated with a hospital admission. Our objective is to determine the *physician-determined diagnosis* for a given disease D based on the discharge summary. However, the ground truth for physician-determined diagnosis is not available. Instead, each admission has two proxy diagnosis labels: D_i^{ICD}, indicating whether disease D appears in the ICD coding for the admission, and D_i^{Doctor}, indicating whether disease D appears in the discharge diagnosis section of the discharge summary. Given a dataset of discharge summaries

$$\mathcal{D} = \{(S_i, D_i^{\text{ICD}}, D_i^{\text{Doctor}})\}_{i=1}^{N}, \qquad (1)$$

our goal is to learn a model $f_\theta : \mathcal{S}_i \rightarrow \{0, 1\}$ that predicts the physician-determined diagnosis D_i^{PDD} directly from discharge summary S_i.

3.2 Diagnosis Proxies and Gray Zone

ICD-Coded Diagnosis (D^{ICD}) ICD codes are standardized diagnostic codes assigned by trained medical coders based on patient's medical record during admission. For a given disease D, we derive a binary proxy label from the ICD diagnosis codes associated with the hospital admission. Specifically, we assign $D^{\text{ICD}} = 1$ if the admission contains any ICD code corresponding to disease D, and $D^{\text{ICD}} = 0$ otherwise.

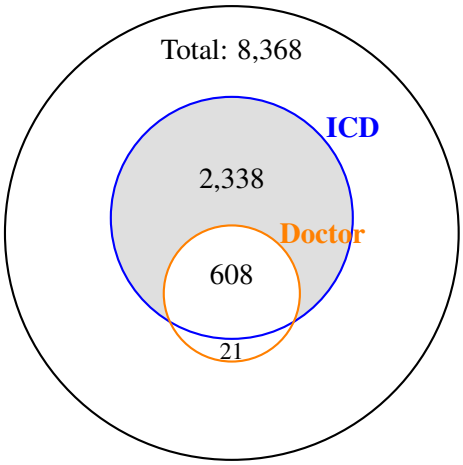

Figure 1: Venn diagram illustrating the agreement between ICD-coded AKI (ICD) and doctor-documented AKI (Doctor) across 8,368 ICU admissions. The blue circle represents admissions with ICD $= 1$, while the orange circle represents admissions where the discharge diagnosis explicitly mentions AKI. The gray region between the two circles denotes gray-zone admissions

Doctor-Documented Diagnosis (D^{Doctor}) A discharge summary is a clinical note written by physicians at the end of a hospital admission summarizing the patient's clinical course, diagnoses, and treatments. The *Discharge Diagnosis* section of the discharge summary typically lists the primary and secondary diagnoses.

Because discharge diagnoses are written as free text, disease names may appear in multiple surface forms. To identify whether disease D was documented in the discharge diagnosis list, we extracted the Discharge Diagnosis section and matched terms corresponding to disease D and its common synonyms or abbreviations. If a match is found, we assign the label $D^{\text{Doctor}} = 1$; otherwise $D^{\text{Doctor}} = 0$.

Agreement Between ICD and Doctor To better understand the relationship between the two diagnosis proxies, we examine their agreement in an ICU cohort containing 8,368 admissions for Acute Kidney Injury (AKI). Similar patterns are observed for other diseases (Appendix B). Figure 1 presents a Venn diagram illustrating the overlap between D^{ICD} and D^{Doctor}.

We observe from Figure 1 that among 8,368 admissions, only 21 fall into the category where disease D is documented in the discharge diagnosis section (Doctor $= 1$) but not coded by the hospital (ICD $= 0$). This suggests that when physicians explicitly document the target disease, coders al-

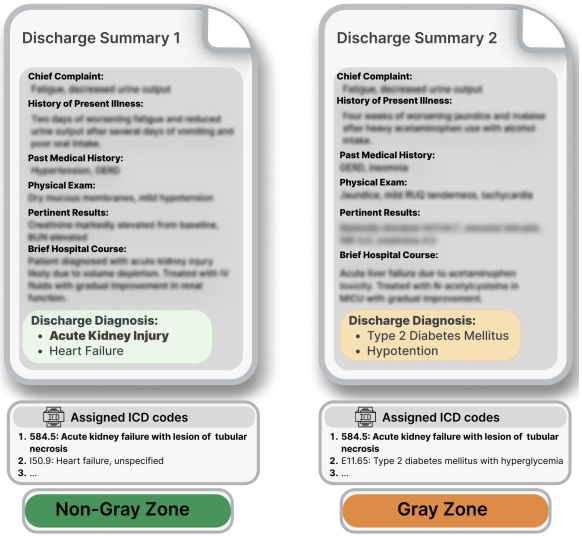

Figure 2: Illustration of gray zone area between two diagnosis proxies. When ICD coding indicates AKI but the discharge diagnosis does not (right), the admission falls into the gray zone.

most always assign the corresponding ICD code. However, among the 2,338 admissions coded as disease D by medical coders, only 26% are also documented in the discharge diagnosis section. This discrepancy defines gray zone.

Gray Zone Definition We define a *gray-zone admission* as an admission where the two proxy labels disagree, i.e., $D_i^{\text{ICD}} \neq D_i^{\text{Doctor}}$, as shown in Figure 2. This uncertainty arises not necessarily because the notes are more difficult to interpret, but because of noise in the weak supervision signals.

3.3 Deep learning model for extraction of physician-determined diagnosis

Our goal is to train a neural network that extracts physician-determined diagnosis for a target disease D without the need to create expert-labeled training data. Recent advances in deep learning, particularly with large language models (LLMs), have demonstrated strong capabilities in extracting diagnostic information from clinical text such as predicting ICD codes (Douglas et al., 2025; Masud et al., 2023; Huang et al., 2022). The state-of-the-art language model designed for ICD code prediction is PLM (Pretrained Language Model) (Huang et al., 2022), a small transformer-based model that is capable of handling arbitrary document length. Given that ICD code prediction aligns closely with our objective, we used PLM as the architecture of our neural network. We describe PLM model in more details in Appendix A.

Although we used PLM in our methodology, our training framework is model-agnostic and can be applied to any encoder-based architecture capable of processing clinical narratives.

Prior work on ICD code prediction (Huang et al., 2022) shows that training with supplementary labels improves performance. Following this approach, we train the model to predict the K most frequent ICD codes in the training dataset and formulate the task as a multi-label classification problem. Therefore, we have K ICD codes and one label for physician-determined diagnosis for disease D which we refer to as PLM-PDD.

A key challenge, however, is the absence of ground truth for physician-determined diagnosis for supervised training. We propose a strategy that leverages D^{ICD} and D^{Doctor} as two sources of weak supervision to train the model.

3.4 Latent Variable Framework

D^{ICD} and D^{Doctor} are informative but imperfect indicators of the physician-determined diagnosis. Instead of treating either proxy as ground truth, we model the physician-determined diagnosis as a latent variable.

Let $z_i \in \{0, 1\}$ denote the latent physician-determined diagnosis for admission i. Our objective is to learn a model that predicts the latent diagnosis z_i from the discharge summary. We first use PLM model parameterized by θ to compute the probability that the physician determined the presence of disease D: $\pi_i = P_\theta(z_i = 1 \mid S_i)$.

The two proxy labels are assumed to be generated from the latent diagnosis through noisy observation processes:

$$D_i^{\text{ICD}} \mid z_i \sim \text{Bernoulli}(\alpha_{z_i}),$$
$$D_i^{\text{Doctor}} \mid z_i \sim \text{Bernoulli}(\beta_{z_i}),$$

where $\alpha_1 = P(D^{\text{ICD}} = 1 \mid z = 1)$ and $\alpha_0 = P(D^{\text{ICD}} = 1 \mid z = 0)$ describe the behavior of ICD coding, and $\beta_1 = P(D^{\text{Doctor}} = 1 \mid z = 1)$ and $\beta_0 = P(D^{\text{Doctor}} = 1 \mid z = 0)$ describe the behavior of physician-written discharge diagnoses.

Given the clinical biases between proxy diagnosis labels, as discussed in Section 3.2, we model asymmetric clinical biases between D^{ICD} and D^{Doctor} by imposing strong constraints on two of the parameters. If disease D truly occurred, ICD coding almost always records it, so we assume $\alpha_1 \approx 1$. On the other hand, physicians rarely list disease D in the discharge diagnosis when it did not occur, which implies $\beta_0 \approx 0$. Therefore, we fix

these two parameters and only learn the remaining parameters α_0 and β_1 during training.

Given the discharge summary S_i, the likelihood of the observed proxy labels is obtained by marginalizing over the latent diagnosis:

$$P(D_i^{\text{ICD}}, D_i^{\text{Doctor}} \mid S_i) = \sum_{z_i \in \{0,1\}} P_\theta(z_i \mid S_i) \cdot P(D_i^{\text{ICD}} \mid z_i) P(D_i^{\text{Doctor}} \mid z_i).$$

For the target disease D, training minimizes the negative log-likelihood of the two observed proxy labels across all admissions:

$$\mathcal{L}_{\text{PDD}} = -\sum_i \log P(D_i^{\text{ICD}}, D_i^{\text{Doctor}} \mid S_i).$$

In addition to the target disease, the PLM model is trained to predict the top K most frequent ICD codes in the dataset as auxiliary tasks. Let $Y_i = (Y_{i1}, \ldots, Y_{iK})$ denote the vector of ICD code indicators for admission i, where each element Y_{ik} indicates the presence of ICD code k. For auxiliary ICD labels, we use the standard likelihood:

$$P(Y_i \mid S_i) = \prod_{k=1}^{K} P(Y_{ik} \mid S_i).$$

The corresponding negative log-likelihood is

$$\mathcal{L}_{\text{ICD}} = -\sum_i \log P(Y_i \mid S_i).$$

The final training objective is: $\mathcal{L} = \mathcal{L}_{\text{PDD}} + \mathcal{L}_{\text{ICD}}$.

3.5 Implications of the Constraints

The fixed constraints $\alpha_1 \approx 1$ and $\beta_0 \approx 0$ lead to a clear interpretation of the likelihood in relation to the gray zone. In the non-gray zone, the likelihood reduces to $\pi_i \beta_1$ when both proxies are positive and $(1 - \pi_i)(1 - \alpha_0)$ when both are negative, which are maximized at $\pi_i = 1$ and $\pi_i = 0$ respectively. This means the model treats the agreed label as ground truth in the non-gray zone. In the gray zone, the likelihood becomes $\pi_i(1 - \beta_1) + (1 - \pi_i)\alpha_0$, and the model's prediction depends on the learned parameters α_0 and β_1. This is precisely the desired behavior: confident supervision where the proxies agree, and learned uncertainty where they disagree.

4 Dataset and Experimental Setup

4.1 Dataset

In this study, we used MIMIC-IV, the fourth release of the Medical Information Mart for Intensive Care, a publicly available and de-identified electronic health record database from Beth Israel Deaconess Medical Center covering the years 2008–2022. MIMIC-IV provides patient health information including discharge summaries, diagnostic codes, and laboratory results, which are the primary data sources used in this study. The dataset also contains additional information such as demographics, medication administration records, and procedures. In total, the dataset includes 546,028 hospital admissions across 223,452 patients.

4.2 Target Disease

In this study, we evaluated our methodology on AKI as a representative example. AKI is a critical condition characterized by a sudden decrease in kidney function. AKI is one of the most challenging diseases to diagnose because the diagnostic guidelines are difficult to implement in practice (Fatehi and Ch, 2017), and no single therapy has yet been identified to prevent or mitigate AKI (Liu et al., 2016; Zarbock et al., 2024).

4.3 Cohort Selection

Because most retrospective studies of AKI have focused on ICU cohorts (Case et al., 2013; Jentzer et al., 2020; Park et al., 2010), and AKI is substantially more prevalent and severe among ICU patients (Pickkers et al., 2021), our objective is to extract Acute Kidney Injury (AKI) diagnoses from ICU admissions using discharge summaries. We therefore selected all adult ICU admissions in MIMIC-IV with an available discharge summary and meeting a set of clinical inclusion criteria (Appendix C) which resulted in a cohort of 41,840 ICU admissions. We reserved 20% of these ICU admissions as a held-out test set resulting in 8,368 ICU admissions (Cohort 1), and used the remaining admissions for training and validation.

To obtain gold-standard labels, we sampled 150 admissions from Cohort 1 for expert adjudication. To do so, we over-sampled the gray zone, where we anticipated the highest rate of AKI misclassification from ICD and discharge diagnoses. We randomly selected 90 admissions from the gray zone and 60 from the non-gray zone (Cohort 2).

Finally, to increase the amount of training data for PLM model, we augmented the ICU training cohort with additional non-ICU admissions that has a discharge summary, resulting in a total of 331,604 discharge summaries (Cohort 3). These augmented examples are distributionally different from ICU cases, as ICU hospitalizations tend to involve more

complex clinical courses and are reported differently in discharge summaries. We therefore include these additional admissions to increase training data and investigate whether such augmentation improves model performance (Appendix D).

4.4 Expert Adjudication

Three nephrologists and one ER doctor provided physician-determined AKI for admissions in Cohort 2. To obtain physician-determined AKI diagnosis from each expert, we removed the Discharge Diagnosis section from each note and asked each expert to answer the following questions:

1. Based on the note, do you think the note writer thought the patient had AKI? (Yes / No)

2. Please highlight parts of the note which provide positive or negative evidence about AKI.

For each admission, we collected the labels and rationales from all four experts. In cases where all reviewers agreed, the label was accepted as the gold-standard. For cases of disagreement, a consensus meeting was held to adjudicate the final label, which was then considered as the gold-standard label for physician-determined AKI diagnosis.

4.5 Baselines

We refer to the predicted physician-determined AKI label produced by our model as **PLM-PDD**. During training, in addition to the physician-determined AKI label, the PLM model is trained to predict the $K = 50$ most frequent ICD codes in the dataset as auxiliary tasks. We compare our approach with several baselines:

AKI-ICD. An admission is classified as ICD-coded AKI (AKI-ICD $= 1$) if it contains any of the following diagnosis codes associated with AKI.

AKI-Doctor A proxy label derived from the discharge diagnosis section of the discharge summary.

KDIGO. A rule-based clinical definition of AKI based on serum creatinine measurements (Levin et al., 2013). This method has been widely used in retrospective studies for AKI detection.

PLM-ICD. The PLM model is fine-tuned to predict the same K ICD codes and the AKI label. For the AKI label, supervision is provided with ICD-coded AKI label, following the standard practice in ICD coding prediction (Huang et al., 2022).

5 Results

5.1 Comparison of AKI labels vs. Gold-standard Label

Given the adjudicated physician-determined AKI provided by the experts, we assessed the accuracy of different AKI diagnosis labels, including AKI-ICD, AKI-Doctor, PLM-ICD, PLM-PDD and KDIGO on Cohort 2 and then examine which AKI label best aligns with gold-standard label.

We first divided Cohort 2 into three groups based on the agreement pattern between AKI-ICD and AKI-Doctor, as shown in Table 1. For each group, we reported the number of agreements with the adjudicated physician-determined AKI across all AKI diagnosis methods, covering both the non–gray zone and gray zone discharge summaries.

In the subset where AKI-ICD and AKI-Doctor agree, most of the methods show very high agreement with the expert. Specifically, PLM-PDD has only one disagreement with the expert, and AKI-ICD and AKI-Doctor have five disagreements. This strongly supports our hypothesis that when AKI-ICD and AKI-Doctor agree, they have very few disagreements with physician-determined AKI.

In cases where AKI-ICD and AKI-Doctor disagree, clear differences emerge. The expert agreed with PLM-PDD in 77 out of 90 cases (86%), PLM-ICD in 65 cases (72%), AKI-ICD in 62 cases (68%), and AKI-Doctor in 35 cases (32%). Importantly, PLM-PDD significantly outperforms its respective weak supervision labels in the gray zone, demonstrating the advantage of our methodology.

Given that Cohort 2 is a biased sample from Cohort 1, we report population-weighted accuracy. This metric calculates the accuracy within each group and then takes a weighted average based on that group's size in Cohort 1. Figure 3 reports the population-weighted accuracy using bootstrapping. We used bootstrapping to ensure robust estimate: in each of 1,000 iterations, we sampled 60 admissions from the non–gray zone and 90 from the gray zone, and computed the weighted accuracy. We then report the mean and standard deviation across these iterations. The height of each bar indicates the mean weighted accuracy, and the error bar represents the standard deviation across 1,000 iterations.

PLM-PDD achieves the highest weighted accuracy (94.6%), followed by PLM-ICD (89.4%), AKI-ICD (85.9%), AKI-Doctor (78.0%) and KDIGO (75.3%). Additionally, PLM-PDD exhibits the lowest standard deviation among all AKI labels.

5.2 PLM Interpretability Analysis

As shown in the previous section, PLM-PDD achieved the highest agreement with gold-standard label for physician-determined AKI. In this section, we analyze the interpretability of PLM predictions

Table 1: Number of agreements with the gold-standard physician-determined AKI label over Cohort 2

Group	# Cohort 1	# Cohort 2	ICD	Doctor	KDIGO	PLM-ICD	PLM-PDD
ICD=0, Doctor=0	6009	30	27	27	23	28	**29**
ICD=1, Doctor=1	608	30	28	28	26	**30**	**30**
ICD=1, Doctor=0	1730	90	62	28	60	65	**77**

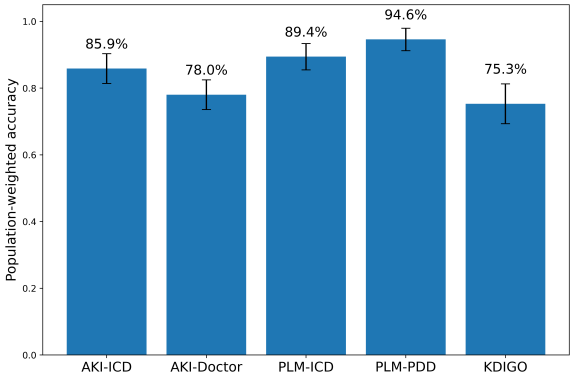

Figure 3: Bootstrapped population-weighted accuracy for different AKI Labels based on Cohort 2

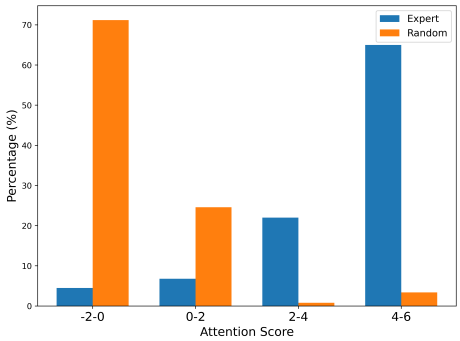

Figure 5: Comparison of attention scores between expert-highlighted and random phrases in Cohort 2

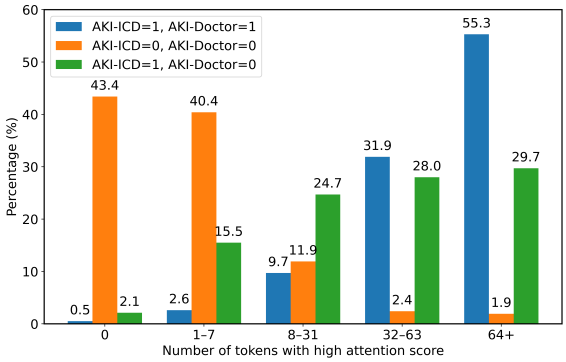

Figure 4: Distribution of number of high-attention tokens across non–gray zone positives, non–gray zone negatives, and gray zone cases (cohort 1)

using attention analysis. We investigate two questions regarding the interpretability of PLM:

Where PLM-PDD model pays attention to in discharge summary? We performed this analysis on Cohort 1 which includes 8,368 hospital admissions. For each discharge summary, we passed the text through PLM-PDD and extracted the raw attention scores for each token. These scores indicate how strongly the model attends to each token when making its prediction for PLM-PDD.

The attention analysis shows that PLM-PDD focuses on clinically meaningful concepts related to AKI diagnosis (e.g., creatinine, oliguria), treatment (e.g., dialysis, hemodialysis, CVVH, IVF), and comorbidities (e.g., CKD, CRF, ESRD). These findings indicate that PLM-PDD not only aligns closely with expert annotations, but also learns to prioritize the same clinical concepts that physicians rely on when diagnosing and managing AKI. Additional details are provided in the Appendix D.

Another important question is how PLM-PDD's attention is distributed across the discharge summary and how it differs between AKI and non-AKI cases. To investigate this, we estimated the number of AKI-related tokens as those with attention scores greater than 2.5, in each discharge summary. This threshold was empirically chosen as we identified it a reliable threshold to indicate AKI-related tokens. We plotted a histogram of frequency of tokens with attention score greater than 2.5 in three types of Cohort 1 admissions: confident positives (AKI-ICD = 1, AKI-Doctor = 1), confident negatives (AKI-ICD = 0, AKI-Doctor = 0), and gray zone, as illustrated in Figure 4.

As shown in Figure 4, discharge summaries labeled AKI-positive by both ICD and physician often contain many high-attention tokens, with 55.3% of these cases containing over 64 high-attention tokens. In contrast, more than 50% of discharge summaries labeled as AKI-negative contain no AKI-related tokens. Gray zone cases show a broad and variable distribution of high-attention tokens. In some discharge summaries, the model identifies little to no AKI-related evidence, while in others, it highlights numerous relevant tokens.

Does PLM-PDD's attention mechanism align with the expert's reasoning? We considered the Cohort 2 in which experts reviewed 150 discharge summaries. We extracted 356 highlighted phrases of varying lengths that impacted their assessment of AKI and extracted highest attention score of the

Table 2: Ablation study evaluating the contribution of auxiliary ICD supervision and gray-zone area.

Model	W. Accuracy	F1-score
PLM-PDD (Full)	94.6	0.920
PLM-PDD (No-ICD)	88.3	0.714
PLM-PDD (No-Gray)	86.9	0.652

model for each phrase.

In order to interpret these scores, we compare the distribution of these scores to a matched set of randomly selected phrases. To do so, for each expert-highlighted phrase length, we randomly selected a phrase of equal length from a randomly chosen discharge summary among the same 150 samples. We calculated the maximum attention score within each of the random phrases and compared the distribution the scores to those from the expert-highlighted phrases, as shown in Figure 5.

We could observe that over 60% of expert-highlighted phrases fall in the 4–6 attention score range, whereas over 70% of random phrases fall below 0. In contrast, only a small fraction of random phrases receive attention in range 4–6. These findings demonstrate that PLM-PDD's attention mechanism is highly aligned with expert clinical reasoning, as it assigns significantly higher attention to text segments highlighted by experts.

5.3 Ablation Study on PLM-PDD

The comparison between PLM-PDD and PLM-ICD in Section 5.1 shows the benefit of modeling dual noisy labels over standard ICD supervision. In this section, we further investigate other components of our methodology to understand which parts contribute to the final performance. In particular, we consider the following model variants and Table 2 reports the performance of these variants compared to the full PLM-PDD model.

PLM-PDD (No-ICD). To evaluate the role of multitask objective, we train a variant of the model that removes all auxiliary ICD supervision and optimizes only the PLM-PDD loss for target disease.

PLM-PDD (No-Gray). We consider removing all gray-zone cases entirely from the training data while still keeping the auxiliary ICD objectives.

5.4 Ablation Study: Importance of each section in Discharge Summary

In this section, we conducted an ablation study to gain insight into which part of discharge summary contributed the most for PLM-PDD prediction over Cohort 2. To do so, from discharge notes, we extracted different sections, including Chief Complaint, History of Present Illness (HOPI), Past Medical History, Pertinent Results (PR), Brief Hospital Course (BHC), and Discharge Medications. Each section was fed independently into the PLM to obtain the PLM-PDD label. We then measured weighted accuracy and F1-Score between the PLM-PDD and adjudicated physician-determined AKI (Table 3).

Table 3: Results for inputting various sections of discharge summary to PLM-PDD

Section	**W. Accuracy**	**F1-Score**
Full document	94.6	0.920
Allergies	80.2	0.000
Chief Complaint	80.2	0.000
History of Present Illness	83.7	0.473
Past Medical History	80.2	0.000
Pertinent Results	85.8	0.620
Brief Hospital Course	88.6	0.843
Discharge Medications	80.5	0.069
BHC+HOPI	93.4	0.896
Full document - PR	94.6	0.920

Among individual sections, the BHC contributed the most for PLM-PDD. The HOPI and PR were the next most informative sections. To assess whether combining sections could enhance performance, we combined the BHC and HOPI (BHC + HOPI), achieving F1-Score nearly equivalent to those obtained with the full discharge summary. To assess the influence of lab results within the discharge summary, we evaluated model performance after removing the PR section. The performance remained identical to that of the full document, indicating that PLM-PDD relies primarily on the clinical narrative rather than on lab values.

6 Conclusion

In this paper, we proposed an approach to train a deep learning model to automatically extract physician-determined diagnosis. We re-framed ICD codes and discharge diagnoses as complementary but noisy label sources and introduced a weakly supervised large language model (LLM) framework based on a latent variable model in order to automate the diagnosis extraction. Our experiments showed that PLM-PDD achieved the highest agreement with expert nephrologist labels and its attention mechanism closely aligned with expert clinical reasoning. While this study focused on AKI, this approach can be applied to other diseases to improve the quality of diagnosis labels for predictive modeling in clinical NLP.

7 Limitations

This study has several limitations. First, our study is based on the MIMIC-IV dataset from a single U.S. hospital system. Second, because adjudicated physician-determined labels were available only from our nephrology experts, we selected AKI as the case study. While this choice is clinically relevant and illustrative, additional evaluations on other diseases with domain-specific experts are required to confirm the generalizability of our approach. Furthermore, because the PhysioNet Credentialed Data Use Agreement explicitly prohibits sharing access to the data with third parties and requires strict zero-data-retention guarantees, we could not evaluate cloud-based large language models (e.g., GPT-4) as baselines. This constraint highlights the importance of locally deployable weakly supervised approaches for practical clinical deployment.

8 Ethical Considerations

This study uses de-identified clinical records from MIMIC-IV. Access to the dataset was obtained through PhysioNet after completing the required human-subjects protection training and signing the PhysioNet Credentialed Health Data Use Agreement. The agreement specifies that credentialed users may access the data solely for lawful scientific research purposes and may not share the data or provide access to third parties. Consistent with PhysioNet guidance, this restriction includes transmitting the data to external services or APIs, such as hosted large language model platforms. The dataset contains de-identified electronic health record data derived from routine clinical care.

For this study, 150 discharge summaries were annotated by coauthors of the paper, all of whom are medical doctors and members of the research team. The annotators were not recruited as research participants, and no data about the annotators themselves (e.g., demographics or behavior) was collected or analyzed; the annotations were produced solely to create reference labels for evaluating an automatic annotation model. Therefore, the annotators do not constitute human subjects in the context of this research. We consulted our university's Institutional Review Board (IRB), which determined that the study is exempt from IRB review because it involves only de-identified data and does not constitute human subjects research.

References

Ilyas Aden, Christopher HT Child, and Constantino Carlos Reyes-Aldasoro. 2024. International classification of diseases prediction from mimiic-iii clinical text using pre-trained clinicalbert and nlp deep learning models achieving state of the art. *Big Data and Cognitive Computing*, 8(5):47.

Denis Agniel, Isaac S Kohane, and Griffin M Weber. 2018. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361.

Norah Hamad Alhumaidi, Doni Dermawan, Hanin Farhana Kamaruzaman, and Nasser Alotaiq. 2025. The use of machine learning for analyzing real-world data in disease prediction and management: Systematic review. *JMIR Medical Informatics*, 13(1):e68898.

James Case, Supriya Khan, Raeesa Khalid, and Akram Khan. 2013. Epidemiology of acute kidney injury in the intensive care unit. *Critical care research and practice*, 2013(1):479730.

Centers for Medicare & Medicaid Services. 2018. Icd-10 general equivalence mappings: Fact sheet. Technical report, U.S. Department of Health & Human Services.

Steven G Coca, Paul Bauling, Tracy Schifftner, Clancy S Howard, Isaac Teitelbaum, and Chirag R Parikh. 2007. Contribution of acute kidney injury toward morbidity and mortality in burns: a contemporary analysis. *American Journal of Kidney Diseases*, 49(4):517–523.

Julia F Costich, Dana B Quesinberry, Lara K Daniels, and Ashley Bush. 2022. Trends in icd-10-cm–coded administrative datasets for injury surveillance and research. *Southern medical journal*, 115(11):801.

Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. 2022. Automated clinical coding: what, why, and where we are. npj digit. *Medicine*, 5(8).

James C Douglas, Yidong Gan, Ben Hachey, and Jonathan K Kummerfeld. 2025. Less is more: Explainable and efficient icd code prediction with clinical entities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30835–30847.

Michael R Ehmann, Jeremiah S Hinson, Steven Menez, Aria Smith, Eili Y Klein, and Scott Levin. 2023. Optimal acute kidney injury algorithm for detecting acute kidney injury at emergency department presentation. *Kidney Medicine*, 5(2):100588.

Pasquale Esposito, Francesca Cappadona, Marita Marengo, Marco Fiorentino, Paolo Fabbrini, Alessandro Domenico Quercia, Francesco Garzotto, Giuseppe Castellano, Vincenzo Cantaluppi, and

Francesca Viazzi. 2024. Recognition patterns of acute kidney injury in hospitalized patients. *Clinical Kidney Journal*, 17(8):sfae231.

Xuehui Fan, Jing Xu, Ruixue Ye, Qiu Zhang, and Yulong Wang. 2025. Retrospective cohort study based on the mimic-iv database: analysis of factors influencing all-cause mortality at 30 days, 90 days, 1 year, and 3 years in patients with different types of stroke. *Frontiers in Neurology*, 15:1516079.

Pedram Fatehi and Hsu Ch. 2017. Evaluation of acute kidney injury among hospitalized adult patients. *UpToDate, topic last updated Oct*, 2.

Thomas Ferté, Sébastien Cossin, Thierry Schaeverbeke, Thomas Barnetche, Vianney Jouhet, and Boris P Hejblum. 2021. Automatic phenotyping of electronical health record: Phevis algorithm. *Journal of Biomedical Informatics*, 117:103746.

Harriet Finnegan and Nicola Mountford. 2025. 25 years of electronic health record implementation processes: scoping review. *Journal of Medical Internet Research*, 27:e60077.

Raquel C Greer, Yang Liu, Deidra C Crews, Bernard G Jaar, Hamid Rabb, and L Ebony Boulware. 2016. Hospital discharge communications during care transitions for patients with acute kidney injury: a cross-sectional study. *BMC health services research*, 16(1):449.

CW Huang, SC Tsai, and YN Chen. 2022. Plm-icd: automatic icd coding with pretrained language models, arxiv. *arXiv preprint arXiv:2207.05289*.

Jacob C Jentzer, Thomas Breen, Mandeep Sidhu, Gregory W Barsness, and Kianoush Kashani. 2020. Epidemiology and outcomes of acute kidney injury in cardiac intensive care unit patients. *Journal of Critical Care*, 60:127–134.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Keith A Joiner, Jianjing Lin, and Juan Pantano. 2024. Upcoding in medicare: where does it matter most? *Health Economics Review*, 14(1):1.

Christopher Joos, Kevin Lawrence, Aubrey E Jones, Stacy A Johnson, and Daniel M Witt. 2019. Accuracy of icd-10 codes for identifying hospitalizations for acute anticoagulation therapy-related bleeding events. *Thrombosis Research*, 181:71–76.

Dmytro Khadzhynov, Danilo Schmidt, Juliane Hardt, Geraldine Rauch, Peter Gocke, Kai-Uwe Eckardt, and Kai M Schmidt-Ott. 2019. The incidence of acute kidney injury and associated hospital mortality: a retrospective cohort study of over 100 000 patients at berlin‘s charité hospital. *Deutsches Ärzteblatt International*, 116(22):397.

Afifa Khaled, Mohammed Sabir, Rizwan Qureshi, Camillo Maria Caruso, Valerio Guarrasi, Suncheng Xiang, and S Kevin Zhou. 2025. Leveraging mimic datasets for better digital health: A review on open problems, progress highlights, and future promises. *arXiv preprint arXiv:2506.12808*.

Adeera Levin, Paul E Stevens, Rudy W Bilous, Josef Coresh, Angel LM De Francisco, Paul E De Jong, Kathryn E Griffith, Brenda R Hemmelgarn, Kunitoshi Iseki, Edmund J Lamb, and 1 others. 2013. Kidney disease: Improving global outcomes (kdigo) ckd work group. kdigo 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney international supplements*, 3(1):1–150.

Katherine P Liao, Jiehuan Sun, Tianrun A Cai, Nicholas Link, Chuan Hong, Jie Huang, Jennifer E Huffman, Jessica Gronsbell, Yichi Zhang, Yuk-Lam Ho, and 1 others. 2019. High-throughput multimodal automated phenotyping (map) with application to phewas. *Journal of the American Medical Informatics Association*, 26(11):1255–1262.

Kathleen D Liu, Anitha Vijayan, Mitchell H Rosner, Jing Shi, Lakhmir S Chawla, and John A Kellum. 2016. Clinical adjudication in acute kidney injury studies: findings from the pivotal timp-2* igfbp7 biomarker study. *Nephrology Dialysis Transplantation*, 31(10):1641–1646.

Yang Liu, Hua Cheng, Russell Klopfer, Matthew R Gormley, and Thomas Schaaf. 2021. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953.

M A Masud, Chen-Kun Hsu, Hsiang-Chih Lin, and Chia-Hsiu Wu. 2023. Applying deep learning model to predict diagnosis code of medical records. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 246–253. IEEE.

Catherine A McCarty, Rex L Chisholm, Christopher G Chute, Iftikhar J Kullo, Gail P Jarvik, Eric B Larson, Rongling Li, Daniel R Masys, Marylyn D Ritchie, Dan M Roden, and 1 others. 2011. The emerge network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC medical genomics*, 4(1):13.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):26094.

Jonathon Mitchell, Michael R Ehmann, Scott Levin, Xihan Zhao, Steven Menez, Chirag R Parikh, Eili Y Klein, and Jeremiah S Hinson. 2025. Patterns in emergency clinician management of acute kidney injury. *American journal of kidney diseases: the official journal of the National Kidney Foundation*, 85(1):124–127.

Dennis G Moledina, Olivia Belliveau, Yu Yamamoto, Tanima Arora, Kyle A Carey, Matthew Churpek, Melissa Martin, Caitlin M Partridge, Sherry G Mansour, Chirag R Parikh, and 1 others. 2021. Variation in best practice measures in patients with severe hospital-acquired acute kidney injury: a multicenter study. *American Journal of Kidney Diseases*, 77(4):547–549.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1101–1111. Association for Computational Linguistics.

Thanh-Tung Nguyen, Viktor Schlegel, Abhinav Kashyap, Stefan Winkler, Shao-Syuan Huang, Jie-Jyun Liu, and Chih-Jen Lin. 2023. Mimic-iv-icd: A new benchmark for extreme multilabel classification. *arXiv preprint arXiv:2304.13998*.

Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. 2005. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639.

Woo Young Park, Eun Ah Hwang, Mi Hyun Jang, Sung Bae Park, and Hyun Chul Kim. 2010. The risk factors and outcome of acute kidney injury in the intensive care units. *The Korean journal of internal medicine*, 25(2):181.

Sarah A Pendergrass and Dana C Crawford. 2019. Using electronic health records to generate phenotypes for research. *Current protocols in human genetics*, 100(1):e80.

Peter Pickkers, Michael Darmon, Eric Hoste, Michael Joannidis, Matthieu Legrand, Marlies Ostermann, John R Prowle, Antoine Schneider, and Miet Schetz. 2021. Acute kidney injury in the critically ill: an updated review on pathophysiology and management. *Intensive care medicine*, 47(8):835–850.

Hude Quan, Vijaya Sundararajan, Patricia Halfon, Andrew Fong, Bernard Burnand, Jean-Christophe Luthi, L Duncan Saunders, Cynthia A Beck, Thomas E Feasby, and William A Ghali. 2005. Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data. *Medical care*, 43(11):1130–1139.

Ammar Sammani, Tasnim Abidin, Matthias Hügle, Maarten van Mourik, Cees Visser, Pim van der Harst, Folkert W Asselbergs, and 1 others. 2021. Automatic icd-10 coding of discharge diagnoses from electronic patient records in cardiology. *npj Digital Medicine*, 4(1):1–8.

Christopher M Sauer, Li-Ching Chen, Stephanie L Hyland, Armand Girbes, Paul Elbers, and Leo A Celi. 2022. Leveraging electronic health records for data science: common pitfalls and how to avoid them. *The Lancet Digital Health*, 4(12):e893–e898.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2017. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604.

Taylor D Smith, Victor Ortiz Soriano, Javier A Neyra, and Jin Chen. 2019. Identifying kdigo trajectory phenotypes associated with increased inpatient mortality. In *Proceedings. IEEE International Conference on Healthcare Informatics*, volume 2019, pages 10–1109.

Christian Strauss, Hendrik Booke, Lui Forni, and Alexander Zarbock. 2024. Biomarkers of acute kidney injury: from discovery to the future of clinical practice. *Journal of Clinical Anesthesia*, 95:111458.

Karen Triep, Alexander Benedikt Leichtle, Martin Meister, Georg Martin Fiedler, and Olga Endrich. 2022. Real-world health data and precision for the diagnosis of acute kidney injury, acute-on-chronic kidney disease, and chronic kidney disease: observational study. *JMIR medical informatics*, 10(1):e31356.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for icd coding from clinical text. *arXiv preprint arXiv:2007.06351*.

WQ Wei and JC Denny. 2015. Extracting research-quality phenotypes from electronic health records to support precision medicine. genome med. 2015; 7 (1): 41.

Renske Wiersema, Sakari Jukarainen, Ruben J Eck, Thomas Kaufmann, Jacqueline Koeze, Frederik Keus, Ville Pettilä, Iwan CC van der Horst, and Suvi T Vaara. 2020. Different applications of the kdigo criteria for aki lead to different incidences in critically ill patients: a post hoc analysis from the prospective observational sics-ii study. *Critical Care*, 24(1):164.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, and 1 others. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

Sean Yonamine, Chu Jian Ma, Rolake O Alabi, Georgia Kaidonis, Lawrence Chan, Durga Borkar, Joshua D Stein, Benjamin F Arnold, and Catherine Q Sun. 2024. Comparison of diagnosis codes to clinical notes in classifying patients with diabetic retinopathy. *Ophthalmology Science*, 4(6):100564.

Sheng Yu, Yumeng Ma, Jessica Gronsbell, Tianrun Cai, Ashwin N Ananthakrishnan, Vivian S Gainer, Susanne E Churchill, Peter Szolovits, Shawn N Murphy, Isaac S Kohane, and 1 others. 2018. Enabling phenotypic big data with phenorm. *Journal of the American Medical Informatics Association*, 25(1):54–60.

Alexander Zarbock, Lui G Forni, Marlies Ostermann, Claudio Ronco, Sean M Bagshaw, Ravindra L Mehta,

Rinaldo Bellomo, and John A Kellum. 2024. Designing acute kidney injury clinical trials. *Nature Reviews Nephrology*, 20(2):137–146.

A PLM Model

Huang et al. (Huang et al., 2022) introduced a framework called PLM (Pretrained Language Model) to automate the classification of discharge summaries into ICD codes using pretrained language models. In this paper, the authors first pretrain a transformer-based encoder, such as BERT and RoBERTa, on a masked language modeling (MLM) task using biomedical-specific datasets including PubMed and MIMIC-III. This pretraining improves the model's ability to capture semantic relationships between clinical terms and helps it learn domain-specific language patterns. As a result, the encoder is better suited for downstream tasks.

Discharge summaries are long documents that typically exceed the 512-token input limit of traditional transformer-based encoders such as BERT. PLM addresses this challenge by splitting each discharge summary into consecutive, equal-length, non-overlapping segments. These segments are then passed through a shared transformer encoder to obtain token-level embeddings. The resulting embeddings are processed by a label attention mechanism that generates a separate document embedding for each label, allowing the model to focus on different parts of the text depending on the label being predicted. Finally, these representations are fed into a linear layer to perform multi-label classification.

Suppose we have num_labels ICD codes (or some diseases). This model splits full document into $|s|$ consecutive segments with T tokens:

$$s = \{s_1, s_2, \ldots, s_{|s|}\} \in R^{|s| \times T}$$

Each segment s_i is passed through a shared transformer-based encoder, such as BERT or RoBERTa, to obtain token embeddings. These embeddings are concatenated across all segments to form a 3D tensor $H \in R^{|s| \times T \times E}$:

$$H = \text{concat}(\text{Encoder}(s_1), \ldots, \text{Encoder}(s_{|s|}))$$

where E is the embedding dimension of encoder. The tensor H is then reshaped into a 2D matrix $H \in R^{|s|T \times E}$. Then, two trainable matrices $V \in R^{E \times E}, W \in R^{E \times \text{num_labels}}$ (two feed forward neural network) are utilized to project H to a new representation space, followed by a softmax:

$$Z = \tanh(HV) \in R^{|s|T \times E},$$
$$A = \text{softmax}(ZW) \in R^{|s|T \times \text{num_labels}}$$

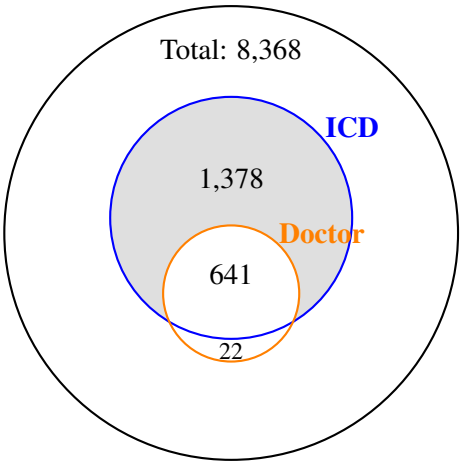

Figure 6: Venn diagram showing the agreement between ICD-coded Heart Failure (HF) ($D^{\text{ICD}} = 1$) and doctor-documented HF ($D^{\text{Doctor}} = 1$) across 8,368 ICU admissions. The gray area represents the gray zone between ICD and Doctor labels

in which A is interpreted as importance of each token for ICD-code i which we call label-aware attention. Matrix A is then used to compute a weighted sum of the original token embeddings, resulting in label-specific document representations

$$D = A^T H \in R^{\text{num_labels} \times E}$$

Each row D_i represents the embedding of the document with respect to ICD code i. The model also uses a learnable label embedding matrix $L \in R^{\text{num_labels} \times E}$, where each row L_i represents the embedding of ICD code i. Finally, the probability of assigning ICD code i to the document is computed by taking the dot product of L_i and D_i, followed by a sigmoid activation:

$$P_i = \text{sigmoid}(\langle L_i, D_i \rangle)$$

where P_i is the probability of patient being diagnosed with ICD code i.

B Gray Zone For Other Diagnoses

To examine whether the disagreement between ICD coding and physician-documented diagnoses is specific to AKI or reflects a more general phenomenon in EHR data, we performed a similar analysis for another common clinical condition: heart failure (HF) on the same ICU-Cohort with $8,368$ admissions and the results are shown in Figure 6.

Because expert-adjudicated labels were available only for AKI, we did not train or evaluate the PLM-PDD model for heart failure. Instead, this analysis serves to illustrate that the gray-zone phenomenon motivating our framework is not unique to AKI and arises for other diagnoses in EHR data.

C Implementation Details

Cohort Selection We selected adult admissions (age ≥ 18) meeting the following inclusion criteria: at least one ICU visit, at least two serum creatinine (SCr) measurements, at least one recorded urine output (UO), recorded weight, exactly one ICU stay per admission, and no record of renal transplant at discharge.

For training and evaluating the model, we removed the Discharge Diagnosis section from the discharge summary.

AKI-ICD The ICD codes associated with AKI are : ICD-9 codes 584, 584.5, 584.6, 584.7, 584.8, and 584.9, or ICD-10 codes N17, N17.0, N17.1, N17.2, N17.8, and N17.9.

KDIGO The KDIGO guidelines (Levin et al., 2013) define Acute Kidney Injury (AKI) using three criteria: (1) an increase in serum creatinine (SCr) by ≥ 0.3 mg/dL within 48 hours, (2) an increase to ≥ 1.5 times the baseline within seven days, and (3) urine output (UO) < 0.5 mL/kg/hour for at least six hours. Because MIMIC-IV lacks pre-admission data for most patients (except for a small subset of patients with prior admissions), we followed prior studies in the literature (Wiersema et al., 2020; Moledina et al., 2021; Esposito et al., 2024) and relied solely on serum creatinine (SCr) values recorded during the admission.

In this study, we adopted the following implementation of KDIGO (Wiersema et al., 2020; Moledina et al., 2021):

We considered only the serum SCr–based KDIGO criteria (1 and 2), excluding the urine output–based criterion (criterion 3). For criterion (2), several baseline definitions have been proposed in the literature. Some studies define baseline using the minimum or the first serum creatinine (SCr) measurement within the preceding seven days (Wiersema et al., 2020; Ehmann et al., 2023; Coca et al., 2007; Moledina et al., 2021), whereas others use the minimum SCr value observed during the entire hospitalization which we refer to as retrospective KDIGO (Triep et al., 2022; Khadzhynov et al., 2019).

In this work, we defined baseline as the minimum SCr value during the hospital stay (retrospec-

tive KDIGO).

PLM In our experiments, we used the RoBERTa-based encoder [1], which was pretrained on biomedical corpora including PubMed abstracts, PMC full texts, and MIMIC-III clinical notes using a masked language modeling (MLM) objective. During fine-tuning the PLM model, we did not freeze the encoder and used the the same hyperparameters used in the paper (Huang et al., 2022).

To ensure consistency across coding systems, ICD-9 and ICD-10 codes were mapped to unified disease categories using the General Equivalence Mappings (GEMS) provided by the Centers for Medicare & Medicaid Services (Centers for Medicare & Medicaid Services, 2018).

D Results

D.1 PLM Interpretability Analysis

We performed this analysis on Cohort 1 which includes 8,368 admissions. For each discharge summary, we passed the text through PLM-PDD and extracted the raw attention scores (ZW), as described in Section 3.3. These scores reflect how much each token contributes to the model's prediction for PLM-PDD. We then computed the average attention score for each unique token across all discharge summaries in Cohort 2. Tokens were then ranked by their average score and The top tokens, along with summary statistics, are presented in Table 4.

The attention analysis shows that PLM-PDD focuses on clinically meaningful concepts related to AKI diagnosis (e.g., creatinine, oliguria), treatment (e.g., dialysis, hemodialysis, CVVH, IVF), and comorbidities (e.g., CKD, CRF, ESRD). These findings indicate that PLM-PDD not only aligns closely with expert annotations, but also learns to prioritize the same clinical concepts that physicians rely on when diagnosing and managing AKI.

D.2 Evaluation of PLM Model on Top 50 ICD Codes

As we mentioned earlier, we constructed Cohort 3 including both ICU and non-ICU cases for training and evaluating our approach. In this section, we will evaluate whether including non-ICU cases was a reasonable design choice.

Our goal is to achieve high accuracy on Cohort 1, which consists of admissions with ICU stays. One option is to train the model on the subset of Cohort 3 that includes only ICU admissions ($N = 30{,}569$), as this subset shares the same distribution as Cohort 1. Alternatively, we could train the model on the full Cohort 3 training set, which includes both ICU and non-ICU admissions ($N = 238{,}754$). Although this full dataset is distributionally biased relative to Cohort 1, it offers a substantially larger number of admission.

We trained the model under two settings: (1) using the training subset of Cohort 3, and (2) using only admissions from training subset of Cohort 3 that has an ICU stay. We then evaluated both models on Cohort 1 and computed the F1-score for each of the 50 ICD codes. The results are presented in Figure 7.

We observe that for all ICD codes, the model trained on Cohort 1 consistently outperforms the model trained only on ICU cases. The macro F1-score of the Cohort 1–trained model is 0.69, compared to 0.59 for the ICU-only model. Specifically, for the AKI-ICD label, the F1-score is 0.84 for the Cohort 1 model versus 0.77 for the ICU model. These results confirm our hypothesis that using Cohort 1 for training the PLM model was beneficial.

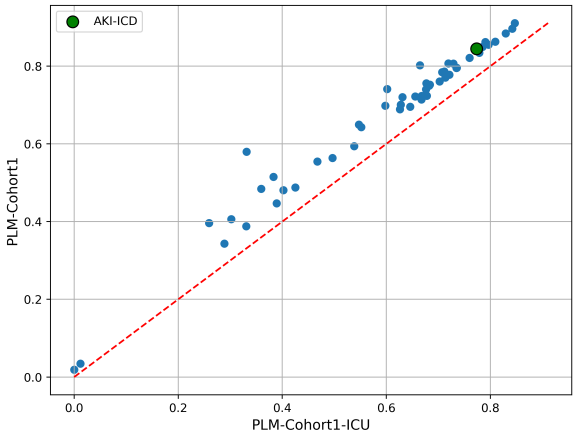

Figure 7: Comparison of PLM model performance when trained on Cohort 3 vs. ICU cases of Cohort 3, evaluated on Cohort 1. Results are shown across the 50 most frequent ICD codes based on F1-score

D.3 Attention score distribution of PLM-PDD for a few AKI-related tokens

In this section, we analyze the attention score distribution of PLM-PDD for a subset of AKI-related tokens. We selected 608 admissions from Cohort 2 where both AKI-Doctor and PLM-PDD labels were positive. As shown in Figure 8, the attention distributions vary across tokens but consistently show

[1] https://github.com/facebookresearch/bio-lm

Table 4: Tokens with the highest average attention scores assigned by PLM-PDD, with summary statistics of their attention scores across Cohort 2.

Token	Mean	Std	Min	Max	# Occurrences
ARF (Acute Renal Failure)	5.706	1.963	-3.798	7.244	305
CVVH (Continuous Veno-Venous Hemofiltration)	5.238	1.442	-2.504	6.821	259
ESRD (End-Stage Renal Disease)	4.774	1.367	-2.748	5.653	614
Oliguria (low urine output)	4.403	2.227	-2.450	7.049	260
CRF (Chronic Renal Failure)	4.354	1.128	-1.448	5.895	51
Dialysis	4.243	1.505	-2.922	6.687	1400
Injury	4.315	3.070	-2.800	7.167	157
CKD (Chronic Kidney Disease)	4.140	0.799	-2.535	5.587	1950
Hemodialysis	3.984	1.605	-2.955	6.099	536
Uric (Uric Acid)	3.683	2.790	-2.763	6.956	370
CRI (Chronic Renal Insufficiency)	3.628	2.143	-3.472	5.119	102
Creatinine	3.371	1.862	-4.197	6.995	22341
Kidney	3.258	2.959	-3.744	7.136	394
Nephrology	3.223	1.853	-2.736	5.989	161
Nephropathy (kidney disease)	2.822	2.161	-1.681	6.809	225
Renal	2.813	2.963	-3.431	7.112	6607
Resuscitation (vital function support)	2.797	1.375	-2.336	6.437	757
Hyperkalemia (high potassium level)	2.731	1.348	-2.593	5.385	399
Fluids (intravenous fluids)	2.693	1.279	-2.321	6.387	195
Casts (urinary casts)	2.669	2.321	-2.857	5.736	109
IVF (IV fluids for hydration)	2.594	1.016	-4.180	6.137	2519
Injured	2.351	3.150	-2.638	6.731	74
Vasopressin (hormone for BP control)	2.338	0.841	-1.339	4.924	150

that a substantial proportion receive high attention scores (greater than 2.5). However, some instances still exhibit lower attention scores, reflecting the fact that these tokens are not always used in AKI-related contexts. For instance, the token "injury" receives high attention when referring to kidney injury, but much lower attention when used in unrelated contexts. This demonstrates that PLM-PDD does not assign high attention to AKI-related tokens uniformly; rather, it modulates attention based on clinical context.

E Use of Generative AI

Generative AI tools were used only for language editing and proofreading. The authors take full responsibility for the content.

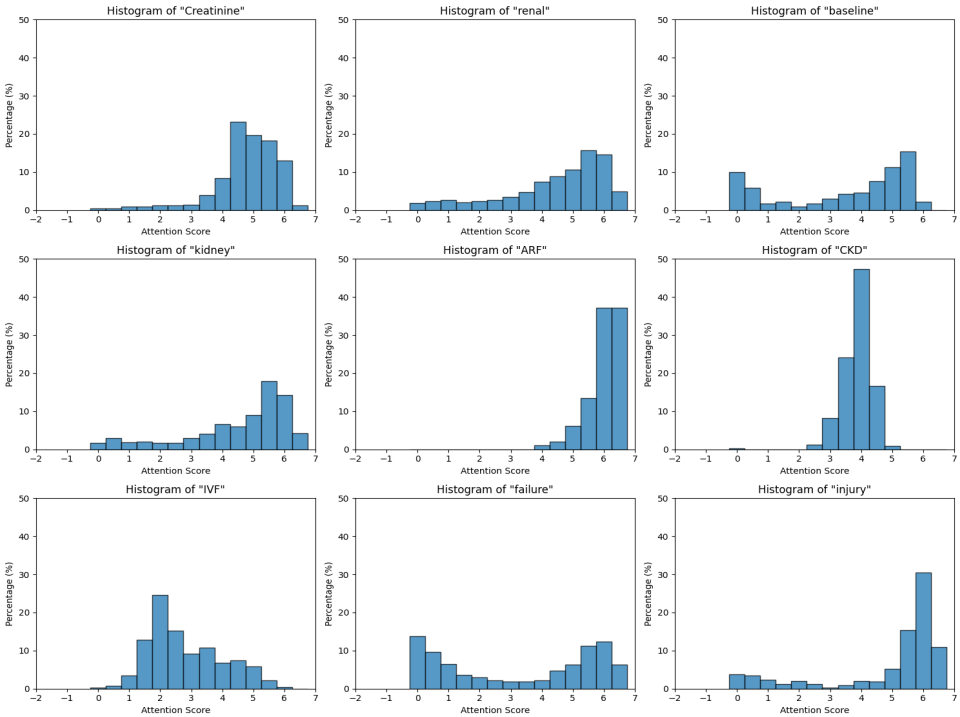

Figure 8: Attention score distribution for a few AKI-related tokens for PLM-PDD Model