Evidence-Grounded LLM Validation of MIMIC-IV ICD Labels

Ahmad Abu Dayeh[1], Hajira Jabeen, and Oya Beyan
University of Cologne
[1]ORCiD ID:https://orcid.org/0009-0009-5665-680

Abstract. Automatically assigning ICD-10 diagnosis codes from discharge summaries is a central multi-label task in clinical NLP, yet widely used benchmarks such as MIMIC contain substantial label noise: many charted codes are not text-grounded in the note or are mis-specified. We present an LLM-based evidence-validation study that examines each (note, code) pair and: (1) determines whether the code is supported by the note, (2) extracts the corresponding evidence quote(s) from the note, and (3) when evidence exists but a more appropriate code can be inferred, suggests an evidence-based replacement. Applying this pipeline to 10,000 notes from the MIMIC-IV corpus, we derive two refined label sets: Evidence-Verified (EV), retaining only text-supported codes, and Evidence-Replaced (ER), substituting some existing codes in EV with better evidence-supported alternatives. We replicate six state-of-the-art ICD coding models (PLM-ICD, LAAT, MultiResCNN, CAML, Bi-GRU, CNN) under identical settings as Edin et al. and evaluate micro-precision, recall, and F1 using paired bootstrap resampling. Our validation study results show that removing unsupported charted codes from MIMIC substantially improves model performance and yields more trustworthy benchmarks for automated medical coding.

Keywords. ICD-10 coding, MIMIC-IV, label noise, large language models

1. Introduction

Automatic assignment of ICD-10 diagnosis codes from free-text discharge notes is a high-impact multi-label clinical NLP task: it supports billing, quality assessment, and downstream research. Progress on this task depends heavily on public benchmarks such as MIMIC [1], yet these datasets are known to contain substantial label noise [2]. Many charted codes are not grounded in the discharge narrative. They might have been introduced by administrative workflows or by information recorded elsewhere, or they are either inappropriately specific or overly broad relative to the note's text. For example, the note states "Follow-up after oncology consultation; patient scheduled for tumor board discussion," yet the chart lists C34.9 (malignant neoplasm of lung, unspecified). While the diagnosis may have been made elsewhere, the discharge summary itself provides no pathological or radiologic evidence of malignancy. Such labels lead models to associate generic phrases like "oncology follow-up" with specific cancer codes, confounding model training through noisy supervision and evaluation through biased performance estimates [2].

[1] Corresponding Author: Ahmad Abu Dayeh, Email: aabudaye@uni-koeln.de.

To address this, we present a Large Language Model (LLM)-based evidence-validation study that examines each (note, code) pair and performs three tasks: (1) determines whether the code is supported by the note, (2) extracts the corresponding evidence quote(s) from the note, and (3) when evidence exists but a more appropriate code can be inferred, suggests an evidence-based replacement accompanied by a brief rationale. Applying this pipeline to a 10,000-note subsample of the MIMIC-IV dataset, we generate two alternative label sets derived from the original MIMIC labels: an Evidence-Verified (EV) label set *retaining* only codes with explicit textual support, and an Evidence-Replaced (ER) set derived from EV, by *substituting* some existing codes in EV with better evidence-supported alternatives. We reuse Edin et al. [3]'s preprocessing and replicate six state-of-the-art automated ICD coding models to measure how evaluation changes when label evidence is taken into account.

In this paper, we investigate how micro-F1, precision, and recall differ across six replicated state-of-the-art models when trained and evaluated against (i) the Original MIMIC (OM) labels, (ii) the Evidence-Verified (EV) labels, which retain only codes supported by note evidence, and (iii) the Evidence-Replaced (ER) labels, which adjust EV with LLM-suggested substitutes. By quantifying these differences, we assess whether published performance on MIMIC reflects true extractable clinical evidence in notes, and we estimate the impact of label quality on model benchmarking. The results aim to guide researchers and practitioners toward more robust evaluation practices and clearer interpretations of automated medical coding performance.

2. Methods

Figure 1. Overview of the two main workflows.

For our validation study, as shown in Figure 1, we conducted two linked workflows: (1) Label Refinement — an LLM-based validation and correction pipeline that produced refined ICD label sets; and (2) Model Replication — where six well-known ICD code prediction models (PLM-ICD, LAAT, MultiResCNN, CAML, Bi-GRU, and CNN) were trained and tested separately on both the original and refined labels. All models used the MIMIC-IV discharge notes and followed the preprocessing pipeline of Edin et al. [3] for consistency. Two key deviations were introduced: model training and testing were performed on a 10,000-note subsample rather than the full dataset, and the target labels were limited to ICD-10-CM diagnosis codes, excluding procedure codes.

The full code, reproducible pipelines, and results are publicly available on GitHub: https://github.com/aabudayeh/MIMIC_LLM__Validation

2.1. Data sourcing and preprocessing

Source data were MIMIC-IV discharge notes with their associated ICD-10-CM diagnosis codes (v2.2). We followed Edin et al. [3]'s reproducible preprocessing pipeline for tokenization, variable labelling, and note-level label mapping. From the processed corpus (~122k notes), we sampled a 10,000-note subset for LLM review. We filtered notes based on diagnosis code count, retaining only those with 5–10 codes to ensure meaningful clinical content and adherence to the LLM's token limits.

2.2. Label Refinement

Figure 2. Comparison of target label sets for the 9,528-note samples.

Each note–code pair in the 10,000-note subsample of MIMIC-IV was reviewed by the LLM(DeepSeek-V3-0324) configured to act as a clinical coding expert. The model received as input the full discharge summary, the set of originally assigned ICD-10-CM codes for that note, and candidate alternative codes drawn from the same top-level ICD-10-CM category to support fine-grained differentiation. For each charted code, the LLM produced: (a) an evidence judgment indicating whether the note text supports the code; (b) one or more evidence quotes directly extracted from the note; and (c) when evidence existed but a more appropriate code could be inferred. These structured outputs were parsed and stored programmatically in JSON format.

Using the LLM outputs, we derived two refined label sets in addition to the original MIMIC (OM) labels (Figure 2). The Evidence-Verified (EV) set retains only codes with explicit textual support. The Evidence-Replaced (ER) set is derived from EV by substituting certain codes with more precise, evidence-supported alternatives. All three label sets (OM, EV, ER) were saved at the note level for downstream experiments.

2.3. Model Replication

We replicated six established ICD-coding models (PLM-ICD, CAML, LAAT, MultiResCNN, Bi-GRU, and a CNN baseline) using the same preprocessing, patient-level data splits (train 73%, validation 11%, test 16%), and model architecture choices reported by Edin et al [3]. Each model was trained independently on each label set (OM, EV, ER) using the same training regimen per architecture, resulting in 6 × 3 = 18 total training runs.

2.4. Evaluation and statistical analysis

Models were evaluated on the held-out test partition using micro-precision, micro-recall, and micro-F1 scores, as these metrics account for class imbalance by aggregating predictions across all labels, a standard practice in multi-label ICD coding tasks. Each model was assessed against its own label set to quantify overall ecosystem improvement. Furthermore, to assess whether observed metric differences across label sets were statistically meaningful, bootstrap confidence intervals (2000 resamples) were computed for micro-F1, precision, and recall for all models except PLM-ICD. Paired bootstrap deltas with 95 % confidence intervals (CI) and two-sided p-values were used to assess significant differences between label sets ($\alpha = 0.05$).

3. Results

Figure 3. Comparison of Micro-Averaged Precision, Recall, and F1 Scores Across Models and Label Sets.

Figure 4. ΔPrecision, ΔRecall, and ΔF1 with 95% CIs ($p < 0.05$ marked with *).

Figure 3 summarizes the precision, recall, and F1 scores across models and label sets, while Figure 4 highlights the pairwise performance differences between each refined label set and the original MIMIC labels, indicating where changes are statistically significant. Across models, EV consistently improves micro-F1 and precision relative to OM, with recall generally higher except for RNN and CNN. In contrast, ER preserves most precision gains but often trades off recall, yielding higher F1 for LAAT and MultiResCNN, little net change for CAML, and lower F1 for RNN and CNN. The paired bootstrap results in Figure 4 confirm these patterns: EV-OM F1 improvements are statistically significant for all models, while ER-OM effects are significantly positive for LAAT and MultiResCNN, negative for RNN and CNN, and nonsignificant for CAML. PLM-ICD follows the same trend, achieving its highest scores under EV and a decline under ER.

4. Discussion

The performance ranking of the models in this study (PLM-ICD > LAAT > CAML > MultiResCNN > RNN > CNN) closely mirrors that of Edin et al. [3], suggesting that the observed trends are robust across data and label conditions. Across all models, moving from OM to EV consistently and significantly improves measured performance because EV strips away label assignments that are not text-grounded, reducing false positives and sharpening the learning signal. This manifests as universal precision gains and statistically significant F1 improvements across architectures, including PLM-ICD. These precision gains are expected: removing noisy, weakly supported codes eliminates many false positives, making models more conservative in their predictions. Conversely, recall may stagnate or drop slightly (in simpler models like CNN) because the cleaned labels, often removing frequent but unsupported codes, reduce label coverage and teach models to predict fewer, more confident codes.

In contrast, ER substituting codes with finer-grained replacements further increases label specificity but also sparsifies the target space. This may suggest why recall drops significantly with ER. High-capacity models such as LAAT and MultiResCNN sustain these changes with improved precision and maintain F1 gains, whereas simpler CNN and RNN models underperform due to considerably reduced recall.

Future work will extend this analysis by manually validating a subsample of the LLM judgments, scaling from 10k to the full 122k notes, and comparing outputs across newer LLMs (e.g., DeepSeek vs LLaMA-3-70B) and prompt configurations. Cross-testing models trained on OM against EV-verified and noise-stratified subsets will further clarify how label quality affects generalization and real-world robustness.

5. Conclusions

Overall, findings from our validation study suggest that existing ICD coding models perform better than previously reported once evaluation is based on evidence-grounded labels. The improvement under EV reflects a more faithful measure of what the models actually learn from the discharge notes. By contrast, the lower performance under OM mainly results from noisy or unsupported charted codes that obscure true model capability. These results highlight the importance of utilizing evidence-verified benchmarks for NLP research, as they offer a more accurate assessment of model performance and reliability in real-world clinical documentation.

References

[1] \Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data. 2023 Jan 3;10(1).

[2] Searle T, Ibrahim Z, Dobson R. Experimental Evaluation and Development of a Silver-Standard for the MIMIC-III Clinical Coding Dataset. arXiv (Cornell University). 2020 Jan 1;

[3] Edin J, Junge A, Havtorn JD, Borgholt L, Maistro M, Ruotsalo T, et al. Automated Medical Coding on MIMIC-III and MIMIC-IV: A Critical Review and Replicability Study. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval [Internet]. 2023 Jul 18 [cited 2023 Dec 4]; Available from: https://arxiv.org/pdf/2304.10909.pdf