RuCCoD: Towards Automated ICD Coding in Russian

Anonymous ACL submission

Abstract

This study investigates the feasibility of automating clinical coding in Russian, a language with limited biomedical resources. We present a new dataset for ICD coding, which includes diagnosis fields from electronic health records (EHRs) annotated with over 10,000 entities and more than 1,500 unique ICD codes. This dataset serves as a benchmark for several state-of-the-art models, including BERT, LLaMA with LoRA, and RAG, with additional experiments examining transfer learning across domains (from PubMed abstracts to medical diagnosis) and terminologies (from UMLS concepts to ICD codes). We then apply the best-performing model to label an in-house EHR dataset containing patient histories from 2017 to 2021. Our experiments, conducted on a carefully curated test set, demonstrate that training with the automated predicted codes leads to a significant improvement in accuracy compared to manually annotated data from physicians. We believe our findings offer valuable insights into the potential for automating clinical coding in resource-limited languages like Russian, which could enhance clinical efficiency and data accuracy in these contexts.

1 Introduction

The explosion of medical data driven by technology and digitalization presents a unique opportunity to enhance healthcare quality. With the adoption and implementation of electronic health records (EHRs), accurate and timely data utilization is crucial for effective treatment and disease management. Central to this process is the assignment of International Classification of Diseases (ICD) codes, which is essential for medical documentation, billing (Sonabend et al., 2020), insurance (Park et al., 2000), and research (Bai et al., 2018; Lu et al., 2022; Shang et al., 2019).

Although ICD code assignment is crucial for EHRs, it poses significant challenges. Human coders must navigate a wide array of medical terminology, subjective interpretations, and time pressures, all while staying updated with constantly changing classification standards (Burns et al., 2012; O'Malley et al., 2005; Cheng et al., 2009). Coding errors can lead to misdiagnosis, ineffective treatment, diminished trust in the healthcare system, and negative public health outcomes. Furthermore, errors in manual coding in the ICD system, result in financial repercussions, accounting for 6.8% of the total payments (Manchikanti, 2002).

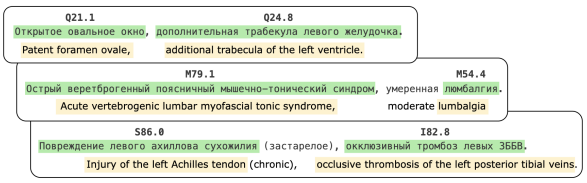

Figure 1: Examples of ICD code assignments by annotators: each entity in green is annotated with its ICD code above and its English translation (in yellow).

Despite extensive research on ICD coding using neural networks (Li and Yu, 2020; Zhou et al., 2021; Yuan et al., 2022a; Baksi et al., 2024; Boyle et al., 2023; Mullenbach et al., 2018a; Cao et al., 2020; Yuan et al., 2022a; Yang et al., 2022; Huang et al., 2022), significant challenges persist for non-English languages. These include low inter-coder agreement, limited labeled data, variability in clinical notes, the hierarchy of ICD codes, and reliance on incomplete input data. To address these issues, we introduce a novel dataset for automatic ICD coding in Russian.

Previous studies have primarily focused on English-language datasets, specifically MIMIC-III/IV (Goldberger et al., 2000; Johnson et al., 2023). Despite being one of the top ten languages in terms of concept name count within the Unified Medical Language System (UMLS) (Bodenreider, 2004) biomedical metathesaurus, Russian remains underdeveloped in the clinical domain. The Rus-

sian segment of the UMLS comprises only 1.96% of the vocabulary and 1.62% of the source counts found in the English UMLS (NIH). Recent corpora, such as RuCCoN (Nesterov et al., 2022a) and NEREL-BIO (Loukachevitch et al., 2023), focus on concepts within the Russian UMLS.

In this work, we explore two closely related tasks: **ICD coding** and **Diagnosis Prediction** (**DP**). As seen in Fig. 2, the tasks take non-overlapping input and complement each other: *ICD coding* normalizes a free-form doctor's diagnosis conclusion into a set of relevant ICD codes while the DP task is to directly predict ICD-agreed diagnoses from EHRs in one pass without relying on the doctor's textual diagnosis conclusion. Although we formulate *ICD coding* as an entity normalization task and *DP* as multilabel classification, both tasks are sometimes referred to as **ICD coding**. Unlike prior classification-based *ICD coding* research (Li and Yu, 2020; Vu et al., 2020; Wang et al., 2024), we explore a more challenging scenario in which a diagnostic model, acting as an independent medical expert, predicts diagnoses from patient data only without relying on the doctor's diagnosis conclusion. Thus, we term the classification task **diagnosis prediction** as it better reflect the problem's nature and does not create a confusion with linking-based *ICD coding* (Lavergne et al., 2016; Névéol et al., 2017; Coutinho and Martins, 2022).

For *ICD coding*, we present **RuCCoD** (**Ru**ssian **ICD** **Co**ding **D**ataset), a novel dataset in Russian[1], labeled by medical professionals based on concepts from the ICD-10 CM (Clinical Modification) system (Sec. 3.1). Second, we establish a **comprehensive benchmark** for state-of-the-art models, including a BERT-based (Devlin et al., 2019) pipeline for information extraction, a LLaMa-based (Touvron et al., 2023) model with Parameter Efficient Fine-Tuning (PEFT) and with retrieval-augmented generation (RAG). Furthermore, we evaluate transfer learning of models trained on UMLS concepts and similar biomedical datasets (PubMed abstracts (Loukachevitch et al., 2023), clinical notes (Nesterov et al., 2022a). The results suggest that the ICD's fine-grained hierarchical structure hinders generalization from other clinical sources (Sec. 4).

For *diagnosis prediction*, we perform a set of experiments on **RuCCoD-DP**, a large in-house dataset of 865k EHRs from 164k patients. When training a diagnostic model, we experiment with

[1]We will release this dataset upon acceptance.

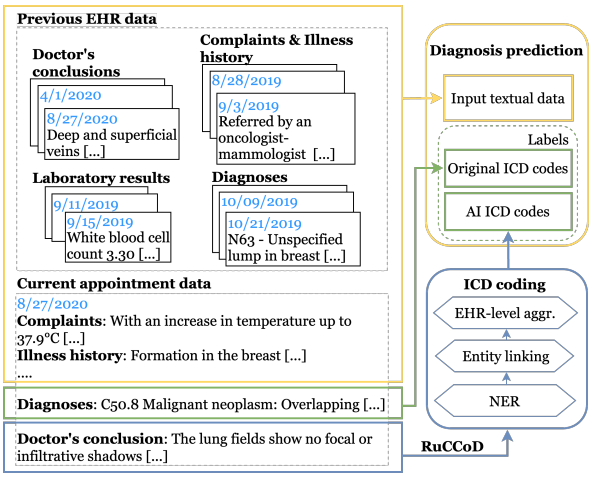

Figure 2: Schematic description of *ICD coding* (in blue) and *diagnosis prediction* tasks (in yellow). Diagnosis prediction uses prior EHR data and current visit details, excluding the doctor's conclusion, which is used for ICD coding to generate *AI-assigned ICD codes*. Both original and AI ICD code lists are then used as targets to train different diagnosis prediction models.

ICD codes assigned by doctors during patient appointments, as well as the **AI-assigned ICD codes** (Sec. 5), that is, diagnoses assigned by automatically linking an EHR diagnosis conclusion with a top-performing *ICD coding* model on RuCCoD (see Fig. 2). Our experiments have revealed that pre-training on automatically assigned ICD codes gives a huge weighted F1-score growth of 28% for **diagnosis prediction** compared to physician-assigned ICD codes indicating the difficulty of ICD-guided diagnosis formalization for physicians and great potential of AI-aided diagnosing. Our work provides a foundation and guidance for ICD-related research in low-resource clinical languages.

2 ICD-Related Tasks

Task: ICD coding is akin to Entity Linking (EL), where the objective is to assign a set of unique ICD codes to the latest patient appointment based on textual diagnosis conclusion written by a doctor. The task aims to help a physician normalize diagnosed diseases to a set of codes from the complex formal ICD hierarchy. We model the *ICD Coding* as an *information extraction* pipeline with three components: (1) *Nested Entity Recognition* (NER) and (2) EL followed by (3) *EHR-level code aggregation*. Step (3) minimizes NER influence on pipeline metrics by omitting NER spans. The approach aligns with real-world ICD applications, where the primary objective is accurate assignment of ICD codes (i.e., disease recognition), and impre-

cise NER outputs are not impactful.

ICD Coding: EHR-level Code Aggregation Given an EHR, we perform EL on NER predictions. Let $L_p = (c_1^p, c_2^p, \ldots, c_n^p)$ and $L_t = (c_1^t, c_2^t, \ldots, c_m^t)$ denote the lists of predicted and ground truth ICD codes, respectively. In standard EL, each list may contain duplicated disease mentions (i.e., $c_i^t = c_j^t$ for $i \neq j$). We remove duplicates from both lists resulting in unique code sets S_p and S_t such that $c_i^p \neq c_j^p$ and $c_i^t \neq c_j^t, \forall i \neq j$. Finally, micro-averaged classification metrics are computed from True Positives (TP), False Positives (FP) and False Negatives (FN): $TP = S_p \cap S_t$; $FP = S_p \setminus S_t$; $FN = S_t \setminus S_p$.

Diagnosis Prediction is a multi-label classification task that outputs likely diagnoses (ICD codes) for the current doctor appointment from a patient's *past medical history*, including complaints, test and examination results from previous appointments. In our study, each EHR contains a doctor's diagnosis conclusion. A major challenge for ICD-grounded applications is that this conclusion is a free-form text, and its normalization to ICD might introduce sensitive errors. Conversely, automatic *Diagnosis Prediction* is constrained to output ICD-compliant diagnoses by task design.

ICD Coding vs. Diagnosis Prediction While *ICD Coding* only observes the current appointment's diagnosis conclusion written by a doctor, the goal of *Diagnosis Prediction* is to actually write the diagnosis conclusion (i.e., make an **AI diagnosis conclusion**). Here, the motivation is to offer a doctor an independent, AI-driven opinion, potentially beneficial for decision-making in complex cases. Hence, the **two tasks are complementary by design**, using non-overlaping EHR parts: *ICD Coding* leverages the latest diagnosis while *Diagnosis Prediction* observes an entire patient's history except for the latest diagnosis conclusion.

3 ICD Datasets

3.1 RuCCoD: ICD Coding Dataset

For **ICD coding**, we release **RuCCoD**, the first dataset of Russian EHRs with disease entities manually linked to ICD-10. In this section, we describe the data collection and annotation pipeline and provide important statistics.

Data Collection As a **source** for RuCCoD, we utilize diagnosis conclusions from the records of a major European city's Medical Information System. Before starting the annotation process, we implemented a meticulous de-identification protocol to protect data privacy. Medical professionals invited to annotate the dataset first conducted a comprehensive manual review of all diagnoses. Their task was to identify and remove any personal or identifiable information manually. This thorough process guarantees compliance with privacy regulations and ensures the dataset is suitable for research use.

	Train	Test
# of records	3000	500
# of assigned entities	8769	1557
# of unique ICD codes	1455	548
Avg. # of codes per record	3	3

Table 1: Statistics for the RuCCoD training and testing sets on ICD coding of diagnosis.

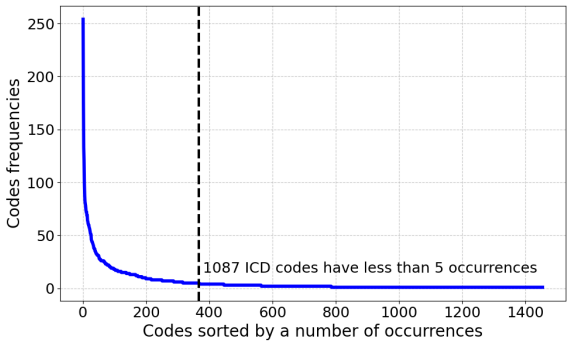

Figure 3: Distribution of ICD code frequencies in the RuCCoD train set.

Annotation Process and Principles The labeling team consisted of three highly qualified experts with advanced education in different fields of medicine, two of whom hold Ph.D. degrees, with every annotation further validated by a fourth expert, a Ph.D. holder in medicine. Grounded in the ICD-10 CM (Clinical Modification) system, the team aimed to identify all nosological units in a diagnosis conclusion and assign the most accurate ICD code to each. An annotation example is shown in Fig. 1. The dataset was randomly split into 3,000 training and 500 testing records. Each expert independently annotated 1,000 training records for diverse labeling, while all three annotated the same 500 test records for consistency. An ICD code was accepted if at least two annotators agreed. Annotation guidelines are in Appx. A.

Inter-Annotator Agreement We assessed annotation consistency among experts using the *Inter-Annotator Agreement* (IAA) metric, defined as the

	Original Dataset	Linked Dataset	Manual Test Set
Number of records	865539	865539	494
Number of unique patients	164527	164527	450
Number of unique ICD codes	3546	3546	394
Avg. number of ICD codes per patient	3 ± 2	5 ± 2	4 ± 2
Avg. number of EHR records before current appointment	(15, 36, 73)	(15, 36, 73)	(17, 36, 77)
Avg. length of EHR records per one appointment	(77, 167, 316)	(77, 167, 316)	(86, 176, 320)
Patient's age	(59, 67, 74)	(59, 67, 74)	(60, 67, 75)
Percentage of male patients	69	69	71

Table 2: Statistics for the randomly split training and testing sets of RuCCoD-DP for diagnosis prediction. Values in brackets show the 25th, 50th, and 75th percentiles.

ratio of accepted codes to the total unique codes assigned per record (Luo et al., 2019). Among ICD codes, the IAA value was 50%, indicating moderate agreement. Low intercode agreement among human annotators reflects both the subjectivity of manual ICD coding and the inherent complexity of the task. Our observation is well correlated with existing studies that have reported a fair to moderate agreement on terminal ICD codes, with Kappa values ranging from 27% to 42%, corresponding to agreement rates of 29.2% and 46.8%, respectively (Stausberg et al., 2008). The reported accuracy of coding exhibits significant variability, ranging from 53% to 98% (Campbell et al., 2001) and from 41.8% to 88.87% (Hosseini et al., 2021). The obtained low IAA metric highlights both the challenges and strengths of our annotation process.

Dataset Statistics Statistics of train and test splits of the RuCCoD dataset are provided in Tab. 1. Despite the large number of ICD codes, especially in the training set, their distribution is uneven. Fig. 3 shows the distribution of ICD codes within the RuCCoD train set. While a small number of codes dominate the dataset, appearing from 50 to 250 occurrences, most codes are rare, with 1,087 codes occurring fewer than 5 times. This stark disparity underscores the challenges of dealing with real-world medical data, where frequent diagnoses are well-represented, but rare conditions remain significantly under-sampled.

3.2 RuCCoD-DP: Diagnosis Prediction Dataset

To explore AI-guided **Diagnosis Prediction**, we collect **RuCCoD-DP** (**RuCCoD** for **D**iagnosis **P**rediction), a corpus of real-world EHRs.

Dataset Construction RuCCoD-DP includes doctor appointments from 2017 to 2021, divided into four parts: (i) patient complaints and anamnesis, (ii) lab test results, (iii) appointment summary (including assigned ICD codes), and (iv) appointment history. Although RuCCoD and RuCCoD-DP share a common source, we ensure both sets to have no overlapping appointments and patients.

Paired Human-AI ICD Codes ICD has a fine-grained disease hierarchy introducing a significant challenge even for a qualified doctor to formalize a correctly diagnosed disease . For instance, a *H10 Conjunctivitis* disease group has 8 specifications including: *H10.0 mucopurulent*, *H10.1 acute atopic*, *H10.2 other acute*, and *H10.3 unspecified acute* conjunctivitis. Thus, doctor-assigned ICD codes in real-world EHRs can expose substantial errors even if a general disease is diagnosed correctly. To address the issue, we consider two ICD code sets for each EHR: (i) real-world ICD codes originally written by physicians within the EHR (**doctor-assigned codes**); (ii) automatically assigned ICD codes predicted by a neural model trained on *RuCCoD* (**AI codes**). **AI codes** (i.e., *AI-assigned diseases*) are assigned to an EHRs by applying our top-performing BERT-based NER+EL *ICD Linking* pipeline (Tab. 3) to a doctor's real-world diagnosis conclusion (see Fig. 2). Our pipeline extracts diseases (NER) and links (EL) them to ICD codes and then the found diseases are assigned to the given EHR labels for *ICD Coding*. Thus, *AI codes* are designed to aid in the formalization of the human-written diagnosis to the ICD code set while relying only on the written conclusion of the physician. Notably, the two coding types rely on the same underlying free-form diagnosis conclusions.

Original and Linked RuCCoD-DP We will refer to RuCCoD-DP variations sharing the same appointments yet different in ICD code assignment method (either *doctor-assigned* or *AI-based*) as **original** and **linked** datasets, respectively. In other words, a single textual appointment entry has two

distinct labels sets. To prevent ICD codes distribution shift between *original* and *linked* data, we retained the ICD codes overlapping between these two sets. For each appointment sample, its textual input included the concatenation of chronologically sorted all prior appointments.

Diagnosis Prediction Test Set The collection of two sets of labels allows exploration of whether manual or generated ICD labels are more reliable for model training. For a fair comparison of the labeling approaches, we manually labeled a *common test set* from a subset of the original appointment dataset's test set. We formed it by selecting a subset from the test part of the original appointment dataset. For annotation, we adopted the same annotation methodology as for the RuCCoD dataset (Sec. 3.1). The IAA between the doctors was 50% for exact ICD codes and 74% for ICD groups. The final statistics for *original*, *linked* datasets as well as the *common manual test* is summarized in Tab. 2.

4 ICD Coding Evaluation

For **ICD coding** experiments, we experiment with the following approaches: 1) a fine-tuned BERT-based pipeline for information extraction, 2) a large language model (LLM) with Parameter-Efficient Fine-Tuning (PEFT), and 3) LLM with retrieval-augmented generation (RAG). All three systems use the same dictionary, with 17,762 pairs of codes and diagnoses (refered to as *ICD dict*) compiled from the Ministry of Health data. In addition, LLM-based systems used a train set as a dictionary as well. See the Appx. G for a list of the LLMs used. See related work in Appx. B.

4.1 Models

BERT-based IE Pipeline Our Information Extraction (IE) pipeline uses sequential NER and EL modules. The NER module, employing a softmax layer, extracts relevant entities, and the EL module then links these entities to ICD codes based on semantic similarity with ICD dictionary entries. For NER, we utilize the pre-trained RuBioBERT (Yalunin et al., 2022), and for EL, we employ the multilingual state-of-the-art models SapBERT (Liu et al., 2021a,b), CODER (Yuan et al., 2022b), and BERGAMOT (Sakhovskiy et al., 2024). We fine-tuned models on training EL sets via synonym marginalization proposed in *BioSyn* (Sung et al., 2020). For more details, see Appx G.

LLMs with PEFT We explored the capabilities of LLMs for clinical coding using PEFT with Low-Rank Adaptation (LoRA) (Hu et al., 2021). The pipeline included two steps: NER and EL, following the structure of BERT-based IE pipeline described earlier. For NER stage, models were fine-tuned on RuCCoD using task-specific prompts (Appx. H). The predictions were validated by exact string matching and Levenshtein distance with a threshold ≤ 2 chosen empirically to optimize the robustness of the spelling without overcorrecting semantically distinct entities. For EL, a RAG approach was implemented to link extracted entities to ICD codes. The retrieval component was built using three strategies: (1) BGE embeddings (Chen et al., 2024) on the *ICD dict*, (2) BGE embeddings on the *ICD dict* combined with RuCCoD training entities, and (3) BERGAMOT embeddings (Sakhovskiy et al., 2024) fine-tuned on RuCCoD with the *ICD dict*.

We adopted the FAISS index (Douze et al., 2024) to retrieve the top-15 most similar dictionary entries for each entity extracted in the NER stage. The final ICD code was assigned using an LLM to select the closest match from the retrieved candidates (prompt in Appx. H). To address class imbalance, diagnosis lists were shuffled during training, forcing models to learn contextual code-discrimination. Fine-tuning parameters followed standard LoRA configurations (Tab. 4, Appx. G).

Zero-shot LLM with RAG As an ablation study, we evaluated the same pipeline as in the PEFT stage but without fine-tuning to isolate the LLMs' inherent capabilities. We used only the fine-tuned BERGAMOT embeddings from strategy (3) for retrieval, retaining the FAISS index and prompts (Appx. H). The LLM selected ICD codes from retrieved candidates if no direct match was found, replicating the EL process from the PEFT stage. This setup allowed us to quantify the contribution of fine-tuning versus zero-shot inference.

4.2 Evaluation Methodology

On RuCCoD, our evaluation includes conventional NER and EL as well as end-to-end document-level code assignment with EHR-level code aggregation (Sec. 2). To recall, document-level metrics is an entity position-agnostic NER+EL task composition with explicitly removed EHR-level ICD code duplicates. For instance, a language model successfully diagnoses a patient by assigning the correct ICD

Model	Precision	Recall	F-score	Accuracy
Supervised with various corpora for NER and EL				
BERT, NER: NEREL-BIO + RuCCoD, EL: RuCCoD	**0.512**	0.529	0.520	0.352
BERT, NER: RuCCoN + RuCCoD, EL: RuCCoD	0.471	**0.543**	0.504	0.337
BERT, NER: RuCCoD, EL: RuCCoD	0.510	0.542	**0.525**	**0.356**
LLM with RAG (zero-shot with dictionaries)				
LLaMA3-8b-Instruct, NEREL-BIO	0.059	0.053	0.056	0.029
LLaMA3-8b-Instruct, RuCCoN	0.164	0.15	0.157	0.085
LLaMA3-8b-Instruct, ICD dict.	0.379	0.363	0.371	0.228
LLaMA3-8b-Instruct, ICD dict. + RuCCoD	0.465	0.451	0.458	0.297
LLM with tuning				
Phi3_5_mini, ICD dict.	0.394	0.39	0.392	0.244
Phi3_5_mini, ICD dict. + RuCCoD	0.483	0.477	0.48	0.316
Phi3_5_mini, ICD dict. + BERGAMOT	0.454	0.448	0.451	0.291

Table 3: Entity-level code assignment metrics on RuCCoD's test set. The best results are highlighted in **bold**. We also refer to Appx. D, E, F on more experiments with different LMs, corpora, and terminologies.

code when it finds at least one of three mentions of the corresponding ICD disease within an EHR. For all three tasks, we report accuracy and the micro-averaged precision, recall, F1-score.

For EL, we use a retrieval-based approach (Liu et al., 2024; Yuan et al., 2022b; Sakhovskiy et al., 2024) and evaluate retrieval accuracy: $acc@k = 1$ if a correct ICD code is retrieved at rank $\leq k$. We consider two evaluation scenarios: (i) *strict* score assessing exact match between a predicted ground truth codes; (ii) *relaxed* score with each code being truncated to higher-level disease group (e.g., *H10.0 mucopurulent* is truncated to *H10 Conjunctivitis*).

4.3 Results

4.3.1 Transfer Learning

First, we performed cross-domain experiments on EL to see how variability in entities and terminology affects the performance. Since UMLS includes the ICD system, we automatically map UMLS CUIs to ICD codes for evaluation. Cross-domain transfer results with entity linking models on RuCCoD, RuCCoN, NEREL-BIO and their union are presented in detail in Appx. D. The evaluation has revealed the following key observations.

Maleficent Cross-Domain Vocabulary Extension While extension of ICD vocabulary consistently gives a slightly improved acc@1 in a zero-shot setting, additional synonyms introduce severe noise in a supervised setting. Specifically, a significant drop of 8.1%, 8.4%, and 14.3% acc@1 is observed for SapBERT, CODER, and BERGAMOT, respectively. Even in an unsupervised setting, vocabulary extension drops acc@5 by 5.2% and 6.8% for SapBERT and BERGAMOT, respectively.

Complicated Cross-Terminology Transfer Both training on RuCCoN and NEREL-BIO as well as merge of these corpora with RuCCoD do not lead to improvement over zero-shot coding. The finding indicates that training on either datasets does not easily transfer to our dataset as well as the specificity and high complexity of hierarchical ICD coding within the EL task.

Complexity of Fine-Grained ICD Coding The 15% gap in acc@1 between *strict* and *relaxed* evaluations shows the challenging nature of semantically similar diseases within the same therapeutic group.

Transfer learning for NER is Feasible A NER model trained on the disease-related entities from NEREL-BIO gained an F1 score of 0.62 on RuCCoD's test set. The model trained on a combined dataset of NEREL-BIO and RuCCoD achieved scores of 0.72. Similar results were observed with RuCCoN. We also evaluated BINDER, which uses a RuBioBERT backbone and treats NER as a representation learning problem by maximizing similarity between vector representations (Zhang et al.). However, BINDER's performance was 1.5% lower than RuBioBERT's, which gained the best F1 score of 0.77 with a softmax classifier. NER transfer for disease entities is significantly better than for entity

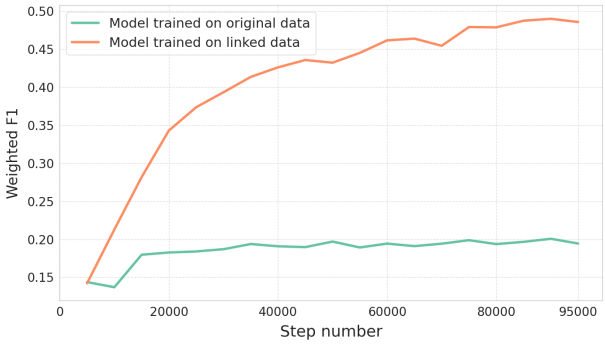

Figure 4: Comparison of weighted F1 scores on the manual diagnosis prediction test set for models trained on *original* and *linked* datasets at different training steps.

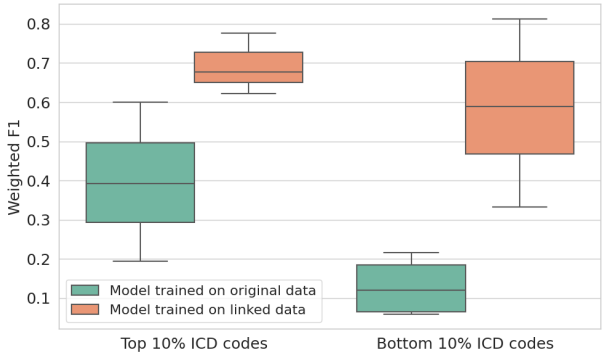

Figure 5: F1 score distribution for top and bottom 10% frequent ICD codes in the common test set.

linking (EL), with the best results obtained from RuCCoD (full results in Appx. C).

4.3.2 End-to-end ICD coding

In the next experiments, we evaluated an end-to-end ICD coding quality on raw texts, in which models were fine-tuned on either RuCCoN or NEREL-BIO or utilized entity dictionaries from these datasets, are presented in Tab. 3. As seen from the results, training on datasets from other domains gives limited performance and the best ICD coding results are observed for models trained with ICD data from RuCCoD data on all three set-ups.

Extended RAG results are in Appx. E. Fine-tuning LLMs improves performance across all tasks, exceeding LLM + RAG results in zero-shot settings. Use of RuCCoD significantly enhances metrics compared to approaches that rely solely on the ICD dictionary or embeddings. Llama3-Med42-8B and Phi3_5_mini are the most effective models after PEFT tuning (see Appx. F).

5 Diagnosis Prediction Evaluation

5.1 Experimental Set-up

Model We chose the Longformer architecture (Beltagy et al., 2020) due to its strong performance in clinical tasks (Edin et al., 2023). Our Longformer model is initialized from a BERT model pre-trained using private EHRs from multiple clinics and further pre-trained on extended sequences. Training details are in Appx. G.

Evaluation To address the imbalanced long-tail ICD code distribution in the **Diagnosis Prediction** task, we adopt the weighted F1 score for overall evaluation, as it has proven effective used in previous research on the problem (Johnson and Khoshgoftaar, 2019; Blanco et al., 2020). The weight of each class is calculated as the proportion of EHRs sharing the given ICD code in the union of both training datasets. Per-class F1 scores were also measured to explore performance variations across frequent and rare ICD codes. In our experiments, we evaluate the quality of the models trained on *original* and *linked* datasets on the *manual test set*.

5.2 Results

5.2.1 Diagnosis Prediction Learning

To predict ICD codes from doctor's appointments, we fine-tuned two Longformer models, one using the *original* dataset and the other using the *linked* dataset. The weighted F1-scores for the two models against the training count are shown in Fig. 4.

AI-based ICD Coding Improves Diagnosing As seen from Fig. 4, AI-guided ICD coding (*linked* data) significantly outperforms manual coding (*original* data) with the peak weighted F1-score of 0.48. The latter quickly reaches its F1-score plateau at 0.2. The huge performance gap of 0.28 highlights the effectiveness of automatic data annotation for model training. Yet, the finding reveals the complexity of ICD-agreed diagnosis prediction task for professional physicians indicating the necessity of AI-driven assistance.

5.3 Diagnosing Stability to Disease Frequency

Next, we study the diagnosis prediction model's ability to generalize to both frequent and rare disease when trained on *original* and *linked* datasets.

Frequency-Based ICD Test Set Split The test dataset was split into two parts: the 10% most frequent ICD codes and the 10% least frequent ICD codes, with a minimum frequency threshold of 15 instances in the *manual* test set for the less frequent group. The stratification approach is designed to align with the distribution of real-world diagnoses assigned and carefully verified by clinicians.

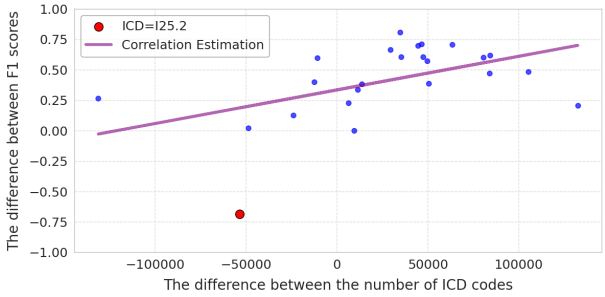

Figure 6: Dependency between differences in the number of codes in original and linked train sets and corresponding F1 scores differences on the common test.

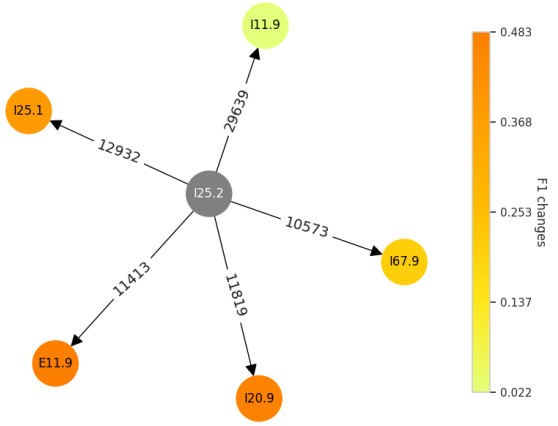

Figure 7: The relationship between transitions from I25.2 and F1 improvements: numbers on the arrows indicate transition frequency, while node color intensity represents the magnitude of F1 metric change.

Diagnosing Improvement is Frequency-Robust Fig. 5 presents the F1 scores spread for individual ICD codes (diseases) grouped by frequency groups. The model trained on *linked* data outperforms the one trained on *original* data for both rare and frequent codes. The ∼6x median F1 score improvement for the bottom 10% codes (0.6 vs. 0.1) underscores the difficulty of manually assigning ICD codes for infrequent diseases. For frequent codes, the training on *linked* data gives about 0.3 median F1 growth over *original* data (∼0.7 vs 0.4) with a significantly lower score deviation (indicated by smaller interquartile distance). Thus, pretraining on automatically labeled data enhances diagnosis prediction for both rare and common diseases, reducing variability for the latter.

5.4 Disease-Wise Quality Shift Analysis

Linked Data's Improvement Stability Fig. 6 shows how changes in appointment counts from *original* to *linked* data affect the diagnosing F1 score. Notably, F1 scores generally improved for the majority of ICD codes regardless of appointment counts increase or decrease. This suggests improved class balance in the linked dataset, although the effect varies.

Case Study: Diagnosing Degradation In Fig. 6, a sharp F1 score drop is observed for *I25.2 Past myocardial infarction*. Apparently, the disease has been mistakingly re-linked to other errounessly. We studied the case by analyzing which ICD codes has *I25.2* been replaced with. Fig. 7 shows the most frequent transition (I25.2 to I11.9, *Hypertensive heart disease*) yielded minimal F1 improvement (0.02), likely due to symptom overlap. E11.9 (*Type 2 diabetes mellitus*) had the highest gain (0.48) due to clearer distinctions. I25.1 (*Atherosclerotic heart disease*) and I20.9 (*Angina pectoris*) had significant gains (0.38, 0.47), while I67.9 (*Unspecified cerebrovascular disease*) had a moderate gain (0.21). From the results, clearly distinguishable diagnoses yield higher F1 scores compared to those with symptom overlap.

6 Conclusion

In this paper, we presented the first models for multi-label ICD-10 coding of electronic health records (EHRs) in Russian. Our study focuses on two key tasks: information extraction from the diagnosis field of EHRs and diagnosis prediction based on a patient's medical history. The NLP pipeline developed for the first task was utilized to re-annotate EHRs in the training set for the second task. The results demonstrate that fine-tuned LMs significantly enhance performance in predicting ICD codes from past medical history. Specifically, the model trained on automatically linked data exhibited faster learning and better generalization compared to the original dataset, achieving higher weighted F1 scores early in training, while the original model plateaued with minimal improvements. Notably, the linked data model consistently outperformed the original across both frequent and rare ICD classes, achieving higher F1 scores with reduced variability. This suggests that the linked dataset enables effective handling of both common and rare ICD codes. Overall, our findings highlight the importance of a neural pipeline for automating ICD coding and improving the accuracy and informativeness of medical text labeling.

Future research will focus on the integration of additional external medical sources like knowledge graphs to improve ICD code prediction. We plan to study the generalization of LLMs on rare codes.

Limitations

Other biomedical corpora in Russian The most relevant corpora to our study are RuCCoN (Nesterov et al., 2022b) and NEREL-BIO (Loukachevitch et al., 2023, 2024), which link entities from clinical records or PubMed abstracts to the Russian segment of UMLS. We conducted preliminary transfer learning experiments using these two corpora; however, a detailed analysis of the semantic differences among the three corpora has yet to be performed.

Moderate Inter-Annotator Agreement Among ICD codes, the IAA value was 50%, indicating moderate agreement, while for ICD groups, the IAA increased to 74%, reflecting higher consistency at a group-wise level. This disparity suggests that annotators were generally aligned when categorizing broader ICD groups but faced challenges in granular code assignment. This pattern mirrors trends observed in clinical practice, possibly due to ambiguities in documentation and coding guidelines (cf. §3.1). While our terminal code IAA (50%) aligns with the upper bounds of reported expert agreement (29.2%–46.8%) (Stausberg et al., 2008), the residual variability underscores the need for standardized annotation protocols or ensemble approaches to mitigate subjectivity in fine-grained coding.

Clinical Diversity While our dataset is substantial, it may not fully capture the diversity of clinical scenarios and patient demographics. A more varied dataset could improve the robustness and generalizability of the models. Clinical language can vary significantly across different medical specialties and institutions. This variability may impact the model's ability to generalize across various clinical contexts.

Data Imbalance The dataset may suffer from class imbalance, with certain ICD codes being underrepresented. This could affect the model's ability to generalize and accurately predict less common diagnoses.

Ethics Statement

No Personal Patient Data in RuCCoD RuCCoD does not contain any personally identifiable patient information. The dataset consists solely of diagnosis conclusions written by medical professionals, which were manually labeled based on the ICD-10 CM (Clinical Modification) system. Prior to the annotation process, annotators were instructed to ensure that no personal information was included in the conclusions. Their task was to identify and remove any personal or identifiable information manually from these texts. Overall, no patient-related information will be disclosed upon the dataset's release.

Private in-house EHR data in RuCCoD-DP Diagnosis prediction leverages prior EHR data along with details from the current visit. As a source for RuCCoD-DP, we utilize records from the Medical Information System of a major European city. All patients, prior to visiting a doctor, sign a special consent form for the processing of their data. The EHR data, which forms the foundation of RuCCoD, is an in-house dataset that will not be released.

Human Annotations The dataset introduced in this paper involved only new annotations. Dataset annotation was conducted by annotators, and there are no associated concerns (e.g. regarding compensation). Each annotator received a compensation of approximately $12 per hour for their contributions. An estimated 85 hours of annotation work per expert resulted in a total payment of $1,020 per annotator. For context, the minimum monthly wage in Russia for full-time employment is under $200, highlighting the substantial effort and investment in creating this high-quality resource. All annotators were aware of potential annotation usage for research purposes. As discussed in limitations, we believe these new annotated datasets serve as a starting point for the evaluation of LMs on ICD coding in Russian. Our annotations, code, and annotation guidelines will be released upon acceptance of this paper.

Inference Costs Running the complete evaluation experiment on a single V100 GPU takes approximately 7.5h and 11h for a decoder-only and encoder-only LM, respectively, while the LLM with RAG evaluation experiment on a single A100 GPU takes approximately 5.5h.

Potential Misuse The RuCCoD dataset, intended for ICD coding in Russian, may be misused if not handled correctly. Potential issues include inaccurate applications leading to incorrect code assignments and overreliance on automated systems without proper validation. To prevent these problems, it is crucial to provide clear guidelines and

adequate training for doctors on using AI assistants, ensuring compliance with ethical and legal standards in research and healthcare.

Transparency The RuCCoD and all associated annotation materials are being released under the CC BY 4.0 license. It should be noted that the dataset contains only diagnosis codes and no medical histories or personal patient data. Furthermore, all diagnoses have been rigorously verified to ensure complete anonymity, in accordance with the prevailing norms of open research practice. Our GitHub repository and HuggingFace dataset card will include comprehensive documentation on the codebase, the methodology for creating benchmarks, and the human annotation process. The source code for our experiments will be freely available at this anonymized repository: https://github.com/auto-icd-coding/ruccod.

Use of AI Assistants We utilize Grammarly to enhance and proofread the text of this paper, correcting grammatical, spelling, and stylistic errors, as well as rephrasing sentences. Consequently, certain sections of our publication may be identified as AI-generated, AI-edited, or a combination of human and AI contributions.

References

Llama-3.1-8b-instruct. https://ollama.com/library/llama3.1:8b-instruct-fp16.

Med42-8b. https://huggingface.co/m42-health/Llama3-Med42-8B.

Mistral-nemo-instruct-2407. https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407.

Phi-3.5-mini-instruct. https://huggingface.co/microsoft/Phi-3.5-mini-instruct.

Qwen2.5-7b-instruct. https://huggingface.co/Qwen/Qwen2.5-7B-Instruct.

Akhila Abdulnazar, Roland Roller, Stefan Schulz, and Markus Kreuzthaler. 2024. Large language models for clinical text cleansing enhance medical concept normalization. *IEEE Access*, 12:147981–147990.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical bert embeddings. *ArXiv*, abs/1904.03323.

Tian Bai and Slobodan Vucetic. 2019. Improving medical code prediction from clinical text via incorporating online knowledge sources. *The World Wide Web Conference*.

Tian Bai, Shanshan Zhang, Brian L Egleston, and Slobodan Vucetic. 2018. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 43–51.

Krishanu Das Baksi, Elijah Soba, John J. Higgins, Ravi Saini, Jaden Wood, Jane Cook, Jack Scott, Nirmala Pudota, Tim Weninger, Edward Bowen, and Sanmitra Bhattacharya. 2024. Medcoder: A generative ai assistant for medical coding. *Preprint*, arXiv:2409.15368.

Weidong Bao, Hongfei Lin, Yijia Zhang, Jian Wang, and Shaowu Zhang. 2021. Medical code prediction via capsule networks and icd knowledge. *BMC Medical Informatics and Decision Making*, 21.

Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad. 2018. Multi-label classification of patient notes: Case study on icd code assignment. In *AAAI Workshops*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Alberto Blanco, Alicia Pérez, and Arantza Casillas. 2020. Extreme multi-label ICD classification: Sensitivity to hospital service and time. *IEEE Access*, 8:183534–183545.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Zeyd Boukhers, Prantik Goswami, and Jan Jürjens. 2023. Knowledge guided multi-filter residual convolutional neural network for icd coding from clinical text. *Neural Computing and Applications*, 35(24):17633–17644.

Joseph S. Boyle, Antanas Kascenas, Pat Lok, Maria Liakata, and Alison Q. O'Neil. 2023. Automated clinical coding using off-the-shelf large language models. *Preprint*, arXiv:2310.06552.

E M Burns, Emily Rigby, Ravikrishna Mamidanna, Alex Bottle, Paul P Aylin, Paul Ziprin, and Omar Faiz. 2012. Systematic review of discharge coding accuracy. *Journal of public health*, 34 1:138–48.

Sean E Campbell, Mark K Campbell, Jeremy M Grimshaw, and Angela E Walker. 2001. A systematic review of discharge coding accuracy. *Journal of Public Health Medicine*, 23(3):205–211.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Ping Cheng, Annette Gilchrist, Kerin Robinson, and Lindsay Paul. 2009. The risk and consequences of clinical miscoding due to inadequate medical documentation: A case study of the impact on health services funding. *Health Information Management Journal*, 38:35 – 46.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. *JMLR Workshop and Conference Proceedings*, 56:301–318. PMID: 28286600, PMC: PMC5341604.

Isabel Coutinho and Bruno Martins. 2022. Transformer-based models for icd-10 coding of death certificates with portuguese text. *Journal of biomedical informatics*, 136:104232.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaolin Diao, Yanni Huo, Shuai Zhao, Jing Yuan, Meng Cui, Yuxin Wang, Xiaodan Lian, and Wei Zhao. 2021. Automated icd coding for primary diagnosis via clinically interpretable machine learning. *International journal of medical informatics*, 153:104543.

Hang Dong, Víctor Suárez-Paniagua, Huayu Zhang, Minhong Wang, Emma Whitfield, and Honghan Wu. 2021. Rare disease identification from clinical notes with ontologies and weak supervision. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2294–2298. IEEE.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Joakim Edin, Alexander Junge, Jakob Drachmann 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. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Fumitoshi Fukuzawa, Yasutaka Yanagita, Daiki Yokokawa, Shun Uchida, Shiho Yamashita, Yu Li, Kiyoshi Shikino, Tomoko Tsukamoto, Kazutaka Noda, Takanori Uehara, and Masatomi Ikusaka. 2024. Importance of patient history in artificial Intelligence-Assisted medical diagnosis: Comparison study. *JMIR Med Educ*, 10:e52674.

Chufan Gao, Mononito Goswami, Jieshi Chen, and Artur Dubrawski. 2022. Classifying unstructured clinical notes via automatic weak supervision. In *Machine Learning for Healthcare Conference*, pages 673–690. PMLR.

A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–220.

Robert Grout, Rishab Gupta, Ruby Bryant, Mawada A Elmahgoub, Yijie Li, Khushbakht Irfanullah, Rahul F Patel, Jake Fawkes, and Catherine Inness. 2024. Predicting disease onset from electronic health records for population health management: a scalable and explainable deep learning approach. *Front Artif Intell*, 6:1287541.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Nafiseh Hosseini, Khalil Kimiafar, Sayyed Mostafa Mostafavi, Behzad Kiani, Kazem Zendehdel, Armin Zareiyan, and Saeid Eslami. 2021. Factors affecting the quality of diagnosis coding data with a triangulation view: A qualitative study. *The International Journal of Health Planning and Management*, 36(5):1666–1684.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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. Association for Computational Linguistics.

Shaoxiong Ji, Matti Hölttä, and Pekka Marttinen. 2021. Does the magic of bert apply to medical code assignment? a quantitative study. *Computers in biology and medicine*, 139:104998.

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

Justin M. Johnson and Taghi M. Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27.

Byung-Hak Kim and Varun Ganapathi. 2021. Read, attend, and code: Pushing the limits of medical codes prediction from clinical notes by machines. In *Machine Learning for Healthcare Conference*, pages 196–208. PMLR.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Eyal Klang, Idit Tessler, Donald U Apakama, Ethan Abbott, Benjamin S Glicksberg, Monique Arnold, Akini Moses, Ankit Sakhuja, Ali Soroush, Alexander W Charney, David L. Reich, Jolion McGreevy, Nicholas Gavin, Brendan Carr, Robert Freeman, and Girish N Nadkarni. 2024. Assessing retrieval-augmented large language model performance in emergency department icd-10-cm coding compared to human coders. *medRxiv*.

Heejoon Koo. 2024. Next visit diagnosis prediction via medical code-centric multimodal contrastive EHR modelling with hierarchical regularisation. *CoRR*, abs/2401.11648.

Keith Kwan. 2024. Large language models are good medical coders, if provided with tools. *Preprint*, arXiv:2407.12849.

Thomas Lavergne, Aurélie Névéol, Aude Robert, Cyril Grouin, Grégoire Rey, and Pierre Zweigenbaum. 2016. A dataset for icd-10 coding of death certificates: Creation and usage. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 60–69.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Fei Li and Hong Yu. 2020. ICD coding from clinical text using multi-filter residual convolutional neural network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8180–8187.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 565–574, Online. Association for Computational Linguistics.

Haochen Liu, Sai Rallabandi, Yijing Wu, Parag Dakle, and Preethi Raghavan. 2024. Self-training strategies for sentiment analysis: An empirical study. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1944–1954, St. Julian's, Malta. Association for Computational Linguistics.

Yang Liu, Hua Cheng, Russell Klopfer, Matthew R Gormley, and Thomas Schaaf. 2021c. 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.

Ilya Loshchilov and Frank Hutter. 2017. Sgdr: Stochastic gradient descent with warm restarts. *Preprint*, arXiv:1608.03983.

Natalia Loukachevitch, Suresh Manandhar, Elina Baral, Igor Rozhkov, Pavel Braslavski, Vladimir Ivanov, Tatiana Batura, and Elena Tutubalina. 2023. NEREL-BIO: A Dataset of Biomedical Abstracts Annotated with Nested Named Entities. *Bioinformatics*. Btad161.

Natalia Loukachevitch, Andrey Sakhovskiy, and Elena Tutubalina. 2024. Biomedical concept normalization over nested entities with partial UMLS terminology in Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2383–2389, Torino, Italia. ELRA and ICCL.

Chang Lu, Tian Han, and Yue Ning. 2022. Context-aware health event prediction via transition functions on dynamic disease graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4567–4574.

Chang Lu, Chandan Reddy, Ping Wang, and Yue Ning. 2023. Towards semi-structured automatic icd coding via tree-based contrastive learning. *Advances in Neural Information Processing Systems*, 36:68300–68315.

Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. 2019. Mcn: A comprehensive corpus for medical concept normalization. *Journal of Biomedical Informatics*, 92:103132.

Mingyu Derek Ma, Xiaoxuan Wang, Yijia Xiao, Anthony Cuturrufo, Vijay S Nori, Eran Halperin, and Wei Wang. 2025. Memorize and rank: Elevating large language models for clinical diagnosis prediction. *Preprint*, arXiv:2501.17326.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples!

In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Laxmaiah Manchikanti. 2002. Implications of fraud and abuse in interventional pain management. *Pain Physician*, 5(3):320.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018a. 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.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018b. 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. Association for Computational Linguistics.

Alexandr Nesterov, Galina Zubkova, Zulfat Miftahutdinov, Vladimir Kokh, Elena Tutubalina, Artem Shelmanov, Anton Alekseev, Manvel Avetisian, Andrey Chertok, and Sergey Nikolenko. 2022a. RuCCoN: Clinical concept normalization in Russian. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 239–245, Dublin, Ireland. Association for Computational Linguistics.

Alexandr Nesterov, Galina Zubkova, Zulfat Miftahutdinov, Vladimir Kokh, Elena Tutubalina, Artem Shelmanov, Anton Alekseev, Manvel Avetisian, Andrey Chertok, and Sergey Nikolenko. 2022b. RuCCoN: Clinical concept normalization in Russian. pages 239–245.

Aurélie Névéol, Aude Robert, Robert Anderson, Kevin Bretonnel Cohen, Cyril Grouin, Thomas Lavergne, Grégoire Rey, Claire Rondet, and Pierre Zweigenbaum. 2017. Clef ehealth 2017 multilingual information extraction task overview: Icd10 coding of death certificates in english and french. In *Clef (working notes)*, pages 1–17.

NIH. 2023. Nih umls statistics.

Kimberly 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 5 Pt 2:1620–39.

Jong-Ku Park, Ki-Soon Kim, Tae-Yong Lee, Kang-Sook Lee, Duk-Hee Lee, Sun-Hee Lee, Sun-Ha Jee, Il Suh, Kwang-Wook Koh, So-Yeon Ryu, et al. 2000. The accuracy of ICD codes for cerebrovascular diseases in medical insurance claims. *Journal of Preventive Medicine and Public Health*, 33(1):76–82.

Damian Pascual, Sandro Luck, and Roger Wattenhofer. 2021. Towards bert-based automatic icd coding: Limitations and opportunities. *arXiv preprint arXiv:2104.06709*.

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237.

Andrey Sakhovskiy, Natalia Semenova, Artur Kadurin, and Elena Tutubalina. 2024. Biomedical entity representation with graph-augmented multi-objective transformer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4626–4643, Mexico City, Mexico. Association for Computational Linguistics.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5953–5959. International Joint Conferences on Artificial Intelligence Organization.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P. Xing. 2017. Towards automated ICD coding using deep learning. *CoRR*, abs/1711.04075.

Aaron Sonabend, Winston Cai, Yuri Ahuja, Ashwin Ananthakrishnan, Zongqi Xia, Sheng Yu, and Chuan Hong. 2020. Automated ICD coding via unsupervised knowledge integration (unite). *International journal of medical informatics*, 139:104135.

Jürgen Stausberg, Nils Lehmann, Dirk Kaczmarek, and Markus Stein. 2008. Reliability of diagnoses coding with icd-10. *International Journal of Medical Informatics*, 77(1):50–57.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Zeqiang Wang, Yuqi Wang, Haiyang Zhang, Wei Wang, Jun Qi, Jianjun Chen, Nishanth Sastry, Jon Johnson, and Suparna De. 2024. Icdxml: enhancing icd coding with probabilistic label trees and dynamic semantic representations. *Scientific Reports*, 14(1):18319.

Pengtao Xie and Eric Xing. 2018. A neural architecture for automated ICD coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1066–1076, Melbourne, Australia. Association for Computational Linguistics.

Xiancheng Xie, Yun Xiong, Philip S. Yu, and Yangyong Zhu. 2019. Ehr coding with multi-scale feature attention and structured knowledge graph propagation. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.

Keyang Xu, Mike Lam, Jingzhi Pang, Xin Gao, Charlotte Band, Piyush Mathur, Frank Papay, Ashish K. Khanna, Jacek B. Cywinski, Kamal Maheshwari, Pengtao Xie, and Eric P. Xing. 2018. Multimodal machine learning for automated icd coding. *ArXiv*, abs/1810.13348.

Alexander Yalunin, Alexander Nesterov, and Dmitriy Umerenkov. 2022. Rubioroberta: a pre-trained biomedical language model for russian language biomedical text mining. *arXiv preprint arXiv:2204.03951*.

Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. 2022. Knowledge injected prompt based fine-tuning for multi-label few-shot ICD coding. *arXiv preprint arXiv:2210.03304*.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022a. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 808–814.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022b. CODER: knowledge-infused cross-lingual medical term embedding for term normalization. *J. Biomed. Informatics*, 126:103983.

Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. Optimizing bi-encoder for named entity recognition via contrastive learning. In *The Eleventh International Conference on Learning Representations*.

Tianran Zhang, Muhao Chen, and Alex A T Bui. 2020a. Diagnostic prediction with sequence-of-sets representation learning for clinical events. *Artif. Intell. Med. Conf. Artif. Intell. Med.*, 12299:348–358.

Zachariah Zhang, Jingshu Liu, and Narges Razavian. 2020b. Bert-xml: Large scale automated icd coding using bert pretraining. *arXiv preprint arXiv:2006.03685*.

Lingling Zhou, Cheng Cheng, Dong Ou, and Hao Huang. 2020. Construction of a semi-automatic icd-10 coding system. *BMC medical informatics and decision making*, 20:1–12.

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.

A Appendix: Annotation guidelines.

A.1 Task Overview

The task is to review the diagnoses in the BRAT markup system, categorize them into separate entities corresponding to individual nosologic units, and assign each of the selected entities an identifier in the form of an ICD code from the provided clinical modification of the ICD-10-CM classifier. The purpose of the annotation is to assign the correct, most private (to the extent possible from the limited anamnesis cotext) identifier to each nosologic unit represented in the diagnosis.

A.2 Data and resources

Data. The documents you will be annotating are anonymized diagnoses. To facilitate and speed up the annotation process, most nosologic units are highlighted and pre-labeled with an ICD code.

Vocabulary. Each phrase identified in the text as a nosological unit or not highlighted but being such must be associated with a code from the ICD-10. This markup will use the clinical modification of the ICD-10-CM, which includes about 17762 different medical diagnoses.

Additional Resources. Although the markup system is already loaded with the ICD-10, you can use the following additional resources to help you correctly identify the most appropriate ICD code:

- The ICD Code Clinical Modification Version 10 is a Russian-language web service for searching and determining the optimal ICD code, available at: www.mkb-10.com. Registration is not required to access this resource.
- Google - You can use Google if you are unfamiliar with a clinical diagnosis or if you encounter a previously unknown abbreviation or acronym.
- Wikipedia - You can also use Wikipedia to find additional information.

A.3 Task Description

For each selected or unselected piece of text corresponding to a nosological unit, you need to as-

sign an ICD code. *Example*: “Atopic dermatitis in partial remission disseminated form”. The selected text fragment “Atopic dermatitis in partial remission” should be associated with the diagnosis "Other atopic dermatitis" (L20.8). Make sure that no text fragment representing a nosological unit is left without an assigned ICD code, thus ensuring the completeness of the markup.

However, each nosologic unit should correspond to only one code. However, in many cases, the selected nosologic units may correspond to more than one ICD code, in which case you should follow the following rules:

- 1. Select an ICD code that maximizes the specificity of the diagnosis up to subsection X.00.
- 2. If the nosological unit includes modifiers such as “mild”, “severe”, “acute”, “chronic”, indication of degree, stage, etc., the modifier should be taken into account when searching for the appropriate ICD code. However, it is often the case that the classifier will only have a more general diagnosis that does not include the above modifier. In this case, select the optimal ICD code by ignoring the modifier. However, modifiers that are inseparable in meaning from the underlying concept should always be considered when selecting the optimal ICD code (e.g., “Acute myocardial ischemia”).

The following rules should also be followed when marking up:

- If the selected nosological unit is written in the plural and the corresponding ICD code exists in the classifier in the plural, you should select it. Otherwise, you should search for the ICD code in the singular.
- Sometimes in the classifier there are diagnoses that at first glance seem to be absolutely identical, which can be differentiated only by the context of the electronic medical record.

A.4 Annotation Tool

The annotation process is conducted using a specialized web service called brat (https://brat.nlplab.org/). You will be provided with a customized login and password. All necessary information from clinical diagnoses and preliminary markup with ICD codes are entered into the annotation tool. Each document in the brat web service leads to a separate clinical diagnosis.

Each selected text fragment is a nosological unit to be associated with the corresponding ICD code. In order to call the ICD code selection menu, you need to highlight the section of text you are going to mark up or double-click on the green label “icd_code” located above the selected text fragment. If you think that a section of the diagnosis is selected incorrectly or redundantly, you need to correct or delete the corresponding selection.

The window may or may not have a pre-selected ICD code on the Ref line. If specified, compare the correctness of the ICD code specified in the “Ref” line with the selected text fragment specified in the “Text” field. If the ICD code is correct, press the “OK” button and move to the next selected text fragment. If the ICD-code is not specified or is specified incorrectly, double-click the “Ref” line in the “Normalization” field, and the ICD-code search window will open. In the opened window check the correctness of the diagnosis selection for search in the “Query” line and click on the “Search ICD_codes” button. The system will search in the ICD codes classifier and list them. If the system does not find the codes by the specified text fragment, try to change it.

Select the appropriate ICD code and its decoding from the list and press the “OK” button (or double-click on the required ICD code). The system will save your selection and return to the previous window, where you should also click on the “OK” button. The system will remember your selection and you can proceed to annotate the next selected text section.

If you did not find a suitable ICD code in the list of ICD codes found by the system, you can try to change the search phrase in the “Query” field, by which the search is performed, and perform the search again. In most cases, the correct selection of the search phrase allows one to find the most appropriate ICD code in the classifier.

If the built-in search system does not yield results, you can switch to the external directory of ICD codes specified in A2. To do this, click on the magnifying glass icon in the “Normalization” field. You can also go to the Google search engine and Wikipedia web encyclopedia by clicking on the corresponding link in the “Search” field.

If even after changing the search phrase and searching in external resources you cannot find a suitable ICD code, return to the previous menu by clicking on the “cancel” button and delete the identifier located in the ID line in the “Normalization” field in the opened window. The same should

be done if a text section that is not a nosological unit is selected. Deleting the identifier will clear the "Ref" line; this will serve as an indicator that the selected text fragment could not be matched with a suitable ICD code.

B Related Work

In describing our work, we encountered persistent terminological ambiguity arising from overlapping nomenclature for distinct task formulations. For instance, the term "ICD coding" is broadly applied to both (1) multi-label classification of medical texts (e.g., assigning ICD codes to discharge summaries) (Li and Yu, 2020; Vu et al., 2020; Wang et al., 2024) and (2) entity linking, where discrete clinical diagnoses are mapped to specific codes (Lavergne et al., 2016; Névéol et al., 2017; Coutinho and Martins, 2022). This conflation obscures fundamental differences: the former treats coding as document-level prediction to capture all relevant codes for a patient's condition, while the latter focuses on precise alignment of clinical entities (e.g., distinguishing "acute myocardial infarction" from its subtypes) through semantic matching, addressing challenges like synonymy or hierarchical code relationships. To resolve this ambiguity, in our work we propose explicit terminology: **"ICD coding"** refers to multi-label classification of medical texts, whereas **"Medical entity linking"** denotes entity-level code assignment.

ICD coding ICD coding has traditionally relied on established machine learning techniques. Early approaches employed methods such as Support Vector Machines (SVM) with TF-IDF features to represent clinical notes (Perotte et al., 2014). Feature engineering, including gradient boosting for large datasets, also played a significant role in enhancing ICD coding accuracy (Diao et al., 2021). Regular expression-based mapping and adaptive data processing further improved efficiency in specific healthcare settings (Zhou et al., 2020).

The advent of neural networks marked a paradigm shift in ICD coding. Recurrent Neural Networks (RNNs), including LSTMs and GRUs, were utilized to encode EHR data and capture temporal dependencies within clinical notes (Choi et al., 2016; Baumel et al., 2018). Convolutional Neural Networks (CNNs) offered alternative architectures for extracting features from clinical text, with models like CAML demonstrating their effectiveness (Mullenbach et al., 2018b). Subsequent advancements introduced multi-filter CNNs (Li and Yu, 2020) and squeeze-and-excitation networks in CNN (Liu et al., 2021c) to enhance feature extraction. Addressing the challenge of imbalanced code distribution, researchers introduced focal loss (Liu et al., 2021c) and self-distillation mechanisms to improve prediction accuracy for rare codes (Zhou et al., 2021). Other models, like HA-GRUs used the charachter-level information (Baumel et al., 2018). Ensemble models used CNN, LSTM, and decision trees to improve accuracy (Xu et al., 2018).

A crucial line of research has focused on integrating external medical knowledge and the inherent hierarchical structure of ICD codes. Approaches have incorporated medical definitions (Shi et al., 2017), Wikipedia data for rare diseases (Bai and Vucetic, 2019) and medical ontologies (Bao et al., 2021) to enrich term embeddings. Tree-of-sequences LSTMs (Xie and Xing, 2018) and graph neural networks (Cao et al., 2020; Xie et al., 2019) were developed to capture relationships between codes, either through hierarchical structures or co-occurrence patterns. Models like KG-MultiResCNN leveraged external knowledge for relations understanding (Boukhers et al., 2023). Weak supervision was used to overcome the lack of training data (Dong et al., 2021; Gao et al., 2022). Furthermore, domain-specific pre-trained language models (PLMs) such as BioBERT (Lee et al., 2019), ClinicalBERT (Alsentzer et al., 2019), and PubMedBERT (Gu et al., 2021) have shown promise in improving performance on various biomedical tasks. However, adapting these models to the large-scale, multi-label nature of ICD coding presents unique challenges, particularly regarding long input sequences (Pascual et al., 2021; Ji et al., 2021). Recent efforts, such as BERT-XML (Zhang et al., 2020b), have addressed this through input splitting and label attention mechanisms. Read, Attend, and Code (RAC) was proposed by Kim and Ganapathi (Kim and Ganapathi, 2021) and achieved state-of-the-art results. Despite these developments, challenges remain in handling semi-structured text and variability of notes (Lu et al., 2023).

Recent studies have increasingly focused on leveraging attention mechanisms and improving the interaction between clinical note representations and ICD code representations. Models such as LAAT (Vu et al., 2020) and EffectiveCAN (Liu et al., 2021c) have incorporated refined label-aware attention mechanisms. However, the effective application of PLMs to ICD coding requires careful

consideration of input length constraints and the development of robust mechanisms for capturing long-range dependencies. Also, the models need to better understand relationships between different sections of clinical notes (Lu et al., 2023).

Diagnosis prediction Diagnosis prediction using structured EHR data has been extensively studied with deep learning approaches. NECHO (Koo, 2024) improves next-visit diagnosis prediction by centering learning on medical codes and incorporating hierarchical regularization to capture structured dependencies in EHR data. DPSS (Zhang et al., 2020a) enhances predictive robustness by modeling patient records as sequences of unordered clinical events, preserving temporal patterns while mitigating biases introduced by the artificial ordering of medical records. The importance of patient history in EHR-based diagnosis prediction demonstrates that historical records alone can achieve 76.6% accuracy, which increases to 93.3% when structured physical examination and laboratory data are integrated (Fukuzawa et al., 2024). At the population level, applying a Bi-GRU model trained on structured EHR data with SNOMED embeddings to predict chronic disease onset demonstrates the utility of structured clinical histories in early disease identification (Grout et al., 2024). To optimize the use of structured medical codes for diagnosis prediction, MERA (Ma et al., 2025) introduces hierarchical contrastive learning and ranking mechanisms to refine diagnosis classification within large ICD code spaces. These studies collectively illustrate the evolution of EHR-based diagnosis prediction from sequence modeling to hierarchical representation learning, highlighting the role of structured clinical history in improving predictive accuracy.

RAG LLMs face challenges as standalone systems for high-precision tasks such as ICD-linking, primarily due to their limited accuracy in extracting detailed, domain-specific information. Ma et al.(Ma et al., 2023) demonstrated that while LLMs lag behind fine-tuned SLMs in information extraction tasks, they excel in understanding and reorganizing semantic content, making them effective at reranking retrieved information. To overcome the limitations of accuracy and domain specificity, recent approaches have incorporated Retrieval-Augmented Generation (RAG) techniques. RAG combines the structured knowledge of external databases for retrieval with the semantic reasoning strengths of LLMs for reranking, resulting in improved precision and overall task performance.

Klang et al. (Klang et al., 2024) demonstrated the effectiveness of RAG in enhancing LLMs for ICD-10-CM medical coding. Their study revealed that RAG-enhanced LLMs outperform human coders in accuracy and specificity, emphasizing the potential of retrieval mechanisms in improving clinical documentation. Similarly, Kwan (Kwan, 2024) proposed a two-stage Retrieve-Rank system for medical coding, achieving a perfect match rate for ICD-10-CM codes and significantly surpassing vanilla LLMs. The MedCodER framework (Baksi et al., 2024) leverages a pipeline of extraction, retrieval, and reranking, to improve automation and interpretability in ICD-10 coding. It demonstrates SOTA performance on ACI-BENCH by integrating LLMs with semantic search and evidence-based reasoning. Boyle et al. (Boyle et al., 2023) presented a zero-shot ICD coding approach using LLMs and a tree-search strategy, achieving a SOTA on the CodiEsp dataset, particularly excelling in rare code prediction without task-specific training. Abdulnazar et al. (Abdulnazar et al., 2024) applied GPT-4 for clinical text cleansing to enhance MCN. By combining text standardization with RAG, their method improved mapping precision to SNOMED CT in the German language.

C BERT-based NER Results

Tab. 5 presents evaluation results for NER task on the RuCCoD dataset. In the context of NER, RuBioBERT employs a softmax activation function in its output layer. BINDER utilizes RuBioBERT backbone and approaches NER as a representation learning problem by maximizing the similarity between the vector representations of an entity mention and its corresponding type (Zhang et al.). RuBioBERT achieves the highest F1-score of 0.756 when trained on the RuCCoD, suggesting that this dataset is particularly effective for the model. BINDER trained on RuCCoD achieves an F1-score of 0.71, slightly lower than RuBioBERT trained on the same dataset.

D Entity Linking Results

Since there are many datasets for entity linking in the biomedical domain, including corpora in Russian, we explored whether these corpora can be helpful for ICD coding. Additionally, we at-

Task	Model or Approach	LR	# Epochs	BS	Scheduler	WD
NER	RuBioBERT	1e-5	20	32	Cosine (Loshchilov and Hutter, 2017)	0.01
EL	BERGAMOT+BioSyn	2e-5	20	32	Adam (Kingma and Ba, 2015)	0.01
LLM tuning	LoRA	5e-5	33	2	Linear with Warmup	0.01
ICD code prediction	Longformer	5e-5	2	4	Linear with Warmup	0.01

Table 4: Models and training hyperparameters. LR stands for learning rate, BS for batch size, WD for weight decay

Model	**Train Data**	**F1-score**	**Precision**	**Recall**
RuBioBERT	RuCCoD train	0.756	0.75	0.77
RuBioBERT	BIO-NNE train	0.62	0.57	0.67
RuBioBERT	RuCCoD + BioNNE train	0.72	0.75	0.70
BINDER + RuBioBERT	RuCCoD train	0.71	0.72	0.71

Table 5: Evaluation results for NER task on RuCCoD dataset.

Train set	**SapBERT**		**CODER**		**BERGAMOT**	
	@1	@5	@1	@5	@1	@5
Zero-shot evaluation, strict						
ICD dict	0.3327	**0.5712**	0.2631	0.4687	0.3495	**0.6170**
ICD dict+UMLS synonyms	**0.3546**	0.5197	**0.3237**	**0.4765**	**0.3559**	0.5487
Supervised evaluation, strict						
ICD	**0.6132**	**0.8182**	**0.6202**	**0.8169**	**0.6415**	**0.8459**
ICD+UMLS sumonyms	0.5326	0.7382	0.5358	0.7318	0.4984	0.7253
RuCCoN	0.3591	0.5345	0.3598	0.5732	0.3643	0.5313
RuCCoN+ICD	0.3952	0.5732	0.3888	0.6570	0.3817	0.5983
NEREL-BIO	0.3443	0.4913	0.3378	0.5274	0.3353	0.5113
NEREL-BIO+ICD	0.3804	0.5596	0.3804	0.6325	0.3598	0.5525
Zero-shot evaluation, relaxed						
ICD dict	0.4842	**0.6886**	0.3752	0.6190	0.5035	**0.7286**
ICD dict+UMLS synonyms	**0.5551**	0.6867	**0.5055**	**0.6293**	**0.5603**	0.7073
Supervised evaluation, relaxed						
ICD	0.7763	**0.8839**	**0.7872**	0.8743	**0.7917**	**0.8943**
ICD+UMLS sumonyms	**0.7788**	0.8616	0.7714	**0.8860**	0.7449	0.8738
RuCCoN	0.5235	0.6531	0.5429	0.7208	0.5132	0.6564
RuCCoN+ICD	0.5493	0.6602	0.5770	0.7485	0.5571	0.6873
NEREL-BIO	0.4803	0.6067	0.4958	0.6634	0.4778	0.6170
NEREL-BIO+ICD	0.5455	0.6447	0.5474	0.7292	0.5384	0.6505

Table 6: Cross-domain transfer results for biomedical linking models. Evaluation results for linking models trained on RuCOD, RuCCoN, NEREL-BIO as well as their union. *ICD+UMLS synonyms* stands for ICD train set with the vocabulary enriched with ICD disease name synonyms from the UMLS knowledge base. The best results for each model and set-up are highlighted in **bold**.

tempted to enrich the ICD normalization vocabulary with concept names from the Unified Medical Language System (UMLS) metathesaurus which includes the ICD-10 vocabulary. Specifically, for each ICD code, we find its Concept Unique Identifier (CUI) in UMLS and retrieve all concept names that share the same CUI but are adopted from the source vocabularies different from ICD-10. We

employ the following Russian biomedical corpora for experiments on cross-terminology transfer:

RuCCoN (Nesterov et al., 2022a) is a manually annotated corpus of clinical records in Russian. It contains 16,028 mentions linked to 2,409 unique concepts from the Russian subset of UMLS metathesaurus (Bodenreider, 2004).

NEREL-BIO (Loukachevitch et al., 2023, 2024) is a corpus of 756 PubMed abstracts in Russian manually linked to 4,544 unique UMLS concepts. The corpus is specifically focused on two main problems: (i) entity nestedness and (ii) cross-lingual Russian-to-English normalization for the incomplete Russian UMLS terminology. In total, NEREL-BIO provides 23,641 entity mentions manually linked to 4,544 unique UMLS concepts. 4,424 mentions have no concept name representation in the Russian UMLS subset and are linked to 1,535 unique concepts present in the English UMLS only.

We experiment with three state-of-the-art specialized biomedical entity linking models:

SapBERT is a metric learning framework that learns from synonymous UMLS concept names by generating hard triplets for pre-training (Liu et al., 2021a,b).

CODER is a contrastive learning model inspired by semantic matching methods that use both synonyms and relations from the UMLS (Yuan et al., 2022b) to learn concept representations.

BERGAMOT is an extension of SapBERT which learns concept name-based and graph-based concept representations simultaneously and introduces a cross-modal alignment loss to transfer knowledge from a graph encoder to a BERT-based language encoder (Sakhovskiy et al., 2024). The graph encoder is discarded after the pretraining stage and only a BERT encoder is used for inference.

For supervised entity linking, we adopt BioSyn (Sung et al., 2020), a BERT-based framework that iteratively updates entity representations using synonym marginalization. For each dataset, we trained BioSyn with default hyperparameters for 20 epochs.

Relaxed EL Evaluation We assess two entity linking set-ups: (i) **strict** evaluation which implies an exact match between predicted and ground truth codes and (ii) **relaxed** evaluation with all codes being truncated to 3-symbols codes (corresponding to the second level of hierarchy).

The results of cross-terminology entity linking transfer presented in Tab. 8 reveal a few insightful findings related to linking ICD codes.

Vocabulary Extension is not a Cure While extension of ICD vocabulary consistently gives a slightly improved Accuracy@1 in a zero-shot setting, additional synonyms introduce severe noise in a supervised setting. Specifically, a significant drop of 8.1%, 8.4%, 14.3% Accuracy@1 is observed for SapBERT, CODER, and BERGAMOT, respectively. Even in an unsupervised setting, vocabulary extension drops Accuracy@5 by 5.2% and 6.8% for SapBERT and BERGAMOT, respectively.

Complicated Cross-Terminology Transfer Both training on RuCCoN and NEREL-BIO as well merge of these corpora with RuCCoD do not lead to improvement over zero-shot coding. The finding indicates the specificity and high complexity of ICD coding within the entity linking task.

Complexity of Fine-Grained ICD coding The high gap between the strict and supervised evaluation of around 15% Accuracy@1 indicates that distinguishing between semantically similar diseases sharing the same therapeutic group is a major challenge.

E LLM with RAG results

All LLM with RAG experiments were conducted with a temperature setting of 0 for all LLMs and a top-k value of 15 for the number of retrieved entities from similarity search. The LLMs used are specified in Appx. G. For the embedding model, we utilized BERGAMOT. To construct the vector database, we used dictionaries extracted from NEREL-BIO, RUCCON, the ICD dictionary, and the ICD dictionary combined with RuCCoD. The results are presented in Tables 9 and 10 for strict evaluation, and in Tables 11 and 12 for relaxed evaluation.

For the NER task, the ICD dict.+RuCCoD dataset yielded the best results. The Llama3.1:8b-instruct-fp16 model achieved the highest F-score (0.511), precision (0.580), recall (0.456), and accuracy (0.343). Qwen2.5-7B-Instruct and Llama3-Med42-8B followed with F-scores of 0.495 and 0.491, respectively. In contrast, NEREL-BIO and RUCCON datasets showed significantly lower performance, with F-scores below 0.13 and accuracies under 0.07.

For NER+ICD Linking, the same dataset and model led again, with Llama3.1:8b-instruct-fp16 achieving an F-score of 0.268 and accuracy of 0.155. Qwen2.5-7B-Instruct and Llama3-Med42-8B followed closely with F-scores around 0.245. Performance on NEREL-BIO and RuCCon was much lower, with F-scores under 0.022 and accuracies below 0.011.

For ICD Code assignment, Llama3.1:8b-instruct-fp16 also performed best, with an F-score of 0.458 and accuracy of 0.297. Qwen2.5-7B-Instruct and Llama3-Med42-8B also performed well, with F-scores of 0.463 and 0.457. Again, NEREL-BIO and RUCCON datasets exhibited weaker results, with F-scores below 0.15 and accuracies under 0.09.

In summary, the ICD dict.+RuCCoD dataset consistently outperformed others with Llama3.1:8b-instruct-fp16 being the best model. Relaxed evaluation settings produced similar trends.

F LLM with tuning results

The LLM tuning results are in Tab. 7.

For the NER task, Llama3-Med42-8B achieved the highest F-score of 0.642, which corresponds to the highest Precision and Recall among the models. Phi3_5_mini and Mistral-Nemo demonstrated similar performance (F-scores of 0.627 and 0.614, respectively), but slightly lag behind the leader. The Qwen2.5-7B-Instruct model showed the lowest scores across all metrics, with an F-score of 0.565 and an Accuracy of 0.393.

In the NER + ICD linking task, the use of the RuCCoD or BERGAMOT approach significantly improved the linking performance. For instance, Phi3_5_mini achieved the highest F-score of 0.333 when using RuCCoD, and Llama3-Med42-8B reached an F-score of 0.299. Notably, for all models, the use of RuCCoD proved to be more beneficial than the BERGAMOT approach.

In the ICD code assignment task, results also improved significantly with the use of the RuCCoD dataset. Once again, Phi3_5_mini emerged as the top-performing model, attaining an F-score of 0.480 when using RuCCoD. Llama3-Med42-8B and Mistral-Nemo also demonstrated strong results, with F-scores of 0.435 and 0.446, respectively, when using RuCCoD. It is noteworthy that the inclusion of RuCCoD consistently improved Precision and Recall across all models.

Based on the presented results, it can be concluded that for all tasks (NER, NER+Linking, and ICD code assignment), the use of RuCCoD significantly enhances model performance compared to relying solely on the dictionary or embeddings. The top-performing models across all tasks are Llama3-Med42-8B and Phi3_5_mini, indicating their high efficiency in medical tasks following PEFT tuning.

G Implementation Details

Utilized LLMs:

- `Phi-3.5-mini-instruct` (Phi)
- `Qwen2.5-7B-Instruct` (Qwe)
- `Llama3-Med42-8B` (Med)
- `Mistral-Nemo-Instruct-2407` (Mis)
- `llama3.1:8b-instruct-fp16` (Lla)

Diagnosis prediction Each Longformer was trained for two epochs on separate NVidia A100 GPUs, with the fine-tuning process taking approximately one week per model. We provide hyperparameters for these models training in Tab. 4.

Hyperparameters A detailed overview, including parameter values and configurations, is provided in Tab. 4.

H Prompts

The original prompts were in Russian. Below are their translations to English.

NER prompt

```
You will be provided with a text
containing diagnoses. Extract the
diagnoses from this text. Do not
alter the spelling of the
diagnoses in the text. Respond
only in the format of a list: ['
diagnosis1', 'diagnosis2', ...]
Text: {text}
```

Model	Precision	Recall	F-score	Accuracy
NER				
Llama3-Med42-8B, RuCCoD	**0.642**	**0.642**	**0.642**	**0.473**
Qwen2.5-7B-Instruct, RuCCoD	0.567	0.562	0.565	0.393
Phi3_5_mini, RuCCoD	0.632	0.623	0.627	0.457
Mistral-Nemo, RuCCoD	0.631	0.598	0.614	0.443
NER+Linking				
Llama3-Med42-8B, ICD dict.	0.149	0.149	0.149	0.08
Llama3-Med42-8B, ICD dict. + RuCCoD	0.299	0.299	0.299	0.176
Llama3-Med42-8B, ICD dict. + BERGAMOT	0.286	0.286	0.286	0.167
Qwen2.5-7B-Instruct, ICD dict.	0.188	0.186	0.187	0.103
Qwen2.5-7B-Instruct, ICD dict. + RuCCoD	0.281	0.279	0.28	0.163
Qwen2.5-7B-Instruct, ICD dict. + BERGAMOT	0.2	0.198	0.199	0.11
Phi3_5_mini, ICD dict.	0.272	0.268	0.27	0.156
Phi3_5_mini, ICD dict. + RuCCoD	**0.335**	**0.33**	**0.333**	**0.199**
Phi3_5_mini, ICD dict. + BERGAMOT	0.322	0.317	0.32	0.19
Mistral-Nemo, ICD dict.	0.231	0.219	0.224	0.126
Mistral-Nemo, ICD dict. + RuCCoD	0.303	0.287	0.295	0.173
Mistral-Nemo, ICD dict. + BERGAMOT	0.267	0.253	0.26	0.149
Code assignment				
Llama3-Med42-8B, ICD dict.	0.229	0.231	0.23	0.13
Llama3-Med42-8B, ICD dict. + RuCCoD	0.434	0.435	0.435	0.278
Llama3-Med42-8B, ICD dict. + BERGAMOT	0.403	0.405	0.404	0.253
Qwen2.5-7B-Instruct, ICD dict.	0.296	0.295	0.295	0.173
Qwen2.5-7B-Instruct, ICD dict. + RuCCoD	0.456	0.449	0.452	0.292
Qwen2.5-7B-Instruct, ICD dict. + BERGAMOT	0.305	0.303	0.304	0.179
Phi3_5_mini, ICD dict.	0.394	0.39	0.392	0.244
Phi3_5_mini, ICD dict. + RuCCoD	**0.483**	**0.477**	**0.48**	**0.316**
Phi3_5_mini, ICD dict. + BERGAMOT	0.454	0.448	0.451	0.291
Mistral-Nemo, ICD dict.	0.326	0.311	0.319	0.189
Mistral-Nemo, ICD dict. + RuCCoD	0.458	0.435	0.446	0.287
Mistral-Nemo, ICD dict. + BERGAMOT	0.394	0.372	0.383	0.237

Table 7: ICD coding results for finetuned LLMs on RuCCoD. The best results are highlighted in **bold**.

Diagnosis selection prompt

```
You will be given a reference
diagnosis and a list of diagnoses
from a database.
Your task is to determine which
diagnosis from the database best
matches the reference diagnosis.
Try to select the diagnosis
accurately, paying attention to
details. Choose the diagnosis with
 the highest match in terms of
words and meaning.
You can only choose from the
diagnoses in the list.
Pay more attention to the
diagnoses at the beginning of the
list, as they are more likely to
be a better match.
It's better to choose a shorter
diagnosis than one that includes
information not present in the
reference diagnosis.
In your response, write only the
```

Model	Precision	Recall	F-score	Accuracy
NER				
BioBERT, Biosyn, RuCCoD	0.649	0.655	0.653	0.485
BioBERT, RuCCoD	**0.721**	**0.769**	**0.744**	**0.592**
BioBERT, NEREL-BIO	0.588	0.675	0.628	0.458
BioBERT, NEREL-BIO, RuCCoD	0.689	0.713	0.701	0.54
BioBERT, RuCCoN	0.637	0.613	0.625	0.454
BioBERT, RuCCoN + RuCCoD	0.609	0.709	0.655	0.487
NER+Linking				
BioBERT, Biosyn, RuCCoD	0.392	0.396	0.394	0.245
BioBERT, RuCCoD	**0.427**	**0.455**	**0.441**	**0.283**
BioBERT, NEREL-BIO	0.353	0.406	0.377	0.233
BioBERT, NEREL-BIO, RuCCoD	0.406	0.42	0.413	0.26
BioBERT, RuCCoN	0.387	0.372	0.379	0.234
BioBERT, RuCCoN + RuCCoD	0.351	0.409	0.378	0.233
Code assignment				
BioBERT, Biosyn, RuCCoD	0.507	0.508	0.507	0.340
BioBERT, RuCCoD	0.51	0.542	**0.525**	**0.356**
BioBERT, NEREL-BIO	0.466	0.531	0.497	0.33
BioBERT, NEREL-BIO, RuCCoD	**0.512**	0.529	0.52	0.352
BioBERT, RuCCoN	0.508	0.485	0.496	0.33
BioBERT, RuCCoN + RuCCoD	0.471	**0.543**	0.504	0.337

Table 8: Evaluation results for entity-level tasks for BERT-based IE pipeline on RuCCoD corpus. The best results are highlighted in **bold**.

Model	Precision	Recall	F-score	Accuracy
NER: ICD dict.				
Llama3.1:8b-instruct	0.208	0.088	0.124	0.066
Llama3-Med42-8B	0.202	0.084	0.118	0.063
Phi-3.5-mini-instruct	**0.211**	**0.093**	**0.129**	**0.069**
Mistral-Nemo-Instruct-2407	0.198	0.072	0.105	0.055
Qwen2.5-7B-Instruct	0.206	0.087	0.122	0.065
NER: ICD dict. + RuCCoD				
Llama3.1:8b-instruct	**0.581**	**0.456**	**0.511**	**0.343**
Llama3-Med42-8B	0.556	0.441	0.492	0.326
Phi-3.5-mini-instruct	0.543	0.450	0.492	0.326
Mistral-Nemo-Instruct-2407	0.541	0.372	0.441	0.283
Qwen2.5-7B-Instruct	0.566	0.440	0.495	0.329
NER+Linking: ICD dict.				
Llama3.1:8b-instruct	**0.071**	0.067	**0.069**	**0.036**
Llama3-Med42-8B	0.058	0.063	0.060	0.031
Phi-3.5-mini-instruct	0.062	**0.069**	0.065	0.034
Mistral-Nemo-Instruct-2407	0.066	0.056	0.060	0.031
Qwen2.5-7B-Instruct	0.065	0.065	0.065	0.033
NER+Linking: ICD dict. + RuCCoD				
Llama3.1:8b-instruct	**0.272**	**0.264**	**0.268**	**0.155**
Llama3-Med42-8B	0.235	0.261	0.247	0.141
Phi-3.5-mini-instruct	0.228	0.257	0.242	0.137
Mistral-Nemo-Instruct-2407	0.247	0.215	0.230	0.130
Qwen2.5-7B-Instruct	0.244	0.246	0.245	0.140
Code assignment: ICD dict.				
Llama3.1:8b-instruct	0.379	0.363	0.371	0.228
Llama3-Med42-8B	0.310	0.345	0.327	0.195
Phi-3.5-mini-instruct	0.260	0.294	0.276	0.160
Mistral-Nemo-Instruct-2407	**0.413**	0.360	0.385	0.238
Qwen2.5-7B-Instruct	0.401	**0.411**	**0.406**	**0.255**
Code assignment: ICD dict. + RuCCoD				
Llama3.1:8b-instruct	**0.465**	0.451	0.458	0.297
Llama3-Med42-8B	0.434	**0.483**	0.457	0.296
Phi-3.5-mini-instruct	0.409	0.458	0.432	0.276
Mistral-Nemo-Instruct-2407	0.462	0.401	0.429	0.273
Qwen2.5-7B-Instruct	0.461	0.465	**0.463**	**0.301**

Table 9: Evaluation results for NER, Code assignment, and end-to-end entity linking task on RuCCoD for LLM+RAG pipeline.

Model	Precision	Recall	F-score	Accuracy
NER: NEREL-BIO				
Llama3.1:8b-instruct	0.100	0.042	0.059	0.030
Llama3-Med42-8B	0.104	0.043	0.060	0.031
Phi-3.5-mini-instruct	0.098	0.043	0.059	0.031
Mistral-Nemo-Instruct-2407	**0.115**	**0.044**	**0.063**	**0.033**
Qwen2.5-7B-Instruct	0.099	0.043	0.060	0.031
NER: RuCCoN				
Llama3.1:8b-instruct	0.188	0.088	0.120	0.064
Llama3-Med42-8B	0.174	0.079	0.108	0.057
Phi-3.5-mini-instruct	0.172	0.085	0.114	0.060
Mistral-Nemo-Instruct-2407	**0.197**	0.082	0.116	0.061
Qwen2.5-7B-Instruct	0.185	**0.091**	**0.122**	**0.065**
NER+Linking: NEREL-BIO				
Llama3.1:8b-instruct	0.023	**0.020**	0.021	**0.011**
Llama3-Med42-8B	0.018	0.019	0.018	0.009
Phi-3.5-mini-instruct	0.019	**0.020**	0.019	0.010
Mistral-Nemo-Instruct-2407	**0.025**	**0.020**	**0.022**	**0.011**
Qwen2.5-7B-Instruct	0.021	**0.020**	0.020	0.010
NER+Linking: RuCCoN				
Llama3.1:8b-instruct	0.050	**0.046**	**0.048**	**0.025**
Llama3-Med42-8B	0.042	0.044	0.043	0.022
Phi-3.5-mini-instruct	0.038	0.041	0.040	0.020
Mistral-Nemo-Instruct-2407	**0.053**	0.044	**0.048**	**0.025**
Qwen2.5-7B-Instruct	0.048	**0.046**	0.047	0.024
Code assignment: NEREL-BIO				
Llama3.1:8b-instruct	0.059	0.053	0.056	**0.029**
Llama3-Med42-8B	0.045	0.047	0.046	0.024
Phi-3.5-mini-instruct	0.046	0.049	0.047	0.024
Mistral-Nemo-Instruct-2407	**0.062**	0.051	0.056	**0.029**
Qwen2.5-7B-Instruct	0.058	**0.056**	**0.057**	**0.029**
Code assignment: RuCCoN				
Llama3.1:8b-instruct	**0.164**	**0.150**	**0.157**	**0.085**
Llama3-Med42-8B	0.125	0.131	0.128	0.068
Phi-3.5-mini-instruct	0.125	0.134	0.129	0.069
Mistral-Nemo-Instruct-2407	0.156	0.129	0.141	0.076
Qwen2.5-7B-Instruct	0.156	0.152	0.154	0.084

Table 10: Evaluation results for NER, Code assignment, and end-to-end entity linking task on RuCCoD for LLM+RAG pipeline using NEREL-BIO and RuCCoN for vectorstore.

Model	Precision	Recall	F-score	Accuracy
NER: ICD dict.				
Llama3.1:8b-instruct	0.208	0.088	0.124	0.066
Llama3-Med42-8B	0.202	0.084	0.118	0.063
Phi-3.5-mini-instruct	**0.211**	**0.093**	**0.129**	**0.069**
Mistral-Nemo-Instruct-2407	0.198	0.072	0.105	0.055
Qwen2.5-7B-Instruct	0.206	0.087	0.122	0.065
NER: ICD dict. + RuCCoD				
Llama3.1:8b-instruct	**0.581**	**0.456**	**0.511**	**0.343**
Llama3-Med42-8B	0.556	0.441	0.492	0.326
Phi-3.5-mini-instruct	0.543	0.450	0.492	0.326
Mistral-Nemo-Instruct-2407	0.541	0.372	0.441	0.283
Qwen2.5-7B-Instruct	0.566	0.440	0.495	0.329
NER+Linking: ICD dict.				
Llama3.1:8b-instruct	**0.095**	0.088	**0.091**	**0.048**
Llama3-Med42-8B	0.077	0.083	0.080	0.042
Phi-3.5-mini-instruct	0.083	**0.092**	0.087	0.046
Mistral-Nemo-Instruct-2407	0.083	0.070	0.076	0.040
Qwen2.5-7B-Instruct	0.087	0.086	0.087	0.045
NER+Linking: ICD dict. + RuCCoD				
Llama3.1:8b-instruct	**0.378**	**0.362**	**0.369**	**0.227**
Llama3-Med42-8B	0.324	0.354	0.338	0.203
Phi-3.5-mini-instruct	0.323	0.357	0.339	0.204
Mistral-Nemo-Instruct-2407	0.342	0.295	0.317	0.188
Qwen2.5-7B-Instruct	0.343	0.340	0.342	0.206
Code assignment: ICD dict.				
Llama3.1:8b-instruct	0.575	0.561	0.568	0.396
Llama3-Med42-8B	0.523	0.594	0.556	0.385
Phi-3.5-mini-instruct	0.437	0.510	0.471	0.308
Mistral-Nemo-Instruct-2407	**0.598**	0.533	0.564	0.392
Qwen2.5-7B-Instruct	0.595	**0.618**	**0.607**	**0.435**
Code assignment: ICD dict. + RuCCoD				
Llama3.1:8b-instruct	**0.701**	0.684	0.692	0.529
Llama3-Med42-8B	0.644	**0.720**	0.680	0.515
Phi-3.5-mini-instruct	0.627	0.703	0.663	0.496
Mistral-Nemo-Instruct-2407	0.691	0.605	0.645	0.476
Qwen2.5-7B-Instruct	0.700	0.704	**0.702**	**0.541**

Table 11: Relaxed evaluation results for NER, Code assignment, and end-to-end entity linking task on RuCCoD for LLM+RAG pipeline.

Model	Precision	Recall	F-score	Accuracy
NER: NEREL-BIO				
Llama3.1:8b-instruct-fp16	0.100	0.042	0.059	0.030
Llama3-Med42-8B	0.104	0.043	0.060	0.031
Phi-3.5-mini-instruct	0.098	0.043	0.059	0.031
Mistral-Nemo-Instruct-2407	**0.115**	**0.044**	**0.063**	**0.033**
Qwen2.5-7B-Instruct	0.099	0.043	0.060	0.031
NER: RuCCoN				
Llama3.1:8b-instruct-fp16	0.188	0.088	0.120	0.064
Llama3-Med42-8B	0.174	0.079	0.108	0.057
Phi-3.5-mini-instruct	0.172	0.085	0.114	0.060
Mistral-Nemo-Instruct-2407	**0.197**	0.082	0.116	0.061
Qwen2.5-7B-Instruct	0.185	**0.091**	**0.122**	**0.065**
NER+Linking: NEREL-BIO				
Llama3.1:8b-instruct	**0.033**	**0.029**	**0.031**	**0.016**
Llama3-Med42-8B	0.024	0.025	0.025	0.013
Phi-3.5-mini-instruct	0.026	0.028	0.027	0.014
Mistral-Nemo-Instruct-2407	**0.033**	0.027	0.030	0.015
Qwen2.5-7B-Instruct	0.030	**0.029**	0.030	0.015
NER+Linking: RuCCoN				
Llama3.1:8b-instruct	**0.076**	0.069	**0.072**	**0.038**
Llama3-Med42-8B	0.061	0.063	0.062	0.032
Phi-3.5-mini-instruct	0.060	0.064	0.062	0.032
Mistral-Nemo-Instruct-2407	**0.076**	0.062	0.068	0.035
Qwen2.5-7B-Instruct	0.073	**0.070**	**0.072**	0.037
Code assignment: NEREL-BIO				
Llama3.1:8b-instruct	0.114	0.107	0.110	0.058
Llama3-Med42-8B	0.088	0.096	0.092	0.048
Phi-3.5-mini-instruct	0.098	0.110	0.104	0.055
Mistral-Nemo-Instruct-2407	0.121	0.105	0.112	0.059
Qwen2.5-7B-Instruct	**0.125**	**0.126**	**0.125**	**0.067**
Code assignment: RuCCoN				
Llama3.1:8b-instruct	**0.295**	0.282	0.288	0.168
Llama3-Med42-8B	0.254	0.275	0.264	0.152
Phi-3.5-mini-instruct	0.248	0.273	0.260	0.149
Mistral-Nemo-Instruct-2407	0.284	0.244	0.263	0.151
Qwen2.5-7B-Instruct	0.292	**0.294**	**0.293**	**0.172**

Table 12: Relaxed evaluation results for NER, Code assignment, and end-to-end entity linking task on RuCCoD for LLM+RAG pipeline using NEREL-BIO and RuCCoN for vectorstore.