Interpretable Dynamic Rule Attention for Medical Coding

Rimon Paul[†,*], Blessing Ogbuokiri[†]
[†] Responsible and Applied Machine Learning Laboratory (RAML Lab),
Department of Computer Science, Brock University

Abstract

Automatic medical coding maps clinical text to International Classification of Diseases (ICD) codes. While highly accurate, recent neural models operate as black boxes, limiting clinical trust and accountability. We address this by proposing an interpretable, rule-guided attention method for a BioClinicalBERT model fine-tuned on Medical Information Mart for Intensive Care III (MIMIC-III) discharge summaries. Our lightweight approach incorporates domain knowledge via keyword mappings, softly biasing attention toward clinical evidence without restricting the model's learning capacity. Evaluated on the full ICD-9 task, the model improves micro-F1 (0.330 to 0.384), micro-precision (0.391 to 0.420), and micro-recall (0.285 to 0.353). A McNemar test confirms a statistically significant shift in prediction behaviour ($p < 10^{-10}$), while quantitative analysis proves significantly increased attention mass on diagnostic keywords ($p < 10^{-15}$). This transparency incurs minimal computational overhead, utilizing linear-time matching without altering the core transformer architecture. Qualitative visualizations further demonstrate that this rule guidance yields clearer, evidence-aligned decision patterns without sacrificing predictive accuracy.

Keywords: Automatic medical coding, bioclinicalbert, clinical nlp, icd classification, interpretable ai, rule-guided attention

1. Introduction

Automatic medical coding assigns diagnosis codes, such as International Classification of Diseases (ICD) codes (e.g., ICD-9), to clinical documents, including discharge summaries. These codes are essential for clinical documentation, billing, epidemiological research, and healthcare auditing. Recent advances in pretrained transformer models, such as BioClinicalBERT [1], have substantially improved coding accuracy on large-scale electronic health record data. However, despite their effectiveness, these models largely operate as black boxes, making individual predictions difficult to interpret.

Limited interpretability restricts the deployment of automated coding systems in clinical and administrative settings. Clinicians and auditors often require clear justification linking predicted codes to specific clinical evidence. While attention mechanisms are commonly used to provide token-level explanations, standard attention is unguided and may not consistently emphasize clinically meaningful information.

To address this limitation, we propose a rule-guided attention mechanism for automatic medical coding. Our approach integrates lightweight, domain-informed keyword rules into the attention computation of a BioClinicalBERT-based model, softly biasing attention toward clinically relevant phrases without imposing hard constraints on learning.

We evaluate our method on the full ICD-9 coding task using discharge summaries from the MIMIC-III clinical database [2, 3]. Results show that predictive performance remains comparable to a strong baseline, while attention distributions become more aligned with clinically relevant evidence. Statistical significance testing and qualitative visualizations further demonstrate improved interpretability, supporting the use of rule-guided attention for transparent and trustworthy neural medical coding.

[*]rp21qf@brocku.ca

2. Related Work

Early medical coding systems relied on feature engineering and classical classifiers like support vector machines and logistic regression [4]. Modern approaches focus on neural models that learn representations directly from clinical text, achieving substantial performance gains on large datasets.

Transformers dominate clinical natural language processing (NLP). Models like BERT [5] and its variants, including ClinicalBERT and BioClinicalBERT [1], demonstrate strong ICD coding performance by capturing complex contextual relationships, but offer limited interpretability.

To explain predictions, architectures like CAML [6] use per-label attention, but learned weights are often unstable and disconnected from clinical evidence [7]. Post-hoc explainers like LIME [8] and SHAP [9] estimate feature importance, but are computationally expensive and fail to improve internal reasoning.

Hybrid approaches combine neural models with medical knowledge to address these limitations. Purely rule-based systems [10] are interpretable but inflexible and underperform neural models. Consequently, recent efforts integrate domain knowledge into neural architectures via constraints, regularization, or guided learning to balance accuracy and interpretability.

Building on this, we introduce a lightweight, end-to-end differentiable rule-guided attention mechanism. Unlike post-hoc techniques, our computationally efficient method softly biases token-level attention toward clinically relevant phrases, providing intrinsic interpretability without external approximations.

3. Methodology

3.1. Task and Dataset

We formulate medical coding as a multi-label text classification problem. Given a tokenized clinical discharge summary x, we predict a binary label vector $y \in \{0,1\}^{|\mathcal{C}|}$, where $\mathcal{C}$ denotes the full set of ICD–9 diagnosis codes. Each element $y_j = 1$ indicates code j's presence, allowing multiple diagnoses per document.

Experiments utilize the MIMIC–III clinical database [3], comprising de-identified intensive care unit (ICU) discharge summaries. We tackle the full ICD–9 task, which features thousands of codes and severe class imbalance. A detailed discussion outlining our choice to benchmark against ICD–9 rather than newer standards is provided in Appendix A.1. Finally, documents are tokenized via BioClinicalBERT and truncated to a maximum sequence length of $T = 512$ tokens.

3.2. Baseline Model

Our baseline model is BioClinicalBERT [1], a transformer encoder pretrained on large-scale biomedical and clinical text. Given an input document tokenized into a sequence of T tokens, the encoder produces contextualized token representations $H = \{h_1, h_2, \ldots, h_T\}$, where $h_i \in \mathbb{R}^d$ denotes the contextual embedding of the i-th token in the sequence. Each embedding captures the meaning of the token conditioned on its surrounding clinical context. In our implementation, the hidden dimension is $d = 768$.

For baseline prediction, document-level representations are obtained by pooling the encoder outputs and passing them through a linear classifier with sigmoid activation to produce multi-label predictions. The model is fine-tuned end-to-end using binary cross-entropy loss with positive class weighting to mitigate the effects of label imbalance.

3.3. **Rule-Guided Attention Model**

To improve interpretability, we introduce a lightweight rule-guided attention mechanism applied on the final encoder layer, as illustrated in Figure 1. We constructed a domain-informed dictionary mapping ∼20 frequent ICD–9 diagnosis groups (e.g., chronic conditions, acute events) to manually verified clinical keywords (e.g., "diabetes mellitus", "heart failure"). Sourced from established clinical terminology and prior MIMIC–III literature, these rules provide soft guidance toward clinically salient evidence rather than exhaustively encoding medical knowledge.

For an input sequence of T tokens, we generate a binary rule mask $m \in \{0,1\}^T$, where $m_i = 1$ if token i (aligned via tokenizer character offsets) overlaps with a rule-matched keyword span, and $m_i = 0$ otherwise. Token-level unnormalized attention scores $e_i \in \mathbb{R}$ are computed using additive attention [11]: $e_i = v^\top \tanh(Wh_i + b)$, parameterized by a learned projection matrix $W \in \mathbb{R}^{d_a \times d}$, bias $b \in \mathbb{R}^{d_a}$, and attention vector $v \in \mathbb{R}^{d_a}$, with the attention hidden dimension set to $d_a = 128$.

Rule guidance is incorporated by adjusting these scores via an additive bias: $\tilde{e}_i = e_i + \lambda m_i$, where hyperparameter λ controls the adjustment strength. We specifically adopt an additive bias because it softly promotes rule-matched tokens while preserving non-zero attention for all tokens, maintaining broader contextual evidence. In contrast, multiplicative masking risks improperly suppressing non-matched clinical context, while explicit attention regularization introduces complex training loss terms. Our additive formulation remains simple, stable, and end-to-end differentiable.

The adjusted scores are softmax-normalized over all T tokens to yield final attention weights α_i. A document-level context vector, $c = \sum_{i=1}^{T} \alpha_i h_i \in \mathbb{R}^d$, summarizes the attended evidence and is passed through a linear classification layer with sigmoid activation for multi-label prediction. All model components are trained jointly. We empirically select $\lambda = 1.5$ to maximize rule-aligned attention without degrading predictive performance (full sensitivity analysis in Appendix A.2; hardware and complexity details in Appendix A.3).

Average rule attention mass. For a document with attention weights $\alpha \in \mathbb{R}^T$ and mask $m \in \{0,1\}^T$, we define rule attention mass as $\text{Mass}(x) = \sum_{i=1}^{T} \alpha_i m_i \in [0,1]$, which measures the fraction of total attention assigned to rule-matched tokens. We report the average $\text{Mass}(x)$ across a randomly sampled subset of 200 validation documents.

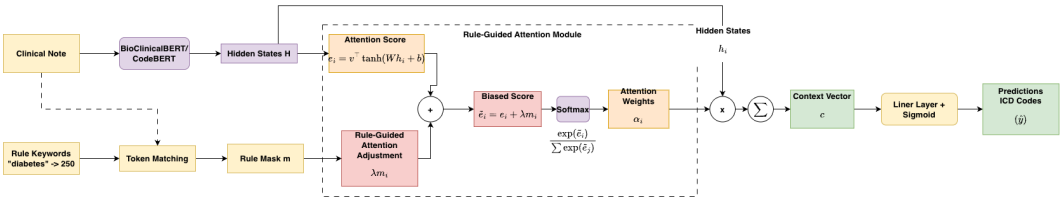


Figure 1. Overview of the rule-guided attention model. Rule-based keyword matches produce a token-level mask that adjusts attention scores prior to normalization.

4. **Results**

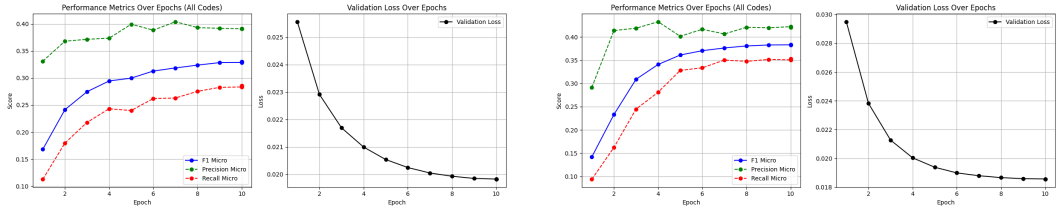
We evaluate the proposed rule-guided attention model on the full ICD coding task using the MIMIC-III discharge summaries. Performance is assessed using standard multi-label classification metrics, along with attention-based analyses to evaluate interpretability.

4.1. **Predictive Performance**

Figure 2 compares the training behaviour of both models. Both converge stably, exhibiting consistent validation loss reductions and steady metric improvements, indicating effective learning on the full ICD label space.

On the held-out test set, the rule-guided model outperforms the baseline in micro-F1 (0.384 vs. 0.330), micro-precision (0.420 vs. 0.391), and micro-recall (0.353 vs. 0.285). Macro-averaged metrics remain low for both models due to the severe label imbalance of rare ICD–9 codes. Because macro-F1 weights all labels equally, it is dominated by these rare codes with limited training signal. Consequently, micro-averaged metrics better reflect accuracy on clinically prevalent diagnoses of greater practical importance.

The rule-guided model's performance gains are most pronounced in early epochs. Although the baseline partially closes this gap over time, this initial acceleration indicates rule-guidance acts as a strong inductive bias. This "warm start" helps the model identify clinically relevant evidence significantly faster, even if both eventually reach comparable predictive capacity.



(a) Baseline BioClinicalBERT *(b)* Rule-guided BioClinicalBERT

Figure 2. Training performance on the full ICD coding task.

4.2. **Attention-Based Interpretability**

Figure 3 presents token-level attention heatmaps for a representative discharge summary. The baseline model distributes attention broadly across the document, including non-diagnostic content. In contrast, the rule-guided model concentrates attention on clinically meaningful phrases, such as diagnoses, symptoms, and key medical events.

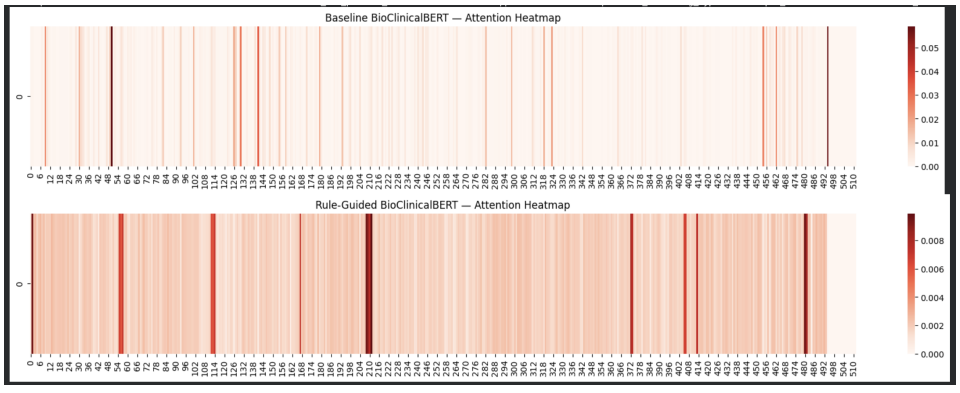


Figure 3. Token-level attention heatmaps comparing baseline (top) and rule-guided (bottom) BioClinicalBERT models. Darker regions indicate higher attention weights.

To further illustrate interpretability, Figure 4 shows attention-weighted text visualizations for the same discharge summary. Tokens receiving higher attention are highlighted more

strongly. The baseline model highlights a wide range of tokens, including structural and less informative text. The rule-guided model produces more focused highlights, emphasizing clinically relevant terms associated with diagnostic reasoning. This qualitative difference demonstrates that rule-guided attention yields clearer and more interpretable evidence linking predictions to clinical text. To further validate these visual patterns quantitatively, we introduce a Rule-Alignment Score (RAS); the full methodology and statistically significant results of this faithfulness analysis are detailed in Appendix A.4. Additionally, access to our publicly available interactive visualization tool is provided in Appendix A.5.

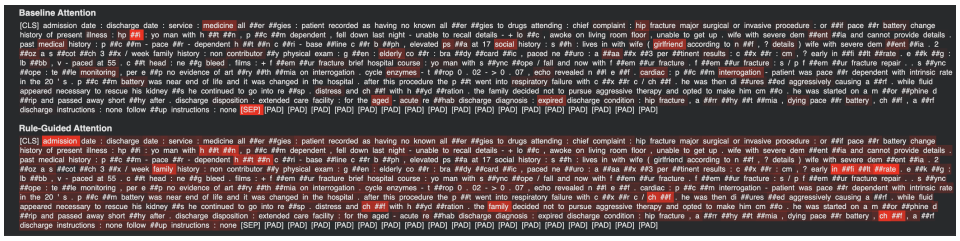


Figure 4. Attention-weighted text visualizations. Top: Baseline BioClinicalBERT. Bottom: Rule-guided BioClinicalBERT. Darker highlights indicate higher attention weights.

4.3. **Statistical Significance Analysis**

Standard evaluation metrics summarize overall performance but obscure individual prediction differences. To assess if the baseline and rule-guided models exhibit statistically distinct decision behaviour, we apply McNemar's test [12] on the full test set. This evaluates the symmetry of prediction disagreements by comparing cases where the baseline is correct but the rule-guided model is incorrect (b), against the reverse (c).

Table 2 (detailed in Appendix A.6) reports the contingency statistics for the full ICD coding task. The large chi-square statistic and corresponding p-value (far below the 0.05 significance threshold) confirm the rule-guided model produces statistically significant differences in prediction behaviour compared to the baseline.

Importantly, this indicates systematically different decision-making rather than universal accuracy superiority. Combined with the aforementioned attention analyses, this finding supports the claim that rule-guided attention meaningfully and interpretably alters model behaviour.

4.4. **Discussion and Implications**

The results show that rule-guided attention changes model behaviour in a meaningful way while preserving strong predictive performance. Although overall accuracy remains comparable to the baseline, McNemar's test indicates that the two models make statistically different predictions, confirming that rule guidance influences decision-making rather than acting as a superficial modification.

Attention heatmaps and highlighted text further demonstrate improved interpretability. Compared to the baseline, the rule-guided model focuses more consistently on clinically relevant phrases, producing clearer and more justifiable evidence for predicted codes. These findings suggest that incorporating domain knowledge into attention mechanisms can improve transparency without sacrificing model flexibility, making the approach suitable for clinical and auditing applications.

5. Conclusion

This work presents a rule-guided attention method for automatic medical coding that softly integrates domain knowledge into a BioClinicalBERT model. This approach improves interpretability and encourages focus on clinical evidence without rigidly constraining learning or sacrificing predictive accuracy.

Experiments on the full ICD task confirm statistically significant behavioural shifts, yielding attention patterns tightly aligned with diagnostic reasoning. Visualizations prove clearer links between predictions and text, directly addressing the limitations of black-box models.

Future work will address three current limitations. First, to overcome the scalability issues of manually curated dictionaries, we will explore data-driven, automated rule generation from corpus statistics. Second, we aim to replace our linear additive bias with learnable gating or non-linear fusion to dynamically balance rules and learned context. Finally, we will expand beyond our ICD-9 benchmark to modern ICD-10 standards and generative large language models for broader clinical applicability.

References

[1] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott. "Publicly available clinical BERT embeddings". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 2019, pp. 72–78.

[2] A. Johnson, T. Pollard, and R. Mark. *MIMIC-III Clinical Database (version 1.4)*. RRID:SCR_007345. 2016. DOI: 10.13026/C2XW26.

[3] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. "MIMIC-III, a freely accessible critical care database". In: *Scientific Data* 3 (2016), p. 160035. DOI: 10.1038/sdata.2016.35.

[4] A. Perotte, R. Pivovarov, K. Natarajan, and N. Elhadad. "Diagnosis code assignment: Models and evaluation metrics". In: *Journal of the American Medical Informatics Association* 21.2 (2014), pp. 231–237.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. 2019, pp. 4171–4186.

[6] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein. "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 (NAACL-HLT)*. 2018, pp. 1101–1111.

[7] S. Jain and B. C. Wallace. "Attention is not explanation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. 2019, pp. 3543–3556.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144.

[9] S. M. Lundberg and S.-I. Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. 2017, pp. 4765–4774.

[10] O. Bodenreider. "The Unified Medical Language System (UMLS): integrating biomedical terminology". In: *Nucleic Acids Research* 32 (2004), pp. D267–D270.

[11] D. Bahdanau, K. Cho, and Y. Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *International Conference on Learning Representations (ICLR)* (2015).

[12] Q. McNemar. "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages". In: *Psychometrika* 12.2 (1947), pp. 153–157. DOI: 10.1007/BF02295996.

Appendix A. Supplementary Materials

A.1. Choice of ICD-9 Standard

We focus on ICD–9 to ensure comparability with prior work, as MIMIC–III provides gold-standard annotations exclusively in this format, defining the majority of established benchmarks. Although newer standards like ICD–10–CA and ICD–11 offer improved clinical granularity, large annotated public datasets remain scarce. Crucially, our rule-guided attention mechanism is standard-agnostic; it can be directly applied to newer versions given appropriate labelled data and dictionaries, which we leave for future work.

A.2. Sensitivity Analysis of Rule Weight (λ)

We study the effect of the rule weight, λ, on both predictive performance and interpretability using a sensitivity analysis. For $\lambda \in \{0.0, 0.5, 1.0, 1.5, 2.0\}$, we train the model for 3 epochs using the same settings as the main experiments and evaluate validation micro-F1.

As detailed in Table 1, micro-F1 remains relatively stable across this range (0.286–0.293), while the average rule attention mass increases monotonically from 0.033 to 0.194, indicating progressively stronger alignment between attention and rule-matched evidence. We select $\lambda = 1.5$ for all main experiments because it yields a substantial increase in rule-aligned attention (Avg Rule Mass = 0.131) without reducing micro-F1 compared to neighbouring values.

λ	Micro-F1 (val)	Avg Rule Mass
0.0	0.2929	0.0334
0.5	0.2887	0.0545
1.0	0.2855	0.0855
1.5	0.2910	0.1309
2.0	0.2933	0.1940

Table 1. Sensitivity analysis of the rule weight λ (3-epoch runs). Micro-F1 remains stable while Avg Rule Mass increases with λ, indicating stronger rule-aligned attention.

A.3. Model Complexity and Training Details

The rule-guided model extends BioClinicalBERT ($\sim$110 million parameters) with a negligible-parameter additive attention component. Because the underlying transformer architecture, input sequence length, and encoder computations remain unchanged, overall computational complexity, memory footprint, and inference costs match the baseline. The added linear-time keyword matching and small attention bias preserve scalability.

Implemented in Python using PyTorch and HuggingFace Transformers, models ran on Google Colab (Linux, single 40 GB NVIDIA A100 GPU) utilizing Mixed-precision (FP16) for efficiency. Both models were trained end-to-end with identical settings: binary cross-entropy loss with positive class weighting (addressing the full ICD–9 task's severe class imbalance), AdamW optimization, a 3×10^{-5} learning rate, and gradient accumulation yielding an effective batch size of 16 from a per-device batch size of 8.

A.4. Quantitative Faithfulness Analysis

To assess interpretability beyond visual inspection, we introduce the Rule-Alignment Score (RAS), a quantitative faithfulness metric that measures the proportion of total attention mass allocated to clinically relevant keywords defined in the rule dictionary. Formally, using

the notation from Section 3, the score for a single document is defined as $\text{RAS} = \sum_{i=1}^{T} \alpha_i m_i$. This metric quantifies the extent to which the model attends to known diagnostic evidence.

We computed the RAS across the entire test set ($N = 5,273$). The baseline BioClinicalBERT model achieved a mean RAS of 5.64%, while the rule-guided model achieved a mean RAS of 6.33%. While the absolute values reflect the natural sparsity of diagnostic keywords within long clinical narratives, the relative increase demonstrates a systematic shift in focus. A paired t-test confirms that this difference is highly statistically significant ($p < 2.13 \times 10^{-15}$). This result provides quantitative evidence that the rule-guided mechanism successfully biases the model's latent attention toward domain-specific clinical concepts, validating the patterns observed in the qualitative heatmaps.

A.5. **Interactive Visualization Tool**

To facilitate interactive inspection of the model's behaviour and explore attention-weighted text visualizations beyond the static examples provided in the main text, we provide a publicly available interactive visualization tool. Interactive Demo: https://medical-coding-explorer-28nsaqkzmhzgruymhs4byk.streamlit.app/

A.6. **McNemar's Test Contingency Statistics**

Statistic	**Value**
b (Baseline correct, Rule-guided wrong)	18,629
c (Baseline wrong, Rule-guided correct)	19,911
Chi-square statistic (χ^2)	42.58
p-value	6.79×10^{-11}

Table 2. McNemar's test results comparing baseline and rule-guided BioClinicalBERT models on the full ICD coding task.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant 2026. The authors also acknowledge the support of the Responsible and Applied Machine Learning Lab (RAML Lab) at Brock University's Department of Computer Science for providing a collaborative research environment and resources that contributed to this work.

Declaration of AI Use

During the preparation of this work, the authors used Gemini to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. No AI generative tools were used to generate scientific content or data, graphs, analysis, or conclusions.