Less is More: Explainable and Efficient ICD Code Prediction with Clinical Entities

James C. Douglas[1], Yidong Gan[1,2], Ben Hachey[1,3], Jonathan K. Kummerfeld[1]
[1]The University of Sydney [2]CSIRO Data61 [3]Beamtree
jonathan.kummerfeld@sydney.edu.au

Abstract

Clinical coding, assigning standardized codes to medical notes, is critical for epidemiological research, hospital planning, and reimbursement. Neural coding models generally process entire discharge summaries, which are often lengthy and contain information that is not relevant to coding. We propose an approach that combines Named Entity Recognition (NER) and Assertion Classification (AC) to filter for clinically important content before supervised code prediction. On MIMIC-IV, a standard evaluation dataset, our approach achieves near-equivalent performance to a state-of-the-art full-text baseline while using only 22% of the content and reducing training time by over half. Additionally, mapping model attention to complete entity spans yields coherent, clinically meaningful explanations, capturing coding-relevant modifiers such as acuity and laterality. We release a newly annotated NER+AC dataset for MIMIC-IV, designed specifically for ICD coding. Our entity-centric approach lays the foundation for more transparent and cost-effective assisted coding.

1 Introduction

Clinical coding is the process of translating free-text patient notes into standardized diagnostic and procedural codes, typically using the World Health Organization's International Classification of Diseases (ICD) (Dong et al., 2022). ICD codes serve important functions in billing, reimbursement, hospital planning, and epidemiological research. However, the coding process is labor-intensive and prone to delays, leading to global backlogs (Alonso et al., 2020; Campbell and Giadresco, 2020).

Assisted coding approaches seek to enhance the speed and reliability of code assignment. Recent work frames the task as an extreme multi-label classification problem, using deep neural models with label-wise token attention (Mullenbach et al., 2018; Vu et al., 2021; Huang et al., 2022). However, most approaches rely on full-length discharge summaries, which are typically lengthy, unstructured, and contain redundant or irrelevant information, a situation sometimes referred to as "note bloat" (Searle et al., 2021; Liu et al., 2022a). This redundancy inflates memory requirements and increases the risk of spurious correlations during training.

Discharge summaries also contain negation, speculation, and institution-specific style variations, which further introduce noise and complicate model training. Transformer-based models, which achieve state-of-the-art performance in ICD coding tasks (Huang et al., 2022; Edin et al., 2023), are particularly hampered by this noise, as their performance scales quadratically with sequence length. Although various architectural refinements have been explored, noise reduction strategies have not been extensively investigated for ICD coding.

In this paper, we propose a text-processing pipeline that leverages clinical Named Entity Recognition (NER) and Assertion Classification (AC) to retain only medical mentions that are relevant to coding. By discarding negated, hypothetical, or otherwise excluded concepts, we aim to further improve the signal-to-noise ratio during supervised classifier training. Our approach yields (1) **comparable coding performance** to a full-text baseline on MIMIC-IV (Goldberger et al., 2000; Johnson et al., 2023a), (2) **faster training** by more than 50%, (3) a **large input length reduction** of approximately 78%, and (4) **improved interpretability** when using entity spans as code evidence.

To enable our approach, we annotate a subset of MIMIC-IV notes using an ICD-aligned schema, capturing disorders, findings (normal/abnormal), procedures, medications and health context, including relevant coding modifiers (e.g., laterality and acuity) within the same span. Additionally, we annotate the assertion statuses of a subset of entities. We release this new dataset to facilitate further work on entity-driven clinical coding.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30835–30847

2 Related Work

Computer-assisted coding tools aim to expedite the coding and auditing process by providing recommendations and guidance (Campbell and Giadresco, 2020). Initially, code prediction models employed rule-based algorithms and traditional machine learning methods reliant on handcrafted features (Kavuluru et al., 2015; Farkas and Szarvas, 2008). As larger Electronic Health Record (EHR) datasets became available, such as MIMIC-III (Goldberger et al., 2000; Johnson et al., 2016) and MIMIC-IV (Goldberger et al., 2000; Johnson et al., 2023b), deep neural networks have become the dominant paradigm for ICD code prediction.

Discharge summaries, which document a patient's hospital stay, constitute the primary source for assigning ICD codes. These documents typically include extensive administrative, narrative, or test-related details, with as little as 10% estimated to be relevant to coding (Zhou et al., 2021). This redundancy underscores the value of filtering out extraneous information. Prior work has explored removing duplicated text, observing negligible or even positive impacts on model performance (Liu et al., 2022a). Our approach extends this line of research by preserving only clinically relevant mentions, further improving the signal-to-noise ratio.

More recently, transformer-based methods, including those employing pre-trained language models (PLMs), have shown competitive results on ICD coding benchmarks. PLM-ICD, which demonstrates state-of-the-art performance over other leading coding models (Edin et al., 2023), couples a medical PLM (RoBERTa-PM (Lewis et al., 2020)) with label-wise attention, which directs model attention to relevant segments of the text for each potential code (Huang et al., 2022). Beyond clinical coding, PLM-ICD achieves strong performance in other extreme multi-label classification tasks in law, news, and general science (Li et al., 2024).

Our work uses PLM-CA, a variant of PLM-ICD that modifies the label-wise attention layer to use a transformer-style cross-attention layer between token and label representations, improving training stability and performance (Edin et al., 2024).

In contrast, NER and AC have been relatively less studied tools for clinical coding. Clinical NER is the process of identifying key medical concepts in text, such as diseases, procedures, and medications, while AC determines the status of these entities (e.g., present, absent, or possible). Clinical entities have proven beneficial in several applications, including epidemiology, predictive modeling, note search, and patient cohort identification (Spasić et al., 2015; Bardak and Tan, 2021; Doerstling et al., 2022; Macri et al., 2022, 2023; Bean et al., 2023; Jani et al., 2024; Kraljevic et al., 2024).

Several approaches to incorporate NER and AC for clinical coding have been attempted, with mixed results. Nath et al. (2023) used NER and AC on the MIMIC-III top 50 split (Mullenbach et al., 2018), assigning codes based on the cosine similarity between static entity embeddings and ICD code descriptions. The approach underperformed PLM-ICD by around 20 percentage points in both micro and macro F1. While the entities in that work were considered in isolation from one another, our work processes entities in groups, preserving their order of appearance in a note. Certain ICD codes depend on interactions across co-morbid conditions, hence preserving surrounding context is crucial.

Meanwhile, DeYoung et al. (2022) extracted disease entities along with neighboring tokens, capturing modifiers such as acuity and severity. These concepts were linked to ICD codes using the Amazon Comprehend Medical API, then re-ranked using a proprietary model. This entity-centered approach outperformed a full-text baseline (CAML) (Mullenbach et al., 2018), highlighting the utility of entities with local context. Similarly, our NER model is trained to extract modifiers (e.g., anatomy, laterality, acuity), implicitly preserving the detail needed for accurate code assignment.

Interpretability remains a major concern in coding applications, where "black box" models such as PLM-CA output code probabilities without supporting evidence. MDACE addresses this by annotating code evidence spans in a subset of MIMIC-III discharge summaries (Cheng et al., 2023). Attention weights have previously been used to highlight the tokens that most influenced a given code prediction (Mullenbach et al., 2018; Liu et al., 2022b). More recently, combining attention weights with a gradient-based method (InputXGrad), termed AttnInGrad, has been shown to produce more plausible evidence than either method alone (Sundararajan et al., 2017; Edin et al., 2024). We use AttnInGrad to identify code-relevant tokens and map them to entities for more coherent explanations.

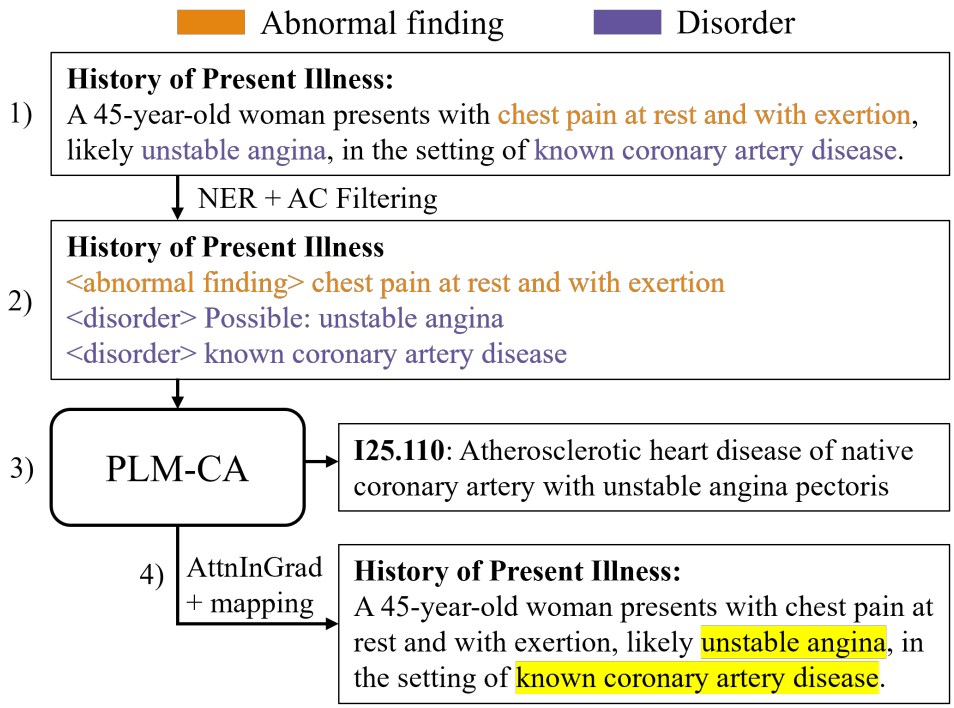

Figure 1: Overview of our entity-based approach to ICD coding. Discharge summaries are processed using NER+AC to yield shorter, clinically relevant representations for model training and inference.

3 Proposed Approach

3.1 Overview

Figure 1 illustrates our overall approach for extracting relevant clinical entities and generating ICD code predictions. In panel (1), we take a discharge summary and apply clinical NER to identify medical concepts and label them according to type (e.g., "abnormal finding" or "disorder"). We also run AC on each entity, filtering out those that are, for example, negated or hypothetical. In panel (2), the remaining entities are concatenated (along with headings) in their original order, forming a concise "entity-only" version of the note. In (3), this consolidated representation is used for training and inference with a PLM-CA model. Finally, AttnInGrad is applied to identify important tokens for each code's prediction (4). These tokens are then mapped to their full entity spans to enhance interpretability.

3.2 Document Processing

Prior to NER, each discharge summary is cleaned and segmented using the MedspaCy extension (v1.2.0) for spaCy (v3.7.6). Document-level preprocessing steps include hard-wrap and whitespace removal, sentence splitting and section heading detection. For faster NER inference, batches of consecutive sentences are processed together, provided the total token count remained under RoBERTa-PM's maximum input length of 512 tokens.

3.3 Named Entity Recognition and Assertion Classification

3.3.1 Named Entity Recognition

We train a RoBERTa-PM model (RoBERTa pretrained on PubMed, PMC and MIMIC-III (Lewis et al., 2020)) with a standard BIO scheme to detect and label six categories:

- **Normal Finding**: normal or physiologic observations.
- **Abnormal Finding**: pathological observations, test results, or symptoms.
- **Disorder**: underlying pathologies or etiologies.
- **Procedure**: diagnostic or therapeutic interventions.
- **Health Context**: broader contextual factors relevant to Chapters V-Z of ICD-10-CM (e.g., socioeconomic status, allergies, injury background).
- **Medication**: brand or chemical drug names.

The annotations are trained to be granular, including modifiers such as acuity (e.g., "acute"), laterality (e.g., "left"), or anatomical site (e.g., "lower lobe") within each entity's span.

3.3.2 Assertion Classification

Each recognized entity is assigned one of five statuses: *present*, *absent*, *possible*, *hypothetical*, or *not associated with the patient*. We adopt an approach from van Aken et al. (2021), in which each entity is enclosed in special tags (e.g., `<entity>`) and reinserted into its source sentence for classification. For this, we again use RoBERTa-PM as a backbone. Entities marked *absent* or *hypothetical* are discarded, while *possible* diagnoses, abnormal findings, and health contexts are retained per ICD-10-CM inpatient guidelines (Centers for Medicare & Medicaid Services (U.S.) and National Center for Health Statistics (U.S.), 2023). Entities labeled *not associated with the patient* are also kept if they are disorders or abnormal findings, as these often denote family history. For procedures and medications, only entities marked *present* are preserved.

3.4 Filtering and Reformatting

Finally, entities marked as *normal finding* are removed, as these are not coded in ICD-10-CM. After AC and filtering, we reinsert the retained entities into a consolidated *entity-only* document. Each entity is prefixed with a learnable token indicating its category (e.g., `<disorder> Diabetes`). If the entity is marked *possible*, we add a textual indicator ("Possible:") to convey uncertainty; if it is *not associated with the patient*, we presumptively prepend "Family history:", e.g., `<disorder> Family history: lung cancer`. Section headings are retained and standardized based on their category (e.g., "Past Medical History," "Physical Exam"), and entities appear in their original sentence order to preserve local context.

3.5 ICD Code Prediction Model

We use PLM-CA (Edin et al., 2024), a variant of PLM-ICD (Huang et al., 2022), for final code prediction.[1] This model was chosen for its open-source status, state-of-the-art performance, and customizability. In brief, PLM-CA employs a RoBERTa-PM encoder with a label-wise cross-attention layer. Each ICD code is represented by a trainable embedding that attends to relevant tokens in the text, producing the probability of that code being assigned.

3.6 Evidence Extraction

For evidence extraction, we use `AttnInGrad` (Edin et al., 2024), which combines token attention weights with a gradient-based measure (*InputX-Grad*) for more robust attributions. When applied to the *entity-only* model, any high-attribution token within an entity span is mapped back to the entire entity. For example, if "Heart" is identified as the top token for a code, then the entire entity "Heart failure with preserved ejection fraction" is flagged as supporting evidence.

4 Experimental Setup

4.1 Datasets

We conduct our experiments on MIMIC-IV (version 2.2), a large publicly available research dataset of de-identified EHR notes from the Beth Israel Deaconess Medical Center (Goldberger et al., 2000; Johnson et al., 2023a). We employ the train/validation/test splits from Edin et al. (2023). We focus on ICD-10-CM (diagnosis) and ICD-10-PCS (procedure) codes documented in MIMIC-IV, which are more contemporary and granular than the ICD-9 codes used in MIMIC-III.

4.2 ICD-Guided Entity Annotation

While prior clinical NER datasets (e.g., i2b2 2010) feature broad "problem", "test", or "treatment" labels (Özlem Uzuner et al., 2011), ICD coding often requires more granularity (e.g., distinguishing underlying etiologies from signs/symptoms) and capturing important modifiers (e.g., the affected body part and severity). Additionally, Chapters V-Z of ICD-10-CM require coverage of external factors such as socioeconomic status and the activities preceding an injury. We therefore developed an ICD-10 grounded annotation schema prioritizing specific entity types with an explicit inclusion of modifiers within each entity's span.

We randomly sampled 400 discharge summaries from MIMIC-IV (stratified by encounter type) and annotated them using Label-Studio (v1.15). One of the authors, who holds a medical degree, served as annotator, labeling entities as *Normal Finding*, *Abnormal Finding*, *Disorder*, *Procedure*, *Health Context*, and *Medication* in a single pass and capturing relevant modifiers within the same span. A

[1] Readers interested in full technical details are encouraged to consult Huang et al. (2022) and Edin et al. (2024).

	Median	IQR
Entities per document	207	[137, 273]
Words per entity	2	[1, 3]
Entity Labels		**Count**
Normal Finding		14,362
Abnormal Finding		26,463
Disorder		11,197
Procedure		7,665
Health Context		7,275
Medication		18,366
Total		85,328
Assertion Labels		
Present		5,180
Absent		2,187
Possible		1,152
Hypothetical		1,796
Not associated with patient		607
Total		10,922

Table 1: Annotation statistics for the MIMIC-IV NER + AC subset.

subset of these entities also received an assertion label from *present, absent, possible, hypothetical, not associated with the patient*. More detailed annotation guidelines and examples are given in Appendix A. Table 1 summarizes key statistics for our 400-note dataset. Notably, the frequency of normal, negated, and hypothetical entities underscores the extent of potentially non-relevant information.

Following the annotation guidelines, a second author annotated a random subset of documents containing roughly 1500 entities. Consistent with earlier clinical NER studies (Deleger et al., 2012), we assessed agreement against the original labels with exact-span F1 and Cohen's Kappa, yielding scores of 0.77 and 0.81, respectively.

To increase training diversity for assertion classification, we also incorporated publicly available data from the i2b2 2010/2012 challenges (Özlem Uzuner et al., 2011; Sun et al., 2013) and the MIMIC-III assertion dataset (van Aken et al., 2021).

4.3 Training Details and Baselines

NER & AC We train separate RoBERTa-PM models for NER and AC using cross-entropy loss, applying class-weighting in AC to mitigate class imbalance. Table 2 shows the shared hyperparameters applied.

Hyperparameter	Value
Batch size	8
Learning rate	3e-5
Weight decay	0.01
LR Scheduler	Linear scheduler with 10% warmup
Max token length	512
Early Stopping	On validation macro-F1

Table 2: Shared hyperparameters for the NER and AC models.

ICD code prediction We train PLM-CA for one run in two main conditions:

- **Full-text (Baseline)**: Entire discharge summaries, preprocessed as in Edin et al. (2024).
- **Entity-only**: The shortened notes from our pipeline.

Aside from a smaller batch size of 1 (with gradient accumulation) to accommodate GPU memory, we leave all PLM-CA hyper-parameters at their published default values (Edin et al., 2024). This setup helps us attribute any observed performance differences more directly to the input-preprocessing strategy. We train until validation Mean Average Precision (MAP) converges, then evaluate on the test split.

4.4 Ablation Experiments

We also explore several ablations on the *entity-only* approach to quantify each component's contribution:

1. **No Assertion Filtering:** Retain all extracted entities (including negated/hypothetical) to measure the impact of AC filtering.
2. **No Entity-Type Tokens:** Omit the special tokens that mark each entity's clinical category (e.g., <disorder>, <abnormal finding>).
3. **Plain Text Category Indicators:** Replace special tokens with plain-text equivalents (e.g., "Disorder:"), removing the need for learnable class tokens.
4. **Randomized Entity Order:** Shuffle entities and headings to test the importance of local context.

In addition, to quantify how each entity category influences coding performance, we retrain PLM-CA on subsets of entities. Mentions of Disorders and Procedures form the minimal subset as

they map directly to ICD-10-CM and ICD-10-PCS codes. From this baseline we test the addition of Abnormal Finding, Health Context, or Medication entities. For each configuration we report performance relative to the full *entity-only* model, along with the resulting median document length.

5 Results

5.1 Performance of NER & AC Models

Tables 3 and 4 report five-fold cross-validation performance for the NER and AC models, respectively. The NER system was recall-oriented, minimizing the number of entities missed by the downstream PLM-CA model. Most mis-labeling occurred within coding-relevant categories, particularly between Disorder and Abnormal Finding, and between Disorder and Health Context. Health context proved to be a difficult category to classify, being highly heterogeneous and often misclassified as disorder or procedure.

Entity Type	Precision	Recall	F1
Disorder	82.4 ± 1.1	84.6 ± 1.0	83.5 ± 0.9
Abnormal Finding	83.8 ± 0.6	87.8 ± 0.6	85.7 ± 0.5
Normal Finding	86.8 ± 0.9	89.6 ± 0.9	88.1 ± 0.8
Medication	85.8 ± 1.2	88.3 ± 0.4	87.0 ± 0.8
Health Context	67.8 ± 1.7	71.3 ± 1.3	69.5 ± 1.3
Procedure	77.8 ± 0.6	82.1 ± 1.1	79.9 ± 0.8

Table 3: NER model performance (exact span match), averaged over five cross-validation splits.

For AC, results were generally high across most labels. However, *hypothetical* and *possible* entities were commonly misclassified as *present*, suggesting that borderline or ambiguous samples are challenging to distinguish.

Assertion Type	Precision	Recall	F1
Present	96.7 ± 0.3	96.8 ± 0.3	96.7 ± 0.2
Absent	96.4 ± 0.2	96.6 ± 0.4	96.5 ± 0.2
Possible	85.3 ± 1.5	84.9 ± 1.1	85.1 ± 1.0
Hypothetical	90.6 ± 1.9	90.5 ± 1.2	90.5 ± 1.2
Not Associated with Patient	96.1 ± 2.0	95.3 ± 2.1	95.6 ± 1.1

Table 4: Assertion classification model performance, averaged over five cross-validation splits.

5.2 Document Length Reduction

Table 5 shows the median and IQR for document length in words before and after entity-based filtering, computed across all documents. Overall, our entity-driven pipeline reduces the median document length from 1,627 to 353 words, representing a roughly 78% reduction in document length. Aggregated across the entire dataset, *entity-only* notes total 47.8M words (versus 212.3M words in the full-text set), a 77.5% reduction.

	Median per note	IQR
Full-text	1627	[1245, 2104]
Entity-only	353	[221, 509]

Table 5: Median document length before and after entity filtering (words).

5.3 Training Efficiency

We measure training time on a single NVIDIA L4 GPU. Table 6 reports the hours per epoch and the total number of epochs until convergence. The *entity-only* approach reduces epoch time by over 50%. While NER+AC inference introduces some overhead, it is applied only once per document.

	Time/Epoch	Epochs	Total (hrs)
Full-text	3.20 hr	9	28.8
Entity-only	1.36 hr	8	10.9

Table 6: Training time per epoch for PLM-CA on MIMIC-IV.

5.4 Performance

We evaluated ICD-10-CM/PCS classification performance using standard metrics from prior research: micro- and macro-averaged AUC-ROC and F1, as well as Exact Match Ratio (EMR), Precision@k, and Mean Average Precision (MAP). Permutation testing (1000 rounds) was applied to the prediction outcomes to test whether the differences between models are statistically significant. Results are presented in Table 7.

Overall, the *entity-only* variant underperforms the *full-text* baseline on both classification and ranking metrics but remains close to the original PLM-ICD model from Edin et al. (2023). The greatest discrepancy is in macro F1, suggesting a difference on rarer codes. We found that each condition excels on different codes; however, no clear semantic or categorical patterns emerged among these discrepancies.

5.5 Evidence Extraction Results

We evaluate *AttnInGrad* explanations on the MDACE dataset (Cheng et al., 2023), which pro-

	Classification					Ranking		
	AUC-ROC		F1		EMR	Precision@k		MAP
	Micro	Macro	Micro	Macro		8	15	
PLM-ICD (Edin et al., 2023)	99.2	96.6	58.5	21.1	0.40	69.9	55.0	61.9
PLM-CA (Full-text)	99.2	96.2	59.6	26.1	0.44	70.9	56.1	63.4
PLM-CA (Entity-only)	99.1	95.9*	58.8*	24.8*	0.45	70.3*	55.3*	62.3*

Table 7: ICD-10-CM/PCS performance on the test set. EMR: percentage of documents with all codes predicted correctly. We include the original PLM-ICD results (full-text) from Edin et al. (2023) for reference (averaged over 10 seeds). Asterisks (*) denote statistical significance at $p<0.05$, following 1,000-round permutation tests comparing the *entity-only* model to the full-text model.

vides ground-truth evidence spans for 302 MIMIC-III discharge summaries. Code classification metrics for the models are described in Appendix B. The attribution threshold for AttnInGrad was tuned using the MDACE validation set to maximize partial span overlap with ground truth (F2 score), giving higher weight to recall, a priority in medical contexts to avoid missing any clinically salient evidence. We also reuse the code prediction thresholds derived from the MIMIC-IV evaluation.

Since the MDACE training set had not been used previously, it was combined with the test set to provide a more robust evaluation. To avoid conflating model accuracy with evidence extraction quality, we analyzed evidence spans only for true-positive predictions. False negative codes are likely exacerbated by coding style differences between MDACE and MIMIC. Only 45.1% of the codes assigned in MIMIC also appeared in MDACE (Cheng et al., 2023), and approximately 11% of MDACE codes were absent in the MIMIC-IV training data.

To accommodate the entity-based approach, position-invariant matching was performed along with normalization (lowercasing, punctuation/whitespace removal). For each code, the evidence was categorized as: *Exact Match* (identical to ground truth), *Superset* (the extracted span contains and expands the ground truth span), *Subset* (the span is a subset of the ground truth span), or *No Overlap*. Table 8 outlines these proportions for the two approaches.

	Exact Match	Superset	Subset	No Overlap
Full-text	65.4%	9.7%	19.5%	5.4%
Entity-only	65.3%	20.4%	6.9%	7.4%

Table 8: Proportion of evidence spans matching MDACE gold references (true positives only).

The *entity-only* model produces more *superset* spans, often capturing relevant modifiers or expansions overlooked in the MDACE annotations. Superset spans extended the ground truth span by a median of 2 words (IQR 3). Examples of these superset spans are in Table 9.

Meanwhile, the proportion of *subset* matches drops from 19.5% to 6.9% when using *entity-only* summaries, mitigating the risk of missing clinically important information. *No Overlap* cases (7.4% for *entity-only*) mostly involved synonyms or abbreviations of the gold span (58.1%), with the remainder being either equally relevant (21.3%), less relevant (13.7%), or insufficient (6.8%) for coding.

An error analysis was performed on 200 randomly selected false positive codes to examine the relevance of their evidence spans. 45% of evidence spans were deemed relevant for the predicted code, fully covering its description. 26.5% were partially relevant, covering part of the code's description. 17% were insufficient, oftentimes capturing the device or medication associated with a code, rather than the code itself. Finally, 11.5% of spans were related to a false negative code, representing situations where the model predicted the correct ICD code chapter, but the wrong specific code.

5.6 Ablation Results

Table 10 shows how various aspects of the pipeline affect performance. Removing assertion filtering or learnable entity category tokens mostly affects macro-F1, suggesting that these features help the model correctly identify rare codes. Randomizing entity order yields the largest drop, underscoring that local context is important.

Further, Table 11 summarizes coding outcomes for various entity subsets. The minimal subset of Disorder + Procedure trims the median note to

ICD Code	MDACE Evidence	Entity Evidence
M43.16: Spondylolisthesis, lumbar region	"spondylolisthesis"	"grade 1 spondylolisthesis at L4–L5"
I25.2: Old myocardial infarction	"MI"	"history of large anterior MI in past"
M75.02: Adhesive capsulitis of left shoulder	"frozen shoulder"	"left frozen shoulder"
C78.7: Secondary malignant neoplasm of liver and intrahepatic bile duct	"liver"	"Melanoma with metastatic disease to lung, bone, and liver"
J12.9: Viral pneumonia, unspecified	"Pneumonia"	"Viral Pneumonia"

Table 9: Examples of *superset* spans from the entity-based model, capturing extra ICD-relevant detail.

Condition	Δ F1 Micro	Δ F1 Macro	Δ MAP
No assertion classification	-0.1	-1.2	-0.2
No entity category prepending	-0.2	-1.2	-0.2
Entity classes as plain text instead of special tokens	-0.1	-1.5	-0.2
Randomized entity/heading order	-1.5	-3.0	-1.7

Table 10: Ablation results on the test set showing the performance change (Δ) relative to the baseline *entity-only* condition.

roughly one-third of the full *entity-only* input, albeit with a sizable drop in performance. Adding Health Context recovers nearly the same performance as Abnormal Findings while keeping the input 28% shorter, giving it the highest performance-per-token payoff. In contrast, Medications delivers only modest gains despite a similar input length to Health Context.

6 Discussion

In this study, we introduce an entity-focused approach for assisted ICD-10-CM/PCS coding that leverages clinical Named Entity Recognition (NER) and Assertion Classification (AC) to remove non-essential text from discharge summaries. Across our evaluations on MIMIC-IV, the PLM-CA (*entity-only*) model achieves performance comparable to a full-text baseline while discarding approximately 78% from the original note. The approach provides several benefits: (1) a substantial reduction in computational overhead, (2) competitive coding performance, and (3) straightforward mapping from model attributions to clinically meaningful entities.

Computational Efficiency Because PLM-CA employs a transformer-based encoder (RoBERTa-PM), its computational complexity scales quadratically with the number of tokens. Although this may be feasible for shorter documents, long clinical notes containing thousands of tokens can render training and inference computationally expensive. By focusing on clinically relevant entities, we reduce training time by more than half. This time efficiency is especially beneficial for hyperparameter tuning or rapid prototyping of new architectures, making large-scale experimentation more feasible and cost-effective.

Further, given the reliance on BERT/RoBERTa encoders with 512-token limits, the entity-based approach reduces reliance on workarounds such as pooling. The framework is also advantageous for including additional documents beyond discharge summaries, such as operation reports.

Coding Performance Despite differences from the full text baseline, the *entity-only* model maintains strong performance. The 1-2 percentage point difference in metrics surpasses findings from an ablation experiment by Wiegreffe et al. (2019), where dictionary-based entity extraction led to a 9-percentage-point decline in micro F1 when substituting entities for raw text. Our findings suggest that a more flexible, neural-driven entity extraction process tailored to ICD coding better preserves predictive signals.

Beyond PLM-CA, the performance of the *entity-*

Condition	Δ F1 Micro	Δ F1 Macro	Δ MAP	Length
Disorders + Procedures	-6.7	-4.0	-8.7	112
Disorders + Procedures + Abnormal Findings	-3.6	-2.6	-4.5	241
Disorders + Procedures + Health Context	-3.5	-2.8	-4.5	173
Disorders + Procedures + Medications	-5.5	-5.0	-7.0	160

Table 11: Test set performance of entity category subsets, relative to the baseline *entity-only* condition. Length refers to median words per document.

only model is competitive with prior full-text models, mirroring or surpassing the performance of the original PLM-ICD model. This competitive performance, despite a major reduction in document length, reflects prior remarks about redundancy and 'bloat' in clinical documentation (Searle et al., 2021; Liu et al., 2022a).

Considering efficiency and performance on a Pareto frontier, the entity-based approach offers a cost-effective balance, sacrificing ~1 percentage point in F1 metrics for greatly increased efficiency. This balance may be desirable in resource-constrained clinical settings or large-scale production environments. Future work could explore additional entity selection or representation strategies to refine this balance.

Model Interpretability Explainability constitutes a major requirement in clinical decision-making settings. Our evaluation using MDACE shows that the entity-based model provides broader yet meaningful spans when mapping attributed tokens to their source entity. The entity-based evidence in many cases included additional modifiers or expansions needed for specific ICD code assignment. In contrast, models that operate on full-text inputs must deal with extracting and aligning disjointed subtokens, complicating the process of reassembling coherent explanations.

The completeness of entity spans is likely advantageous for downstream auditing and code validation. For false positive predictions, the corresponding entity evidence can be reviewed and either confirmed or discarded, helping coders identify overlooked conditions or refine incorrect codes. Importantly, a sizable portion of false positives in our analysis included relevant code evidence, highlighting a possible secondary benefit of detecting overlooked or under-specified codes.

7 Conclusion

We have shown that focusing on clinically relevant entities, coupled with assertion filtering, is an effective strategy to streamline assisted ICD-10-CM/PCS coding on MIMIC-IV. By compressing notes to around one-fifth of their original length, our approach reduces training cost, maintains competitive performance, and provides interpretable explanations. Future work could extend this approach by integrating direct entity-to-code linkage methods. In addition, the reduced input size can accommodate entities from other document sources (e.g., progress notes, surgical or radiology reports), offering a more comprehensive view of a patient encounter. These findings underscore the promise of entity-centric design for real-world clinical informatics, providing a solid foundation for further exploration and refinement. Our newly annotated dataset, aligned with ICD-10 needs, can facilitate further research in entity-based coding and serve as a general-purpose resource for clinical NLP.

8 Limitations

We acknowledge several limitations. First, while entity-based pruning retains vital details for most codes, errors in the NER model could result in contextual loss. Additionally, although our method improves the plausibility of model-generated evidence by linking attributions to entities, it does not directly address faithfulness. Second, while the newly created NER + AC dataset is designed to align with ICD coding conventions and underwent validation, it has not yet been validated by professional clinical coders. Third, although the MIMIC-IV population is large, it is geographically and institutionally specific to Beth Israel Deaconess Medical Center, limiting the generalizability of our findings to other healthcare settings, particularly those outside critical care.

9 Ethical Considerations

This research uses de-identified patient data (MIMIC-IV), which is publicly available under stringent guidelines. Tools based on data from a single institution risk misclassifying certain patient groups; therefore, training with diverse datasets is essential for fairness and accuracy in real-world deployments. Furthermore, any assistive coding tools leveraging approaches described in this work must be integrated responsibly into workflows, with oversight by professional coders and clinicians to minimize the risk of errors that could impact patient care and billing. This includes efforts to mitigate automation bias, where a user may accept code predictions without sufficient scrutiny.

Acknowledgments

This material is partially supported by the Australian Research Council through a Discovery Early Career Researcher Award. We also thank the anonymous reviewers for their constructive feedback on our submission.

References

Vera Alonso, João Vasco Santos, Marta Pinto, Joana Ferreira, Isabel Lema, Fernando Lopes, and Alberto Freitas. 2020. Problems and barriers during the process of clinical coding: a focus group study of coders' perceptions. *Journal of Medical Systems*, 44:1–8.

Batuhan Bardak and Mehmet Tan. 2021. Improving clinical outcome predictions using convolution over medical entities with multimodal learning. *Artificial Intelligence in Medicine*, 117:102112.

Daniel M. Bean, Zeljko Kraljevic, Anthony Shek, James Teo, and Richard J. B. Dobson. 2023. Hospital-wide natural language processing summarising the health data of 1 million patients. *PLOS Digital Health*, 2(5):e0000218.

Sharon Campbell and Katrina Giadresco. 2020. Computer-assisted clinical coding: A narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals. *Health Information Management Journal*, 49(1):5–18.

Centers for Medicare & Medicaid Services (U.S.) and National Center for Health Statistics (U.S.). 2023. Icd-10-cm official guidelines for coding and reporting.

Hua Cheng, Rana Jafari, April Russell, Russell Klopfer, Edmond Lu, Benjamin Striner, and Matthew Gormley. 2023. MDACE: MIMIC documents annotated with code evidence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7534–7550, Toronto, Canada.

Leilani Deleger, Qing Li, Todd Lingren, Melissa Kaiser, Kurtis Molnar, Lisa Stoutenborough, Mark Kouril, Kristin Marsolo, and István Solti. 2012. Building gold standard corpora for medical natural language processing tasks. In *Proceedings of the AMIA Annual Symposium*, volume 2012, page 144.

Jay DeYoung, Han-Chin Shing, Chris (Luyang) Kong, Christopher Winestock, and Chaitanya Shivade. 2022. Entity anchored icd coding. *Amazon Science*.

Steven S. Doerstling, Dennis Akrobetu, Matthew M. Engelhard, Felicia Chen, and Peter A. Ubel. 2022. A disease identification algorithm for medical crowdfunding campaigns: Validation study. *Journal of Medical Internet Research*, 24(6):e32867.

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.

Joakim Edin, Alexander Junge, Jakob D. Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. 2023. Automated medical coding on mimic-iii and mimic-iv: A critical review and replicability study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2572–2582, New York, NY, USA. Association for Computing Machinery.

Joakim Edin, Maria Maistro, Lars Maaløe, Lasse Borgholt, Jakob Drachmann Havtorn, and Tuukka Ruotsalo. 2024. An unsupervised approach to achieve supervised-level explainability in healthcare records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4869–4890, Miami, Florida, USA.

Richárd Farkas and György Szarvas. 2008. Automatic construction of rule-based icd-9-cm coding systems. *BMC Bioinformatics*, 9 Suppl 3(Suppl 3):S10.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

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, Seattle, WA.

Meghna Jani, Ghada Alfattni, Maksim Belousov, Lynn Laidlaw, Yuanyuan Zhang, Michael Cheng, Karim Webb, Robyn Hamilton, Andrew S. Kanter, William G. Dixon, and Goran Nenadic. 2024. Development and evaluation of a text analytics algorithm for automated application of national covid-19 shielding criteria in rheumatology patients. *Annals of the Rheumatic Diseases*, 83(8):1082–1091.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023a. Mimic-iv-note: Deidentified free-text clinical notes (version 2.2).

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023b. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10:1.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Ramakanth Kavuluru, Anthony Rios, and Yuan Lu. 2015. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial Intelligence in Medicine*, 65(2):155–166.

Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Baston, Jack Ross, Esther Idowu, James T. Teo, and Richard J. B. Dobson. 2024. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290.

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

Nan Li, Bo Kang, and Tijl De Bie. 2024. Your next state-of-the-art could come from another domain: A cross-domain analysis of hierarchical text classification. *arXiv preprint*.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2022a. "note bloat" impacts deep learning-based nlp models for clinical prediction tasks. *Journal of Biomedical Informatics*, 133:104149.

Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2022b. Hierarchical label-wise attention transformer model for explainable icd coding. *Journal of Biomedical Informatics*, 133:104161.

Carmelo Macri, Ian Teoh, Stephen Bacchi, Michelle Sun, Dinesh Selva, Robert Casson, and Weng Onn Chan. 2022. Automated identification of clinical procedures in free-text electronic clinical records with a low-code named entity recognition workflow. *Methods of Information in Medicine*, 61(3-04):84–89.

Carmelo Z. Macri, Sheng Chieh Teoh, Stephen Bacchi, Ian Tan, Robert Casson, Michelle T. Sun, Dinesh Selva, and Weng Onn Chan. 2023. A case study in applying artificial intelligence-based named entity recognition to develop an automated ophthalmic disease registry. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 261(11):3335–3344.

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 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana.

Namrata Nath, Sang-Heon Lee, and Ivan Lee. 2023. Application of specialized word embeddings and named entity and attribute recognition to the problem of unsupervised automated clinical coding. *Computers in Biology and Medicine*, 165:107422.

Thomas Searle, Zina Ibrahim, James Teo, and Richard Dobson. 2021. Estimating redundancy in clinical text. *Journal of Biomedical Informatics*, 124:103938.

Irena Spasić, Bo Zhao, Christopher B. Jones, and Kate Button. 2015. Kneetex: an ontology–driven system for information extraction from mri reports. *Journal of Biomedical Semantics*, 6:34.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Betty van Aken, Ivana Trajanovska, Amy Siu, Manuel Mayrdorfer, Klemens Budde, and Alexander Loeser. 2021. Assertion detection in clinical notes: Medical language models to the rescue? In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 35–40, Online.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2021. A label attention model for icd coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Sarah Wiegreffe, Edward Choi, Sherry Yan, Jimeng Sun, and Jacob Eisenstein. 2019. Clinical concept extraction for document-level coding. In *Proceedings of*

the 18th BioNLP Workshop and Shared Task, pages 261–272, Florence, Italy.

Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. 2021. Automatic ICD coding via interactive shared representation networks with self-distillation mechanism. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5948–5957, Online.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

A Annotation Guidelines

The following guidelines outline the standards employed for annotating discharge summaries from MIMIC-IV. Our annotation schema is motivated by the ICD-10-CM/PCS coding framework, ensuring consistency and clinical relevance.

A.1 Entity Categories

We define six primary entity categories: **Normal Finding**, **Abnormal Finding**, **Disorder**, **Health Context**, **Medication**, and **Procedure**. Each category is annotated with neighboring contextual information (e.g., severity, laterality) to support downstream modeling.

A.1.1 Normal Findings

A *Normal Finding* is a physiologically normal observation or test result that would not usually map to an ICD-10-CM code.

- *"The patient is normotensive."*
  Entity: "normotensive"
- *"Normal bowel sounds were auscultated."*
  Entity: "Normal bowel sounds"

A.1.2 Abnormal Findings

An *Abnormal Finding* is any atypical or pathological observation, including explicit statements of abnormal test results. Numeric results alone do not qualify, unless explicitly described as abnormal. When there is a known etiological cause, annotate the cause as a **Disorder**, and the manifestation as an **Abnormal Finding**.

- *"Elevated white blood cell count was noted."*
  Entity: "Elevated white blood cell count"
- *"The patient exhibited jaundice."*
  Entity: "jaundice"

A.1.3 Disorders

A *Disorder* is the underlying pathology or etiology of a clinical finding.

- *"The patient has viral gastroenteritis causing vomiting."*
  Entity: "viral gastroenteritis" (Disorder), "vomiting" (Abnormal Finding)
- *"Hypertension is responsible for the patient's headaches."*
  Entity: "Hypertension" (Disorder), "headaches" (Abnormal Finding)

A.1.4 Health Context

Health context captures contextual circumstances that impact health, such as socioeconomic factors, personal medical history, allergies, injury context, and code status. These often appear following expressions such as "history of" or "status post (s/p)", which should also be included in the entity span.

- *"The patient's condition was due to a motor vehicle accident."*
  Entity: "motor vehicle accident" (Health Context)
- *"The patient is DNR."*
  Entity: "DNR" (Health Context)
- *"The patient smokes 2 packs of cigarettes daily."*
  Entity: "smokes 2 packs of cigarettes daily" (Health Context)
- *"The patient has a history of renal transplant."*
  Entity: "history of renal transplant" (Health Context)

A.1.5 Medications

Any mention of a drug (brand name or generic) together with a relevant action (e.g., "discontinued," "started," "titrated"). Administration route details are included in the span where possible. Dosage numbers are omitted **unless** they appear between the drug name and its action or route or administration.

- *"The patient was started on atorvastatin."*
  Entity: "started on atorvastatin"
- *"Ibuprofen was held."*
  Entity: "Ibuprofen was held"

A.1.6 Procedures

A *Procedure* is any diagnostic or therapeutic intervention (including named devices used during the intervention). Routine laboratory tests are excluded.

- *"Patient underwent angioplasty."*
  Entity: "angioplasty"
- *"The team performed an elective laparoscopic cholecystectomy."*
  Entity: "elective laparoscopic cholecystectomy"

A.2 General Annotation Principles

Entity Spans and Contextual Modifiers. Include relevant modifiers (e.g., acuity, anatomy, laterality) within the entity span.

- *"The patient has acute right lower lobe pneumonia."*
  Entity: "acute right lower lobe pneumonia" (Disorder)
- *"The test showed severe anemia."*
  Entity: "severe anemia" (Abnormal Finding)

Ambiguity between *Abnormal Finding* and *Disorder*: When uncertain, default to Abnormal Finding.

Ambiguity between *Normal Finding* and *Abnormal Finding*: If an ICD-10-CM code can be found for the span, treat it as Abnormal Finding.

Negated or speculative mentions: Still annotate the entity, but do **not** include the negation or speculation cue in the span.

Entity boundary uncertainty: When unsure whether a span contains multiple entities, split it into separate entities.

A.3 Assertion Categories

Each entity is assigned one of the following assertion labels:

- **Present**: Directly stated as existing for the patient.
- **Absent**: Explicitly negated (e.g., "No evidence of pneumonia," "Denies chest pain").
- **Possible**: Mentioned as uncertain or probable (e.g., "Probable pneumonia," "Could be myocardial infarction").
- **Hypothetical**: Speculative or future-oriented (e.g., "If pneumonia were to occur," "Considering intubation").
- **Not associated with patient**: Mention pertains to someone else (e.g., "Family history of diabetes," "Mother has osteoporosis").

B MDACE Classification Performance

Classification performance was measured on the combined MDACE train and test sets, using prediction thresholds from the MIMIC-IV evaluation.

Model	Precision	Recall	F1
Full-text	47.3	64.8	54.7
Entity-only	41.8	65.9	51.1

Table 12: Micro-averaged performance of the *full-text* and *entity-only* models on the MDACE test set.