

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

Vahid Mahzoon¹, Elmira Talebianaraki¹, Solabomi Ojeniyi¹, Tobin Efferen²
Sheetal Koul¹, Zivjena Vucetic¹, Avrum Gillespie¹, Slobodan Vucetic¹

¹Temple University, Philadelphia, PA, USA

²Beckman Coulter, Brea, CA, USA

Correspondence: vahid.mahzoon@temple.edu, slobodan.vucetic@temple.edu

Abstract

Retrospective studies using electronic health records (EHRs) often require diagnosis labels that reflect whether treating physicians considered a disease to be present. Such *physician-determined diagnosis* (PDD) labels are expensive to obtain through expert review, while available alternatives provide only noisy proxies. We exploit two complementary proxy labels: administrative diagnostic codes and free-text disease listings in discharge summaries. We introduce a dual-noisy-label learning framework that models PDD as a latent variable and treats these proxies as noisy observations of it, allowing a pretrained language model to learn from their agreement and disagreement without expert-labeled training data. We evaluate the approach on extracting Acute Kidney Injury (AKI) PDD from MIMIC-IV discharge summaries. Based on an expert-adjudicated evaluation set of 150 summaries, our PLM-PDD model achieves 94.6% population-weighted accuracy, compared with 89.4% for the same model trained using administrative coding, and 85.9% and 78.0% for AKI proxies. Largest gains occur in the *gray-zone*, where the proxies disagree. Ablation experiments show that gray-zone examples and auxiliary ICD supervision contribute to performance, while expert-highlighted evidence receives substantially higher model attention than matched random text. These results demonstrate that complementary noisy EHR labels can be used to learn clinically meaningful diagnosis labels without requiring expert annotation.

1 Introduction

Electronic health record (EHR) databases such as MIMIC-IV have become important resources for retrospective clinical research (Finnegan and Mountford, 2025; Johnson et al., 2023). Researchers use this data to study 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), as well as to develop models that predict disease onset, progression, and patient outcomes across large populations (Miotto et al., 2016; Yang et al., 2022; Alhumaidi et al., 2025). Many such studies require a reliable binary diagnosis label for each hospital admission. However, different labels available in EHR data capture different notions of diagnosis: a patient may satisfy a formula-based clinical definition, receive an administrative code, or be considered to have the disease by the treating physicians, and these labels need not agree. In this work, we focus specifically on whether the treating physicians considered a target disease to be present, as reflected in the clinical narrative. We refer to this target as the *physician-determined diagnosis* (PDD).

Obtaining PDD labels at scale is difficult. Disease-specific clinical criteria can provide useful phenotypes for some conditions (Strauss et al., 2024), but such criteria do not necessarily capture the treating physician’s diagnostic judgment and are not uniformly available across diseases. Alternatively, researchers can review the medical record and adjudicate whether the treating team considered a disease to be present. Such expert review provides a direct way to establish PDD labels, but it is time-consuming and expensive and therefore cannot readily be applied to hundreds of thousands of clinical notes. Moreover, the relevant diagnostic judgment is not always stated explicitly in a single location: it may need to be inferred from descriptions of symptoms, laboratory findings, assessments, treatments, and the patient’s clinical course throughout the discharge summary. Thus, training a supervised NLP model for PDD extraction would itself require a large expert-labeled dataset, recreating the labeling problem that the model is intended to solve.

EHR data do, however, contain routinely available diagnosis signals that can serve as weak supervision. The first is the *ICD-coded diagnosis*.

ICD codes are assigned by medical coders for administrative and billing purposes and are widely available at scale (Pendergrass and Crawford, 2019; Quan et al., 2005; Yonamine et al., 2024; Wei and Denny, 2015). They provide broad diagnostic coverage but are an imperfect proxy for the treating physicians’ diagnostic judgment; administrative coding practices can result in diagnoses being coded even when they do not correspond to the PDD (O’malley et al., 2005; Costich et al., 2022; Joos et al., 2019; Joiner et al., 2024). A second signal comes from the *Discharge Diagnosis section* of the discharge summary, a physician-written free-text list of primary and secondary diagnoses recorded at discharge. An explicit mention of a disease in this section provides strong evidence that the treating physicians considered the disease present, but secondary or less clinically salient diagnoses may be omitted (Greer et al., 2016). Thus, the two signals provide complementary but noisy views of PDD: ICD coding offers broader coverage, whereas the Discharge Diagnosis label is more selective.

We propose a weakly supervised approach that learns from these complementary diagnosis proxies without treating either one as ground truth. We call admissions in which the ICD and Discharge Diagnosis labels disagree the *gray zone*. Our approach models PDD as a latent variable and treats the two proxy labels as noisy observations generated from it. A pretrained language model predicts PDD directly from the clinical narrative, while the latent-variable model explicitly accounts for the different error patterns of the two observed proxies. When the proxies agree, they provide relatively strong weak supervision; when they disagree, the model learns how each proxy relates to the latent PDD rather than simply following either label. This dual-noisy-label framework therefore enables training on routinely available EHR labels without requiring expert-adjudicated PDD labels for supervision. Figure 1 provides an overview of the relationship between the two diagnosis proxies.

We evaluated the approach using Acute Kidney Injury (AKI), a common and critical condition characterized by a sudden decrease in kidney function (Ostermann, 2014; Chanda et al., 2026) as a case study in MIMIC-IV. To obtain an independent evaluation standard, three nephrologists and one emergency medicine physician adjudicated PDD for 150 held-out admissions after the Discharge Diagnosis section had been removed from the notes. We then

compared the PLM-PDD language model with several baselines, including the two proxy labels and a formula-based definition of AKI (Khwaja, 2012), using the expert-adjudicated PDD labels as the evaluation standard. Evaluation also included ablation and attention analyses.

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 physician-determined diagnosis targeted in our work. Nevertheless, ICD codes are often used as diagnosis labels in downstream EHR research.

Another line of work identifies disease phenotypes using rule-based 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 has the target disease phenotype.

In contrast, our method does not require manually constructing disease-specific features or identifying clinical concepts to use as model inputs. Instead, we leverage pretrained transformer models originally developed for ICD code prediction to learn representations directly from the narratives in discharge summaries. Rather than treating each proxy independently, we model the different error patterns of the two proxy labels through a latent-variable framework in which they are noisy observations of the physician-determined diagnosis.

3 Methodology

3.1 Problem Formulation

Let S_i denote a discharge summary associated with a hospital admission, which is a document written at the end of a hospital admission summarizing the patient’s clinical course, diagnoses, and treatments. Our objective is to determine whether the treating physicians considered disease D to be present, as reflected in the summary, which we call the *physician-determined diagnosis* (PDD). However, the ground truth for PDD is not readily 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^{DD} , indicating whether disease D is explicitly stated in the *Discharge Diagnosis* section of the discharge summary, which is typically a free-text list of primary and secondary diagnoses written at discharge time. Given a dataset of discharge summaries

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

our goal is to learn a model $f_\theta(S_i)$ that predicts D_i^{PDD} directly from S_i . For all model training and evaluation, the Discharge Diagnosis section is removed from S_i and is used only to construct D_i^{DD} .

3.2 Diagnosis Proxies and Gray-Zone

ICD-Coded Diagnosis (D_i^{ICD}). ICD codes are standardized diagnostic codes assigned by trained medical coders after discharge based on the patient’s record for the admission. For a given disease D , we assign $D_i^{\text{ICD}} = 1$ if the admission contains any ICD code corresponding to disease D , and $D_i^{\text{ICD}} = 0$ otherwise. For example, for Acute Kidney Injury, the corresponding ICD-9 codes are 584, 584.5, 584.6, 584.7, 584.8, and 584.9. ICD coding

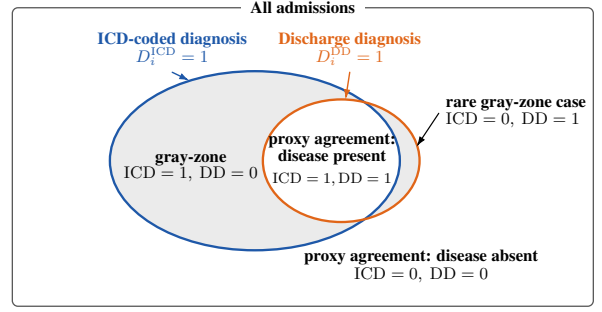


Figure 1: Conceptual relationship between ICD-coded and discharge diagnoses. The gray regions denote admissions in which the two diagnosis proxies disagree.

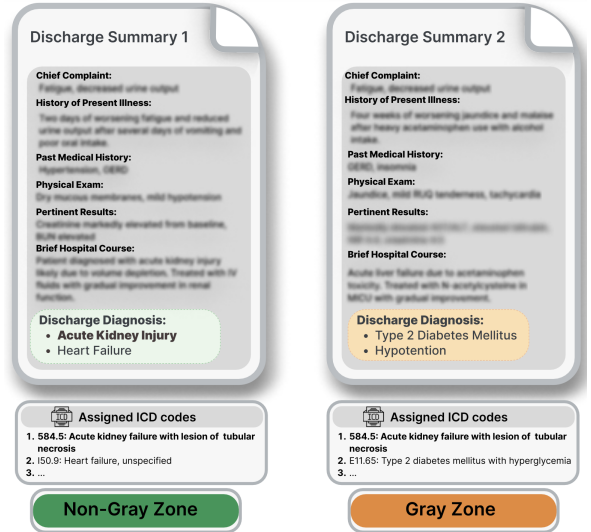


Figure 2: Illustration of discharge summaries outside (left) and within (right) the gray-zone.

provides broad coverage of diagnoses documented in the medical record, but because these codes are assigned for administrative and billing purposes, they are often more lenient than the physician-determined diagnoses.

Discharge-Diagnosis Label (D_i^{DD}). Because the Discharge Diagnosis section is written as free text, disease names may appear in multiple surface forms. To identify whether disease D was explicitly documented in DD section, we match terms corresponding to disease D and its common synonyms or abbreviations. If a match is found, we assign $D_i^{\text{DD}} = 1$; otherwise $D_i^{\text{DD}} = 0$. An explicit mention provides strong evidence that the treating physicians considered the disease present. However, physicians often omit secondary or less clinically salient diagnoses from the Discharge Diagnosis section. Therefore, $D_i^{\text{DD}} = 0$ does not necessarily indicate that the physicians considered the disease absent.

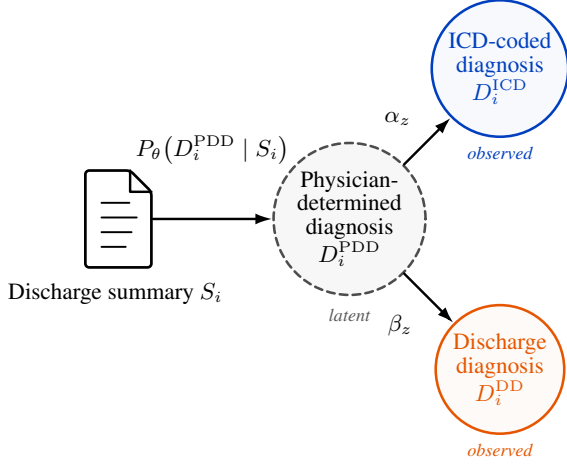


Figure 3: **Latent-variable model for physician-determined diagnosis with dual noisy labels.**

Complementary Noisy Labels and the Gray-Zone. The two proxy labels therefore have complementary error patterns. The ICD label provides broader diagnostic coverage but may contain false positives relative to PDD, whereas the discharge-diagnosis label is more selective but may contain false negatives due to omitted diagnoses. We therefore do not treat either proxy as ground truth.

When the two proxies agree, they provide stronger evidence about the physician-determined diagnosis than either proxy alone. We refer to admissions for which the two proxy labels disagree,

$$D_i^{\text{ICD}} \neq D_i^{\text{DD}}, \quad (2)$$

as the *gray-zone*, as illustrated in Figures 1 and 2. These cases are ambiguous from the perspective of weak supervision because the two available proxy labels provide conflicting signals. Our objective is to learn from both the agreement and disagreement between these complementary proxies rather than choosing either one as the target label.

3.3 Dual Noisy Label Learning

Neither D^{ICD} nor D^{DD} provides ground-truth supervision for PDD. Instead, we model the physician-determined diagnosis $D_i^{\text{PDD}} \in \{0, 1\}$ as a latent variable and treat the two proxy labels as noisy observations of it, as illustrated in Figure 3.

Given discharge summary S_i , the model parameterized by θ estimates

$$\pi_i = P_\theta(D_i^{\text{PDD}} = 1 \mid S_i), \quad (3)$$

the probability that the treating physicians considered disease D to be present.

We assume that the two proxy labels are generated from the latent PDD through separate noisy observation processes and are conditionally independent given D_i^{PDD} :

$$D_i^{\text{ICD}} \mid D_i^{\text{PDD}} = z \sim \text{Bernoulli}(\alpha_z), \quad (4)$$

$$D_i^{\text{DD}} \mid D_i^{\text{PDD}} = z \sim \text{Bernoulli}(\beta_z), \quad (5)$$

where

$$\alpha_1 = P(D^{\text{ICD}} = 1 \mid D^{\text{PDD}} = 1), \quad (6)$$

$$\alpha_0 = P(D^{\text{ICD}} = 1 \mid D^{\text{PDD}} = 0), \quad (7)$$

$$\beta_1 = P(D^{\text{DD}} = 1 \mid D^{\text{PDD}} = 1), \quad (8)$$

$$\beta_0 = P(D^{\text{DD}} = 1 \mid D^{\text{PDD}} = 0). \quad (9)$$

The complementary error patterns described in Section 3.2 motivate asymmetric constraints on these observation processes. We assume that ICD coding has almost perfect coverage when the physician-determined diagnosis is positive, corresponding to α_1 close to 1. Conversely, an explicit disease mention in the Discharge Diagnosis section is very unlikely when the physician-determined diagnosis is negative, corresponding to β_0 close to 0. We therefore fix $\alpha_1 = 1$ and $\beta_0 = 0$ and learn α_0 and β_1 during training. The learned parameter α_0 captures the tendency of ICD coding to produce positive labels when PDD is negative, while β_1 captures how often a positive PDD is explicitly listed in the Discharge Diagnosis section.

Because D_i^{PDD} is unobserved, we obtain the likelihood of the two observed proxy labels by marginalizing over it:

$$P(D_i^{\text{ICD}}, D_i^{\text{DD}} \mid S_i) = \sum_{z \in \{0,1\}} P_\theta(D_i^{\text{PDD}} = z \mid S_i) \cdot P(D_i^{\text{ICD}} \mid z) P(D_i^{\text{DD}} \mid z). \quad (10)$$

Training minimizes the negative log-likelihood of the observed proxy labels:

$$\mathcal{L}_{\text{PDD}} = - \sum_i \log P(D_i^{\text{ICD}}, D_i^{\text{DD}} \mid S_i). \quad (11)$$

Interpretation. The resulting objective has a simple interpretation in terms of the gray-zone. When the two proxies agree, the observation model provides strong evidence in favor of the corresponding PDD label. When they disagree, the model does not automatically follow either proxy. Instead, it weighs alternative explanations for the disagreement according to the learned noise parameters and the evidence in the discharge summary.

For example, when $D^{\text{ICD}} = 1$ but $D^{\text{DD}} = 0$, the disagreement may arise because ICD coding assigns a positive label despite a negative PDD, as captured by α_0 , or because a positive PDD was omitted from the Discharge Diagnosis section, as captured by $1 - \beta_1$. Thus, the framework provides high-confidence supervision where the proxies agree while explicitly modeling uncertainty where they disagree.

3.4 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 language models, 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). PLM (Pretrained Language Model) (Huang et al., 2022) is a transformer-based architecture developed for ICD code prediction that processes long documents by segmenting them. It is a relatively small transformer-based model that is capable of handling arbitrary document length. Given that ICD code prediction is closely related to our objective, we use PLM as the architecture of our neural network. We describe the PLM model in more detail in Appendix A.

Although we use PLM in our methodology, our training framework is model-agnostic and can be applied to any encoder-based architecture capable of processing clinical narratives. In our framework, PLM is used to parameterize $P_\theta(D_i^{\text{PDD}} = 1 \mid S_i)$.

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 top K most frequent ICD codes in the training dataset as auxiliary tasks and formulate the task as a multi-label classification problem. Let $Y_i = (Y_{i1}, \dots, Y_{iK})$ 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

$$\mathcal{L} = \mathcal{L}_{\text{PDD}} + \mathcal{L}_{\text{ICD}}.$$

We refer to the resulting model as **PLM-PDD**.

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 2008–2022. MIMIC-IV includes discharge summaries, diagnostic codes, and laboratory results, which are the primary data sources in this study. 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 Acute Kidney Injury (AKI) as a representative example. AKI is a critical condition characterized by a sudden decrease in kidney function. We selected AKI as our case study because our research team has domain expertise in AKI, enabling us to construct expert-adjudicated PDD labels for evaluation.

4.3 Cohort Selection

We focused our AKI evaluation on Intensive Care Unit (ICU) admissions because AKI is more prevalent and severe among ICU patients, and ICU cohorts are commonly used in retrospective AKI studies (Case et al., 2013; Jentzer et al., 2020; Park et al., 2010; Pickkers et al., 2021).

Applying the clinical inclusion criteria described in Appendix C to adult ICU admissions with an available discharge summary yielded an *ICU cohort* of 41,840 admissions. We randomly reserved 20% of these admissions as a held-out *ICU test set* ($N = 8,368$). The remaining 80% formed the *ICU development set* ($N = 33,472$), which was used for training and validation.

For expert evaluation, we sampled 150 admissions from the ICU test set to form the *adjudicated test subset*. Because disagreement between the two proxy labels is the primary challenge addressed by

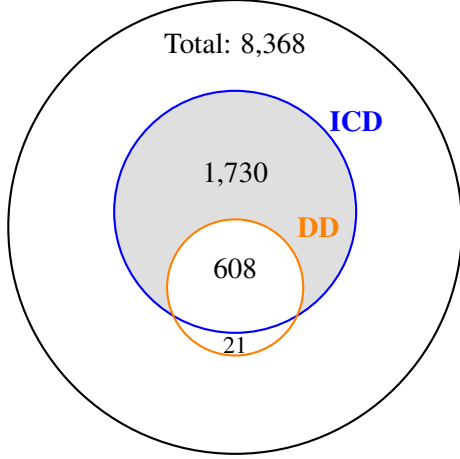


Figure 4: Venn diagram illustrating the relationship between ICD-coded AKI (ICD) and AKI documented in the Discharge Diagnosis section (DD) across 8,368 ICU admissions.

our method, we oversampled the gray-zone: 90 admissions were randomly selected from the gray-zone and 60 from the non-gray zone. Because this sampling strategy does not reflect the prevalence of the gray-zone in the full ICU test set, we will use population-weighted metrics when evaluating on the adjudicated test subset.

Finally, to increase the amount of data available for model training, we augmented the ICU training data with additional non-ICU admissions that had an available discharge summary, yielding an *expanded training cohort* of 331,604 discharge summaries. Because these additional admissions differ from the target ICU population, we evaluate the effect of this augmentation in Appendix D.

4.4 Empirical Agreement Between Weak Labels

We examined the empirical relationship between the two diagnosis proxy labels in the ICU test set. Figure 4 shows the overlap between the ICD-coded diagnosis label D^{ICD} and the Discharge Diagnosis label D^{DD} for AKI across the 8,368 admissions. Of the 2,338 admissions with $D^{\text{ICD}} = 1$, there were 608 with $D^{\text{DD}} = 1$. Conversely, among the 629 admissions with $D^{\text{DD}} = 1$, only 21 did not have a corresponding positive ICD label. Thus, disagreement between the two proxy labels is highly asymmetric: the dominant gray-zone pattern is $D^{\text{ICD}} = 1$ and $D^{\text{DD}} = 0$. This empirical pattern is consistent with the characterization in Section 3.2 of the Discharge Diagnosis label as a more selective proxy than ICD coding.

4.5 Expert Adjudication

Three nephrologists and one emergency medicine physician independently reviewed the 150 admissions in the adjudicated test subset and assigned a PDD label for AKI. As with the model input, we removed the Discharge Diagnosis section from each discharge summary before adjudication. Each expert was asked 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 that provide positive or negative evidence about AKI.

For each admission, we collected the binary label and highlighted evidence from all four experts. Inter-rater agreement among the experts is reported in Appendix D. When all reviewers agreed, we used the unanimous label as the expert-adjudicated PDD. Cases with disagreement were discussed in a consensus meeting to determine the final PDD label. We refer to the resulting consensus labels as the *expert-adjudicated gold-standard PDD labels*. These labels were used only for evaluation and were not used for model training.

4.6 Baselines

We compare PLM-PDD with the two diagnosis proxy labels defined in Section 3.2 and two additional AKI baselines:

ICD proxy (D^{ICD}). The ICD-coded diagnosis label defined in Section 3.2. For AKI, $D^{\text{ICD}} = 1$ if the admission contains an ICD code corresponding to AKI. The complete list of ICD codes is provided in Appendix C.

Discharge Diagnosis proxy (D^{DD}). The explicit-mention label derived from the Discharge Diagnosis section, as defined in Section 3.2. We assign $D^{\text{DD}} = 1$ when AKI or one of its matched synonyms or abbreviations is explicitly listed in the Discharge Diagnosis section. We further assess the recall of the discharge-diagnosis extraction procedure in Appendix C.

KDIGO. A creatinine-based implementation of the Kidney Disease: Improving Global Outcomes (KDIGO) clinical definition of AKI (Khwaja, 2012). This type of rule-based definition has been widely used for AKI detection in retrospective studies. Implementation details, including our definition of baseline serum creatinine, are provided in Appendix C.

PLM-ICD. PLM architecture (Huang et al., 2022) is trained using D^{ICD} as direct supervision for the AKI prediction task, replacing the dual-noisy-label objective used by PLM-PDD. As in PLM-PDD, the model is trained to predict the $K = 50$ most frequent ICD codes as auxiliary tasks.

5 Results

5.1 Comparison with Expert-Adjudicated PDD

We first evaluated how well the two diagnosis proxy labels and the model-based approaches agree with the expert-adjudicated PDD labels for AKI. We compared the ICD proxy D^{ICD} , the Discharge Diagnosis proxy D^{DD} , KDIGO, PLM-ICD, and PLM-PDD on the adjudicated test subset.

We divided the adjudicated test subset according to the agreement pattern between D^{ICD} and D^{DD} , as shown in Table 1. The adjudicated subset contains 30 admissions for which both proxies are negative, 30 for which both proxies are positive, and 90 from the dominant gray-zone pattern, $D^{\text{ICD}} = 1$ and $D^{\text{DD}} = 0$. We did not sample from the rare reverse-disagreement pattern, $D^{\text{ICD}} = 0$ and $D^{\text{DD}} = 1$, because it accounts for less than 0.5% of the ICU test set.

When the two proxy labels agree, both proxies also show high agreement with the expert-adjudicated PDD. Among the 30 admissions for which both proxies are negative, D^{ICD} and D^{DD} each agree with the expert-adjudicated label in 27 cases, while PLM-PDD agrees in 29. Among the 30 admissions for which both proxies are positive, both proxy labels agree with the expert-adjudicated PDD in 28 cases, while PLM-ICD and PLM-PDD agree in all 30 cases. These results support our assumption that agreement between the two proxy labels provides relatively strong supervision.

Clearer differences emerge in the gray-zone. Among the 90 admissions with $D^{\text{ICD}} = 1$ and $D^{\text{DD}} = 0$, PLM-PDD agrees with the expert-adjudicated PDD in 77 cases (86%), compared with 65 cases (72%) for PLM-ICD, 62 cases (69%) for D^{ICD} , 60 cases (67%) for KDIGO, and 28 cases (31%) for D^{DD} . Thus, PLM-PDD shows substantially higher agreement with expert-adjudicated PDD than either weak label in the gray-zone.

Because the adjudicated test subset intentionally oversamples gray-zone admissions, raw accuracy on these 150 cases does not reflect the dis-

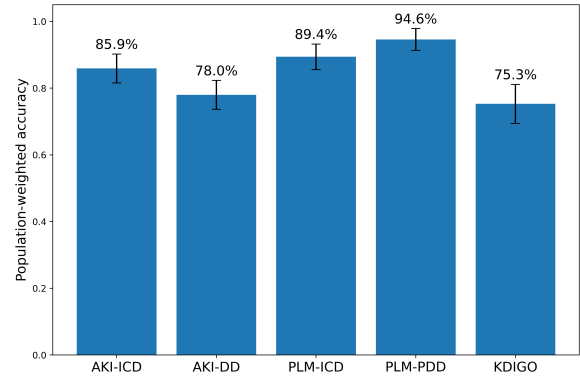


Figure 5: Bootstrapped population-weighted accuracy for different AKI Labels

tribution of the full ICU test set. We therefore report population-weighted accuracy, calculated by estimating accuracy within each sampled proxy-agreement group and weighting it according to that group’s prevalence in the ICU test set.

Figure 5 reports population-weighted accuracy using 1,000 bootstrap iterations. In each iteration, we resampled admissions within the adjudicated proxy-agreement groups and recomputed the population-weighted accuracy. We report the mean and standard deviation across bootstrap iterations.

PLM-PDD achieves the highest population-weighted accuracy (94.6%), followed by PLM-ICD (89.4%), the ICD proxy (85.9%), the Discharge Diagnosis proxy (78.0%), and KDIGO (75.3%). PLM-PDD also exhibits the lowest standard deviation across bootstrap iterations.

5.2 Ablation Study on PLM-PDD

The comparison between PLM-PDD and PLM-ICD in Section 5.1 demonstrates the benefit of modeling the two noisy diagnosis proxies rather than using the ICD proxy as direct supervision. We further investigated two components of our methodology to understand their contribution to PLM-PDD performance. Table 2 reports the performance of the following variants relative to the full model.

No auxiliary ICD supervision. To evaluate the role of the multitask objective, we trained a variant of PLM-PDD that removes the auxiliary prediction of the $K = 50$ ICD codes and optimizes only the PDD objective.

No gray-zone training. To evaluate the contribution of cases in which the two proxy labels disagree, we trained a variant after removing all gray-zone admissions from the training data while retaining

Table 1: Number of agreements with the expert-adjudicated PDD label for AKI in the adjudicated test subset.

Proxy pattern	# ICU test	# Adjudicated	D^{ICD}	D^{DD}	KDIGO	PLM-ICD	PLM-PDD
$D^{ICD} = 0, D^{DD} = 0$	6009	30	27	27	23	28	29
$D^{ICD} = 1, D^{DD} = 1$	608	30	28	28	26	30	30
$D^{ICD} = 1, D^{DD} = 0$	1730	90	62	28	60	65	77

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

Model	W. Accuracy	F1
PLM-PDD (Full)	94.6	0.920
No auxiliary ICD	88.3	0.714
No gray-zone training	86.9	0.652

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+HPI	93.4	0.896
Full document without PR	94.6	0.920

the auxiliary ICD prediction objective.

Removing either component substantially reduces performance. Without auxiliary ICD supervision, population-weighted accuracy decreases from 94.6% to 88.3%, and F1 decreases from 0.920 to 0.714. Removing gray-zone training examples produces an even larger reduction, to 86.9% weighted accuracy and 0.652 F1. These results suggest that both auxiliary ICD supervision and training on proxy-disagreement cases contribute to the performance of PLM-PDD, with the gray-zone examples being particularly important for learning the distinction between the two noisy labels.

5.3 Analysis by Discharge Summary Section

We next investigated which sections of the discharge summary are most informative for PLM-PDD predictions on the adjudicated test subset. We extracted individual sections, including Chief Complaint, History of Present Illness (HPI), Past Medical History, Pertinent Results (PR), Brief Hos-

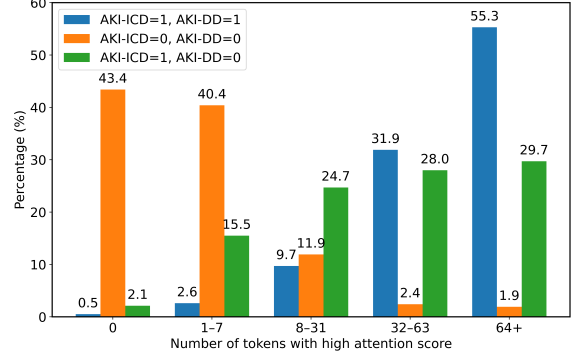


Figure 6: Distribution of number of high-attention tokens across non-gray zone positives, non-gray zone negatives, and gray-zone cases

pital Course (BHC), and Discharge Medications. Each section was independently provided as input to the trained PLM-PDD model, and we measured population-weighted accuracy and F1 against the expert-adjudicated PDD labels (Table 3).

Among individual sections, Brief Hospital Course yields the highest performance, with a weighted accuracy of 88.6% and an F1 of 0.843. History of Present Illness and Pertinent Results are the next most informative individual sections. Combining Brief Hospital Course and History of Present Illness further increases performance to 93.4% weighted accuracy and 0.896 F1, approaching the performance obtained using the full discharge summary.

We also evaluated the model after removing the Pertinent Results section from the full discharge summary. Performance remained unchanged at 94.6% weighted accuracy and 0.920 F1. Thus, PLM-PDD does not depend on the dedicated Pertinent Results section to achieve its full performance, suggesting that substantial diagnostic information is also available elsewhere in the clinical narrative.

5.4 Attention Analysis

We next analyzed PLM-PDD’s label-attention scores to investigate what information in the discharge summaries contributes to its predictions.

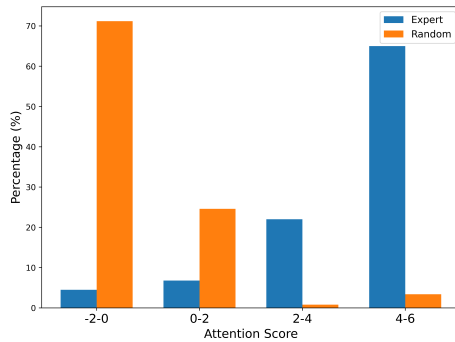


Figure 7: Comparison of attention scores between expert-highlighted and random phrases

We consider two questions: (1) which clinical concepts receive high attention from PLM-PDD, and (2) whether expert-highlighted evidence receives higher attention than comparable randomly selected text.

Which clinical concepts receive high attention?

We performed this analysis on the ICU test set of 8,368 admissions. For each discharge summary, we passed the text through PLM-PDD and extracted the raw label-attention score for each token. These scores are the weights assigned to tokens by the model’s label-attention mechanism when producing its PDD prediction.

The highest-attention tokens include clinically meaningful concepts related to AKI diagnosis (e.g., *creatinine*, *oliguria*), treatment (e.g., *dialysis*, *hemodialysis*, *CVVH*, *IVF*), and relevant comorbidities (e.g., *CKD*, *CRF*, *ESRD*). Additional examples are provided in Appendix D.

We also examined how the number of high-attention tokens varies across proxy-agreement groups. For visualization, we define high-attention tokens as tokens with a raw attention score greater than 2.5. The selection of this threshold is described in Appendix C. Figure 6 shows the distribution of the number of high-attention tokens among admissions where both proxy labels are positive, both are negative, and the dominant gray-zone group $D^{\text{ICD}} = 1, D^{\text{DD}} = 0$.

Admissions for which both proxy labels are positive frequently contain many high-attention tokens: 55.3% contain more than 64 such tokens. In contrast, nearly half of admissions for which both proxy labels are negative contain no high-attention tokens. Gray-zone admissions show a substantially broader distribution. Some contain little high-attention evidence, while others contain

many high-attention tokens, consistent with the heterogeneity of cases for which the two weak labels disagree.

Do expert-highlighted phrases receive higher attention?

We next used the adjudicated test subset, for which experts highlighted portions of each discharge summary that provided positive or negative evidence about AKI. Across the 150 admissions, the experts highlighted 356 phrases of varying lengths. For each highlighted phrase, we extracted the maximum PLM-PDD attention score assigned to a token within that phrase.

To provide a comparison distribution, for each expert-highlighted phrase we randomly selected a phrase of the same length from one of the same 150 discharge summaries and calculated its maximum attention score. Figure 7 compares the resulting distributions.

More than 60% of expert-highlighted phrases have maximum attention scores in the 4–6 range, whereas more than 70% of matched random phrases have scores below 0. Only a small fraction of random phrases receive scores in the 4–6 range. Thus, text spans identified by clinicians as relevant evidence for AKI tend to receive substantially higher PLM-PDD attention scores than matched random text spans. This provides evidence that the model’s attention is associated with clinically relevant evidence identified independently by expert reviewers.

6 Conclusion

In this paper, we proposed an approach to extract physician-determined diagnoses (PDD) from clinical narratives without expert-labeled training data. We modeled the ICD and Discharge Diagnosis labels as complementary noisy observations of a latent PDD and developed a weakly supervised learning framework that exploits their different error patterns. In experiments, PLM-PDD achieved the highest agreement with expert-adjudicated PDD, and expert-highlighted evidence received substantially higher model attention than matched random text. Although we evaluated the approach on AKI, the framework is designed to extend to other diseases for which analogous diagnosis proxies can be constructed, which will be studied in future work.

7 Limitations

This study has several limitations. First, our study is based on the MIMIC-IV dataset from a single U.S. hospital system. The performance and error patterns of the proxy labels may differ across institutions and coding/documentation practices. Second, we evaluated expert-adjudicated PDD only for AKI, a disease for which our research team had the clinical expertise required for adjudication. Evaluation on additional diseases with appropriate domain experts is needed to establish whether the approach generalizes beyond AKI. Finally, the PhysioNet Credentialed Health Data Use Agreement prevented us from transmitting MIMIC-IV records to third-party hosted services, so we could not evaluate cloud-based language models as baselines.

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 150 discharge summaries used for expert adjudication were reviewed by four physician coauthors as part of the research process. No information about the annotators themselves was collected or analyzed. Our university's Institutional Review Board determined that the study was exempt from IRB review because it used de-identified data and did 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, k1479.

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.

Ashis Kumar Chanda, Elmira Talebianaraki, Vahid Mahzoon, Zivjena Vucetic, and Slobodan Vucetic. 2026. Concordance analysis between kdigo definition of acute kidney injury and its coding in clinical practice. *Kidney Medicine*, 8(7):101374.

Steven G Coca, Paul Bauling, Tracy Schiffner, 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):801–805.

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 digital medicine*, 5(1):159.

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.

- Thomas Ferté, Sébastien Cossin, Thierry Schaevebeke, Thomas Barnette, Vianney Jouhet, and Boris P Hejblum. 2021. Automatic phenotyping of electronic 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.
- Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. Plm-icd: Automatic icd coding with pretrained language models. In *Proceedings of the 4th clinical natural language processing workshop*, pages 10–20.
- 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, et al. 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 Khadzhyrov, 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):397–404.
- 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*.
- Arif Khwaja. 2012. KDIGO clinical practice guidelines for acute kidney injury. *Nephron Clinical Practice*, 120(4):c179–c184.
- Katherine P Liao, Jiehuan Sun, Tianrun A Cai, Nicholas Link, Chuan Hong, Jie Huang, Jennifer E Huffman, Jessica Gronsbell, Yichi Zhang, Yuk-Lam Ho, et al. 2019. High-throughput multimodal automated phenotyping (MAP) with application to PheWAS. *Journal of the American Medical Informatics Association*, 26(11):1255–1262.
- 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.
- Jakir Hossain Bhuiyan Masud, Chen-Cheng Kuo, Chih-Yang Yeh, Hsuan-Chia Yang, and Ming-Chin Lin. 2023. Applying deep learning model to predict diagnosis code of medical records. *Diagnostics*, 13(13):2297.
- Catherine A McCarty, Rex L Chisholm, Christopher G Chute, Iftekhar J Kullo, Gail P Jarvik, Eric B Larson, Rongling Li, Daniel R Masys, Marylyn D Ritchie, Dan M Roden, et al. 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, et al. 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 Wiegrefe, 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.
- Marlies Ostermann. 2014. Diagnosis of acute kidney injury: Kidney disease improving global outcomes criteria and beyond. *Current opinion in critical care*, 20(6):581–587.
- 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):181–187.
- 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.
- 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.
- 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](#). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3335–3341. International Joint Conferences on Artificial Intelligence Organization. Main track.
- Wei-Qi Wei and Joshua C Denny. 2015. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome medicine*, 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, et al. 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, et al. 2018. Enabling phenotypic big data with phenorm. *Journal of the American Medical Informatics Association*, 25(1):54–60.

A PLM Model

(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.

Let M denote the total number of prediction labels. The model splits each document into $|s|$ consecutive segments with T tokens:

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

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 $H \in R^{|s| \times T \times E}$:

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

where E is the embedding dimension of the encoder. The tensor H is then reshaped into a 2D matrix $H \in R^{|s| \times T \times E}$. Two trainable matrices $V \in R^{E \times E}$ and $W \in R^{E \times M}$ are then used to compute label-specific attention:

$$Z = \tanh(HV) \in R^{|s| \times T \times E},$$

$$A = \text{softmax}(ZW) \in R^{|s| \times T \times M}.$$

Each column of A contains the label-attention weights over the document tokens for one prediction label. Matrix A is then used to compute a weighted sum of the original token embeddings, resulting in label-specific document representations:

$$D = A^T H \in R^{M \times E}.$$

Each row D_i represents the document embedding associated with label i . The model also uses a learnable label embedding matrix

$$L \in R^{M \times E},$$

where each row L_i is the learned embedding associated with label i . The probability associated with label i is computed by taking the dot product of L_i and D_i , followed by a sigmoid activation:

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

In PLM-PDD, one label-specific representation is used to parameterize $P_\theta(D_i^{\text{PDD}} = 1 \mid S_i)$, while the remaining label-specific representations are used to predict the K auxiliary ICD labels described in Section 3.4. Thus, for PLM-PDD, $M = K + 1$.

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 8.

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.

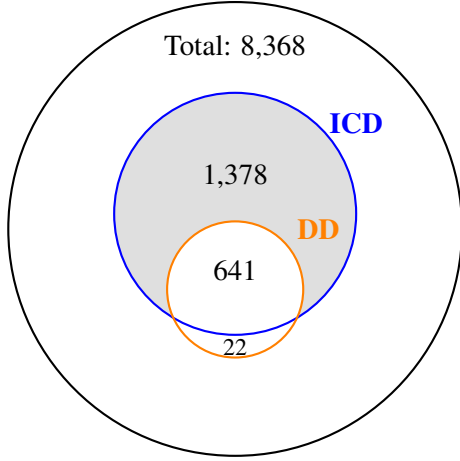


Figure 8: Venn diagram illustrating the relationship between ICD-coded Heart Failure (HF) ($D^{\text{ICD}} = 1$) and HF documented in the Discharge Diagnosis section ($D^{\text{DD}} = 1$) across 8,368 ICU admissions. The gray area represents the gray-zone between the ICD and Discharge Diagnosis labels.

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.

Discharge Diagnosis Extraction The disease terminology set was constructed using UMLS resources, clinical expertise, and manual inspection of hundreds of Discharge Diagnosis sections, and included common AKI surface forms and abbreviations such as *AKI*, *acute kidney injury*, *acute renal failure*, and *ARF*. To assess the recall of the string-matching procedure used to construct D^{DD} , we manually reviewed 400 discharge summaries, including 200 cases for which the matcher assigned $D^{\text{DD}} = 1$ and 200 cases for which it assigned $D^{\text{DD}} = 0$. Two authors reviewed all 400 cases to determine whether AKI was explicitly listed in the Discharge Diagnosis section. All 200 matcher-positive cases contained an explicit AKI mention, and no explicit AKI mentions were missed among the 200 matcher-negative cases.

KDIGO The KDIGO guidelines (Khwaja, 2012) 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 (retrospective KDIGO).

Selection of Attention Threshold To select the attention threshold, we analyzed the distribution of attention scores across tokens and observed that tokens in the higher attention range consistently corresponded to clinically meaningful AKI-related concepts, such as creatinine and dialysis, as also shown in Table 4. Based on this analysis, we selected a threshold of 2.5 to capture this high-attention region, where tokens are more likely to represent AKI-related clinical concepts. We further validated this observation through expert review of representative high-attention tokens.

PLM In our experiments, we used the RoBERTa-based encoder¹, 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 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.²

¹<https://github.com/facebookresearch/bio-lm>

²<https://www.cms.gov/medicare/coding-billing/icd-10-codes/icd-10-cm-icd-10-pcs-gem-archive>

D Results

D.1 Inter-Rater Agreement

To evaluate agreement among the four experts, we computed inter-rater agreement across the 150 adjudicated admissions. Overall agreement was 85.3% with Fleiss’ $\kappa = 0.798$. For the non-gray zone, agreement was 88.3% with $\kappa = 0.844$, while for the gray-zone, agreement was 83.3% with $\kappa = 0.754$. These values indicate near-perfect agreement, particularly given that AKI diagnosis is known to be clinically challenging.

D.2 PLM Interpretability Analysis

We performed this analysis on ICU test set 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 Appendix A. 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 ICU test set. 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.3 Evaluation of PLM Model on Top 50 ICD Codes

As we mentioned earlier, we constructed an expanded training cohort including both ICU and non-ICU cases for training 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 the ICU test set, which consists of admissions with ICU stays. One option is to train the model on the ICU-only subset of the training and validation split of the expanded training cohort ($N = 33,472$), as this subset shares the same distribution as ICU test set. Alternatively, we could train the model on the full training and validation split of the expanded training cohort, which includes both ICU and non-

ICU admissions ($N = 331,604$). Although this full dataset is distributionally biased relative to the ICU test set, it offers a substantially larger number of admissions.

We trained the model under two settings: (1) using the full training and validation split of the expanded training cohort, and (2) using the ICU-only subset of this split. We then evaluated both models on the ICU test set and computed the F1-score for each of the 50 ICD codes. The results are presented in Figure 9.

We observe that for all ICD codes, the model trained on full training split of the expanded training cohort consistently outperforms the model trained only on ICU cases. The macro F1-score of the expanded-training 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 expanded-training model versus 0.77 for the ICU-only model. These results confirm our hypothesis that using expanded training data for training the PLM model was beneficial.

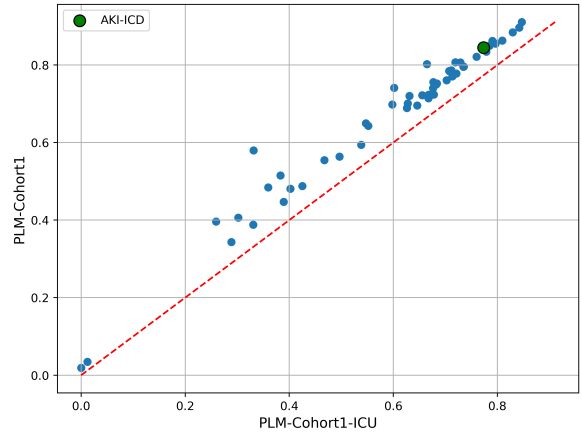


Figure 9: Comparison of PLM model performance when trained on the full training split of the expanded training cohort versus its ICU-only subset, evaluated on the ICU test set. Results are shown across the 50 most frequent ICD codes based on F1-score.

D.4 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 the ICU test set where both AKI-DD and AKI-ICD labels were positive. As shown in Figure 10, the attention distributions vary across tokens but consistently show that a substantial proportion receive high attention scores (greater than 2.5). However, some

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

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

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.

F Code Availability

The implementation of PLM-PDD and the code used for the experiments are publicly available at <https://github.com/vmahzoon76/PLM-PDD>.

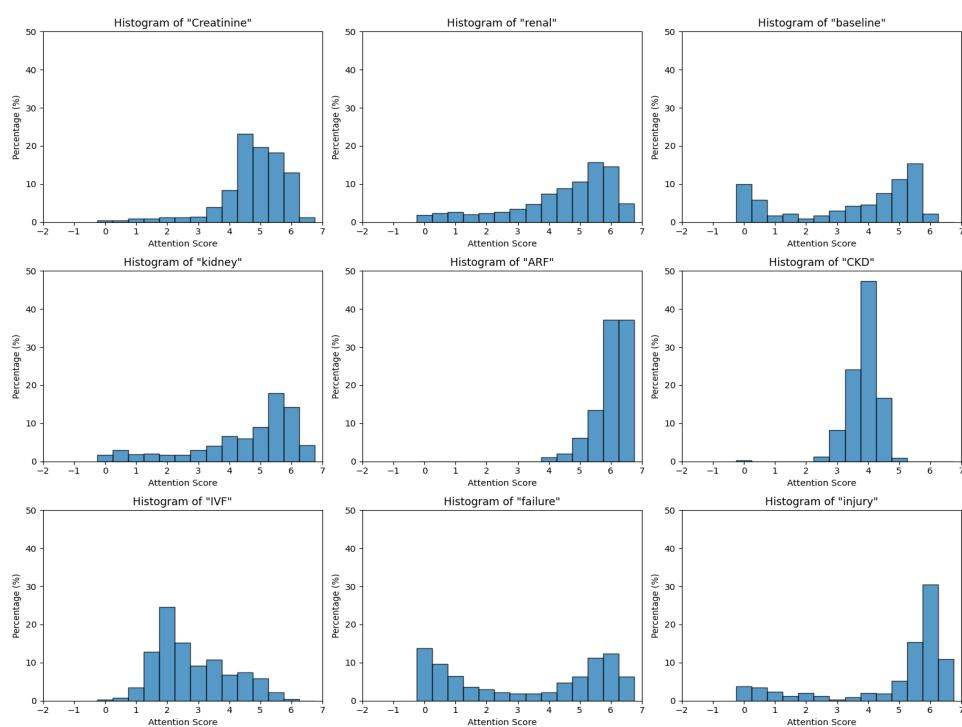


Figure 10: Attention score distribution for a few AKI-related tokens for PLM-PDD Model