Overview of ElCardioCC Task on Clinical Coding in Cardiology at BioASQ 2025

Dimitris Dimitriadis[1,*], Vasiliki Patsiou[2], Eleonora Stoikopoulou[1,†], Achilleas Toumpas[1,†], Alkis Kipouros[1,†], Alexandra Bekiaridou[2], Konstantinos Barmpagiannos[1], Anthi Vasilopoulou[1], Antonios Barmpagiannos[1], Athanasios Samaras[1], Dimitrios Papadopoulos[1], George Giannakoulas[1] and Grigorios Tsoumakas[1,3]

[1]*Aristotle University of Thessaloniki, Greece*
[2]*Elmezzi Graduate School of Molecular Medicine, Northwell Health, Manhasset, NY, USA*
[3]*Archimedes, Athena Research Center, Greece*

Abstract

Automated clinical coding converts unstructured medical narratives into standardized formats like ICD-10, supporting research, data analysis, and healthcare management. While much progress has been made for English texts, languages like Greek remain underexplored, limiting the applicability of such tools in non-English clinical settings. To address this gap, we introduce the ELCardioCC task, part of the BioASQ 2025 challenge. ELCardioCC is a shared task focusing on automated ICD-10 coding of Greek cardiology discharge letters and extraction of supporting text spans. It includes three sub-tasks— named entity recongition, entity linking, and multi-label learning with explainable AI—to promote research in underrepresented languages and support the development of medical natural language processing tools beyond English. The task focuses on (i) assigning ICD-10 codes related to cardiology to Greek hospital discharge letters, and (ii) extracting the exact text spans corresponding to each code. ELCardioCC attracted five participating teams with multiple system submissions. A Greek-language clinical dataset of 1,500 de-identified cardiology discharge letters was created and annotated with ICD-10 codes to support these sub-tasks. The dataset includes both structured metadata and unstructured clinical narratives, with annotations performed by medical professionals using a standardized process. Results indicate that state-of-the-art models can be effectively adapted for Greek clinical texts, demonstrating their practical potential for multilingual medical coding; however, further improvements are necessary to achieve optimal performance and fully address the task's challenges.

Keywords

Clinical coding, named entity recognition, entity linking, multi-label learning, explainable ai, BioASQ shared task

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, accounting for approximately 32% of all global deaths, according to the World Health Organization (WHO)[1]. These conditions comprise a heterogeneous group of pathologies—including ischemic heart disease, heart failure, arrhythmias, and cerebrovascular events—that require long-term management and continuous

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
dndimitri@csd.auth.gr (D. Dimitriadis); spatsiou19@gmail.com (V. Patsiou); stoikopoyloyeleonwra@gmail.com (E. Stoikopoulou); toumpasaxilleas@gmail.com (A. Toumpas); kipourosalkis@gmail.com (A. Kipouros); ampekiaridou@gmail.com (A. Bekiaridou); kostasmparmp@hotmail.gr (K. Barmpagiannos); anthoni2507@gmail.com (A. Vasilopoulou); antonismparmpagiannos@hotmail.com (A. Barmpagiannos); ath.samaras.as@gmail.com (A. Samaras); dpapadopo@csd.auth.gr (D. Papadopoulos); g.giannakoulas@gmail.com (G. Giannakoulas); greg@csd.auth.gr (G. Tsoumakas)
https://dndimitri.eu (D. Dimitriadis); https://intelligence.csd.auth.gr/people/tsoumakas/ (G. Tsoumakas)
0000-0002-9404-0331 (D. Dimitriadis); 0000-0002-3444-4537 (V. Patsiou); 0009-0009-7157-0725 (E. Stoikopoulou); 0009-0008-6028-5622 (A. Toumpas); 0009-0004-2831-2695 (A. Kipouros); 0000-0002-1143-2293 (A. Bekiaridou); 0009-0003-1476-9793 (K. Barmpagiannos); 0009-0009-4956-330X (A. Vasilopoulou); 0009-0006-9976-9080 (A. Barmpagiannos); 0000-0002-3404-2749 (A. Samaras); 0000-0001-6292-2922 (D. Papadopoulos); 0000-0001-7491-6319 (G. Giannakoulas); 0000-0002-7879-669X (G. Tsoumakas)

[1]https://www.who.int/health-topics/cardiovascular-diseases

clinical monitoring. As a result, they generate substantial volumes of medical documentation, often stored as unstructured free-text in electronic health records (EHRs) [1]. The ability to systematically extract and organize clinically relevant information from these records is increasingly critical for real-time surveillance, quality assurance, and translational research.

A central approach to structuring unstructured clinical data is the assignment of standardized diagnostic codes, such as those defined by the International Classification of Diseases, 10th Revision (ICD-10). These codes support data interoperability, facilitate clinical audits and epidemiological studies, and enable downstream applications including health resource allocation, risk stratification, and outcome prediction. However, the manual coding process is labor-intensive, expensive, and susceptible to significant intra- and inter-annotator variability. Prior studies have demonstrated that even experienced coders frequently diverge in how they map free-text narratives to granular ICD-10 categories [2].

To overcome these limitations, *automated clinical coding* has emerged as a promising solution. By applying techniques from natural language processing (NLP) and machine learning (ML), it enables the transformation of free-text medical documents into structured, machine-readable codes [3]. This task presents several modeling challenges, such as recognizing domain-specific medical terminology, resolving ambiguous or polysemous expressions, identifying context-sensitive entities, and managing temporality and negation. In the cardiovascular domain, these complexities are further amplified by the frequent presence of comorbidities—such as hypertension, diabetes, and chronic kidney disease—which often interact and complicate the coding process [4].

Moreover, the high dimensionality of the ICD-10 code space—comprising over 70,000 distinct codes—and the inherently multi-label nature of clinical documentation increase both computational and methodological complexity [5]. These challenges are particularly acute in low-resource environments where large, high-quality annotated corpora are scarce or unavailable. As healthcare systems transition to more data-driven infrastructures, automated coding is increasingly viewed as foundational for scalable, efficient, and equitable health information management. Its implementation is especially critical in high-burden domains like cardiology, where precise and timely codification can influence not only individual patient care, but also broader policy and funding decisions [6].

Despite substantial advances in clinical NLP, existing automated coding systems are almost exclusively developed for English-language corpora [7, 8]. This language bias presents a significant barrier to the adoption of automated tools in multilingual healthcare systems. In Greece, for example, the scarcity of annotated clinical corpora, language-adapted NLP tools, and reliable coding benchmarks has hindered progress in clinical text mining and intelligent documentation. Consequently, Greek hospitals and research institutions remain disadvantaged in their ability to adopt modern AI-driven documentation systems, thereby limiting their participation in international data-sharing initiatives and slowing the development of interoperable health infrastructure [9]. Addressing these disparities is vital for equitable access to medical AI, especially in specialties such as cardiology, where precise diagnosis and tracking of comorbid conditions are critical [10]. Moreover, with the increasing digitization of health records in Greece and other underrepresented countries, there is a timely need for language- and domain-adapted benchmarks that can catalyze the development of accurate, explainable, and generalizable NLP models.

To address this gap, we introduce ELCardioCC, a shared task and competitive benchmark that is part of the BioASQ 2025 challenge [11]. The task focuses on two main objectives: (i) assigning cardiology-related ICD-10 codes to discharge letters from Greek hospitals, and (ii) extracting specific mentions of ICD-10 codes from the discharge letters. Designed as both a research task and a competition, ELCardioCC invites participants to develop and evaluate automated systems on these objectives. To structure the challenge, the task is divided into three sub-tasks. In the Named Entity Recognition (NER) sub-task, participants identify clinical entity mentions in the text. In the Entity Linking (EL) sub-task, those mentions must be linked to the appropriate ICD-10 codes. The third sub-task, Multi-label Learning & Explainable AI (MLC-X), addresses the same goals as NER and EL, but encourages the use of models and methodologies from multi-label learning and explainable AI to enhance both performance and interpretability.

The remainder of this paper is organized as follows: Section 2 provides an overview of the ELCardioCC

task, including its objectives, sub-tasks, evaluation metrics, and baseline systems. Section 3 describes the dataset and presents key statistics. Section 4 outlines the methods, models, and algorithms submitted by participating teams. Section 5 reports the results and performance analysis, while Section 6 presents the related work. Finally, Section 7 concludes the paper and discusses potential directions for future work.

2. Overview of the Shared Task

In this section, we provide a detailed description of the ELCardioCC shared task. We begin with an overview of the task objectives, the dataset, the submission and evaluation process. Next, we define the individual sub-tasks. We then outline the evaluation framework and, finally, present our baselines for each sub-task.

2.1. Description

Participants in the ELCardioCC task were tasked with developing NER, EL and MLC-X systems using a specialized corpus of discharge letters from the cardiac department of a Greek hospital. These discharge letters, which were written in Greek, contained valuable medical information about patients' conditions, treatments, and outcomes. The corpus was annotated with the positions of mentions and their corresponding ICD-10 codes.

The provided datasets were divided into a development dataset and an unseen test dataset. The development dataset consists of letters that came with gold-standard annotations, indicating the positions of mentions in the text and mapping them to their respective ICD-10 codes. In contrast, the test dataset only contained the raw discharge letters without any annotations, leaving participants to apply their models to identify mentions and link them to the appropriate ICD-10 codes.

For submission, participants were required to prepare and submit JSON files containing their results for both tasks. The NER submissions included a list of start and end positions for the mentions identified within the text, while the EL submissions consisted of the identified mentions along with the ICD-10 codes they were linked to. For MLC-X submissions, each letter was accompanied by a list of corresponding ICD-10 codes, as well as a list of start and end positions for terms identified by an explainable AI model as contributing most significantly to the determination of these ICD-10 codes. The Explainable AI sub-task was optional. Each team was allowed to submit up to five different runs for each task, enabling them to explore different strategies or model configurations for better performance.

The evaluation process compared the predictions made by the participating teams against manual annotations provided by clinical experts. Details of the evaluation process and metrics can be found in Section 2.3.

Figure 1 provides an overview of the workflow for the ELCardioCC shared task.

2.2. Sub-Tasks

The task was divided into three main sub-tasks. In the first sub-task, participants focused on NER, where they were required to identify all mentions present in discharge letters, along with their corresponding start and end positions. This phase involved detecting 5 types of mentions: chief complaint, diagnosis, prior medical history, drugs and cardiac echo, which are often complex and context-dependent, especially when working with medical language in Greek.

The second sub-task, EL, involved mapping the recognized mentions to their corresponding ICD-10 codes. Once the mentions were detected either in the first phase or by using any other approach for mention identification, participants had to generate a list of candidate ICD-10 codes for each entity. This process required a deep understanding of both the clinical context and the ICD-10 classification system. The generated ICD-10 codes were expected to accurately represent the identified mentions and provide a standardized classification for the mentions.

In the MLC-X sub-task, participants' systems were tasked with identifying all relevant ICD-10 codes contained within each discharge letter. Unlike the EL sub-task, where participants first needed to

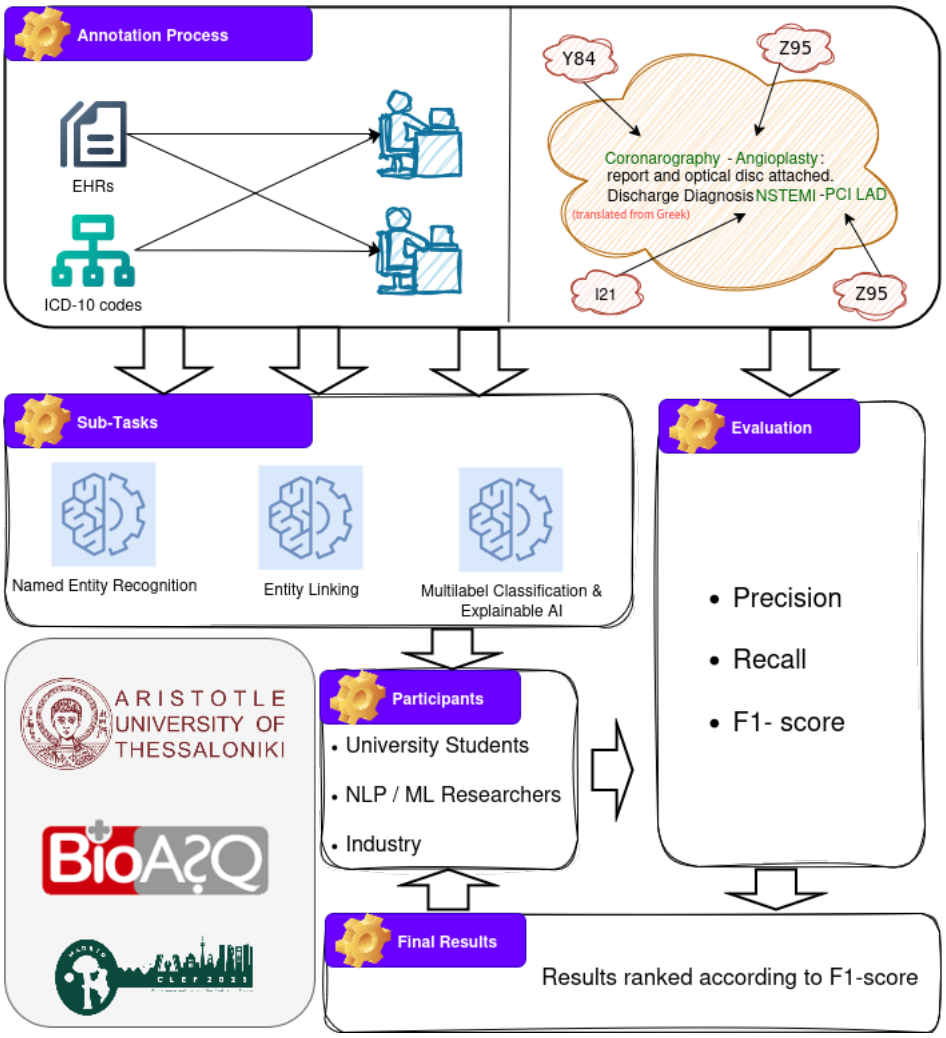


Figure 1: Visual Illustration of the ELCardioCC Shared Task Workflow

recognize specific mentions within the text and subsequently map them to their corresponding ICD-10 codes, the MLC-X sub-task allowed for a more direct approach. Participants were not required to explicitly identify mentions in the text; instead, they could employ alternative techniques. This distinction enabled participants to focus on optimizing classification strategies tailored to the complex, multi-label nature of the problem, emphasizing precision and recall in capturing all applicable codes from the unstructured medical text. In a subsequent optional step, the systems identify the mentions corresponding to the ICD-10 codes using explainable AI techniques. The terms highlighted by these techniques can be regarded as contributing significantly to the identification of the ICD-10 codes.

2.3. Evaluation

During the evaluation, participants' submissions were compared against the ground-truth annotations extracted by clinical experts, which were not provided to the participants. A strict exact match criterion, tailored to each task, was applied to compare ground truth and predicted outputs without any further preprocessing.

Specifically, for the NER sub-task, participants submitted lists of the start and end positions for mentions found in the discharge letters. A prediction was considered correct only if both the start and end positions exactly matched the gold standard annotations. In the EL sub-task, participants were required to both identify text spans and assign ICD-10 codes. Similarly, a prediction was considered correct only if both the span boundaries and the assigned code exactly matched the gold annotations,

with no partial matches accepted.

The evaluation was based on these exact matches and was measured, in both tasks, using Precision, Recall, and F1 Score, defined as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The first part of the MLC-X task did not involve predicting text spans but focused on identifying all relevant ICD-10 codes per document. For evaluation, the predicted and gold codes were treated as sets and compared for each document to determine true positives (correctly predicted codes), false positives (codes predicted but not in the gold data) and false negatives (gold codes not predicted). The precision, recall and F1 score were then calculated based on the aggregated counts across all documents as follows:

$$\text{Precision} = \frac{\sum_{i=1}^{N} |Y_i \cap \hat{Y}_i|}{\sum_{i=1}^{N} |\hat{Y}_i|}$$

$$\text{Recall} = \frac{\sum_{i=1}^{N} |Y_i \cap \hat{Y}_i|}{\sum_{i=1}^{N} |Y_i|}$$

where N is the number of documents, $\hat{Y}_i$ is the set of predicted codes for document i, and Y_i is the corresponding gold set. The F1 score was calculated as before.

For teams providing explanations, the evaluation followed the same criterion as the Entity Linking task and was computed independently as an additional metric. If a team did not provide span-level explanations, no valid span data was available, and their metrics were recorded as zero. The document-level sub-task was designated as the primary basis for team ranking, while the optional span-level sub-task provided an additional challenge for participants and was excluded from the final rankings.

2.4. Baseline Systems

For each of the sub-tasks, we developed one or more baseline systems to provide a clear reference point for evaluating the performance of current approaches on our benchmark dataset. These baselines serve as essential comparisons, helping participants and researchers understand the strengths and limitations of new methods relative to established techniques. By offering well-defined reference results, we aim to support fair, transparent, and reproducible assessment within the challenge.

For the NER sub-task, the baseline system [12] is built on the cased multilingual BERT-base[2] (mBERT) architecture [13]. The model was fine-tuned with a token-level classification head to perform NER under the BIO2 tagging scheme.

Documents were first segmented into sections using a keyword-driven method that exploits the semi-structured format and consistent headers of the discharge letters. Within each section, sentence segmentation was performed using Stanza, chosen for its support of Greek and its robustness in handling clinical abbreviations and irregular formatting. Each sentence was tokenized using the model's native tokenizer, with sequences padded or truncated to a maximum length of 384 tokens. Only a few samples exceeded this length, making it a reasonable cutoff to reduce computational cost.

Training was performed using the AdamW optimizer and a standard cross-entropy loss function for token-level classification. The model was fine-tuned over five epochs with a batch size of 4 and a learning rate of 8×10^{-6}, selected for stable convergence in low-resource settings. A weight decay

[2] https://huggingface.co/google-bert/bert-base-multilingual-cased

of 0.01 was applied to mitigate overfitting, alongside gradient clipping with a maximum norm of 3. Dropout with a probability of 0.1, as set by the default BERT configuration, was used. No learning rate scheduler or advanced regularization techniques were employed, preserving the baseline's simplicity.

Subword-level predictions produced by mBERT were aggregated to the word level by selecting the highest-scoring tag across each word's constituent subword tokens without any post-processing.

For the EL task, the baseline system [14] is also built on the multilingual BERT-base architecture and augmented to reflect the hierarchical structure of the ICD-10 taxonomy. It adopts a hierarchical classification framework with two parallel classification heads, one for coarse-grained (block-level) predictions and another for fine-grained (code-level) predictions.

Mentions are tokenized using a custom scheme that includes five tokens on both sides of the mention span. To maintain input consistency, sequences are padded or truncated to a fixed length of 128 tokens. The mention-level inputs are processed by mBERT, with contextualized representations fed into both classification heads. Special mention markers and masks are employed to ensure the model attends appropriately to the mention span during encoding.

Training is guided by a hierarchical loss function that combines the cross-entropy losses from both parent and child classifiers. A tunable weighting parameter is used to balance the contributions of each level to the overall loss. This dual-level supervision encourages the model to learn both broad and specific label associations, improving performance on fine-grained classes while maintaining coherence with the ICD-10 taxonomy.

The model is optimized using the Adam optimizer, with learning rates tuned independently for each classification head. In particular, a learning rate of 1×10^{-4} is used for the parent classifier and 1×10^{-5} for the child classifier, while the base model uses a lower learning rate of 3×10^{-6}. A moderate weight decay of 0.05 is applied for regularization, alongside dropout. Although only fine-grained (child-level) predictions are used for evaluation, the model is designed to support hierarchical supervision. In practice, training was conducted with a strong emphasis on the child-level loss, and the parent loss was down-weighted in later stages to prioritize fine-grained classification.

Finally, for the MLC-X task, the baseline approach [15] utilized the Greek-BERT transformer[3] [16], adapted and fine-tuned for multi-label classification. The model was fit and trained with 40 output heads, each representing one of the 40 most frequent ICD-10 codes in the training set. This translates to codes that have at least 30 appearances in the training set. Expanding the output layer to cover rarer codes led to similar or worse performance. Through this model architecture, the resulting system is limited to predicting only these common ICD-10 codes within discharge letters, while rarer codes or codes with no appearances in the training set cannot be detected.

The "MLCX1_baseline" system uses the transformer model to perform document-level prediction of ICD-10 codes for each discharge letter in the test set. Due to the length of the discharge letters, each document was first segmented into sections, with predictions then aggregated to document-level. Using a Greek pre-trained language model capable of predicting the most frequent ICD-10 codes provides a solid baseline for evaluation using micro-averaged metrics. However, more advanced systems capable of understanding and predicting a broader range of codes are expected to outperform it.

The "MLCX2_baseline" system extends the first system by adding a rule-based component after the prediction, in order to transform document-level predictions into span-level mentions. For each predicted code within a letter, the system searches the document for common predefined terms and abbreviations associated with that ICD-10 code. If relevant mentions are found, they are included as explanations for the prediction. If no relevant mention is detected, the corresponding code is removed from the output. As is typical of rule-based methods in text processing, the "MLCX2_baseline" system is expected to achieve high precision but lower recall, making it a strong baseline for evaluating explainable AI techniques.

[3] https://huggingface.co/nlpaueb/bert-base-greek-uncased-v1

3. Dataset

We introduce a novel dataset derived from Greek EHRs, specifically discharge letters from the cardiology department of a Greek hospital. The dataset addresses the limited availability of Greek-language clinical corpora, providing a valuable resource for advancing NLP tasks in biomedical applications. It is designed to support NER, EL, and clinical coding, with a focus on mapping medical entities to ICD-10 codes.

The dataset was curated from discharge letters, which capture critical aspects of patient care, including diagnoses, symptoms, medical procedures, and treatment plans. These documents were chosen for their detailed narrative structure, which represents a rich source of unstructured medical data. The dataset reflects the complexities of Greek clinical texts, featuring diverse sentence structures, specialized terminology, and abbreviations typical of healthcare documentation. It is designed to support the development and evaluation of multilingual and monolingual transformer-based models for tasks such as weakly supervised clinical entity recognition and automated clinical coding. The inclusion of commonly observed ICD-10 codes ensures relevance to practical healthcare applications, while the use of Greek-language texts fills a critical gap in existing resources.

A systematic and ethical approach was taken in constructing the dataset. First, all documents underwent de-identification to remove personally identifiable information, ensuring compliance with data protection regulations. Subsequently, a team of 4 medical professionals divided in 2 groups annotated the texts with ICD-10 codes with each group annotating the same set of documents, linking clinical entities to standardized classifications. An annotation tool, Doccano [17], was employed to streamline the process, enabling precise identification of mentions and their relationships. The annotation process included all cardiology-related information from the discharge letter, encompassing details from both the current hospitalization and the patient's past medical history. In particular, every disease and symptom, either current or past, was codified using a precise ICD-10 code. In addition, all medical procedures, such as cardiac catheterization or pacemaker insertion, that occurred during hospitalization were linked to their specific ICD-10 code. Medical recommendations and scheduled procedures that did not take place during the hospitalization were excluded from annotation. Moreover, the findings of diagnostic tests, such as echocardiographic findings, ECGs and x-rays, were annotated in detail whenever applicable. ECG findings were coded with the greatest possible accuracy at admission and discharge, as well as during hospitalization. X-ray reports were precisely annotated despite the fact that they were often missing or not relevant to cardiology in general. Of particular interest was the echocardiographic findings annotation process due to linguistic difficulties. Specifically, annotation of valvular heart disease was difficult due to the distance between the name of the valve and the pathological terms, such as stenosis or regurgitation, within the text. To address this issue, it was decided to annotate the full length of the phrase containing the valve name (first word to be annotated) and the pathological term (last word to be annotated). Furthermore, the annotation process was unable to characterize the severity of valvular heart disease, even though such data were available, due to the lack of specific ICD-10 codes.

The gold standard annotations for each discharge letter were created by merging the annotations provided by the two medical professionals, through the following conflict-resolution process:

- Spans annotated by only one annotator were included in the gold set, by following the assumption that the span simply being missed by the second annotator is much more likely than an irrelevant span being included. Excluding these annotations would thus lead to additional False Positive predictions by the participants.
- Spans annotated identically by both annotators were directly included. In cases of overlapping spans, the annotations were merged into a single span, by selecting the largest span. This approach ensures that additional context is preserved, which can be critical for accurate ICD-10 code assignment. Examples where the larger span provides a more accurate description of the diagnosis include: "coronary heart disease" versus "positive family history of coronary heart disease", "aortic valve stenosis" versus "severe aortic valve stenosis", "stroke" versus "ischemic stroke", and "STEMI" versus "STEMI of inferior wall".

- In instances where the same span was annotated using two different ICD codes, a third medical professional was consulted to determine the most appropriate ICD-10 label, after inspecting the relevant span and context.

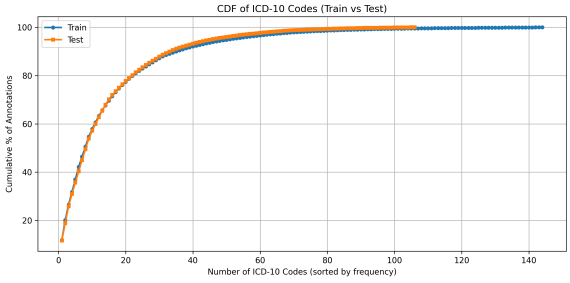
Additionally, to further improve the quality of the gold annotations, the test set discharge letters underwent a second annotation round by the medical professionals.

The final dataset that resulted from the above annotation process was divided into two sets: a training set containing 1,000 annotated discharge letters made available to participants to develop their systems, and a test set containing 500 annotated discharge letters, with the annotations kept hidden from the participants.

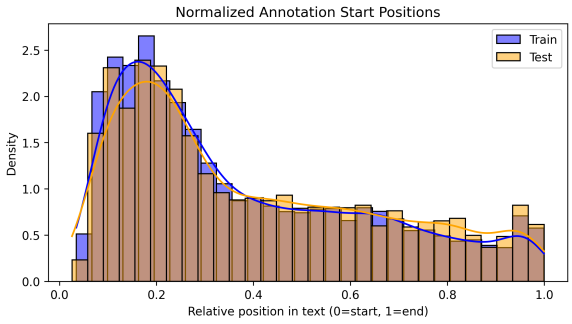
3.1. Key Statistics

The dataset was constructed with 1,000 documents for training and 500 for testing, creating a balanced split suitable for model development and evaluation. The training set contains 10,168 annotations, averaging 10.17 per document (ranging from 1 to 33), and the test set contains 5,696 annotations with a slightly higher average of 11.39 per document (ranging from 2 to 44). Despite these totals, the number of unique mentions is considerably lower, with 2,418 in the training set and 1,320 in the test set, indicating frequent repetition of the same mentions both within and across documents.

This pattern is more clearly seen in the code distribution. The top 10 most frequent codes account for a substantial portion of annotations in both sets (Figure 2(a)). Table 1 presents a detailed breakdown of these codes and their exact counts. In the training set, they represent 5,885 annotations, which is over half of the total. In the test set, the top 10 codes account for 3,255 out of 5,696. Such concentration reflects the dominance of a small subset of clinical concepts throughout the dataset. Additionally, the normalized start positions of annotations reveal clear clustering, with annotations in the first 30% of the documents occurring at roughly twice the frequency compared to the rest (Figure 2(b)). Such clustering shows clinical mentions mainly appear in the early sections, while later parts mostly contain discharge instructions, scheduling, and raw test results.



(a) Cumulative distribution of ICD-10 codes in the train and test sets



(b) Distribution of normalized annotation start positions across the training and test sets

Figure 2: Key Dataset Statistics

Beyond their location in the text, the annotations are also generally short. The average length is 14.31 characters in the training set and 13.48 characters in the test set. These lengths suggest that annotations typically consist of brief clinical mentions of about two to three words. Overall, about 7% of the text in both training and test documents consists of annotated clinical mentions, which is expected given the nature of clinical data and the typical annotation practices in this field. The annotations are based on a predefined labelset of 324 unique ICD-10 codes, deemed sufficient to describe the clinical concepts in the dataset.

From this set, 144 codes appear in the training data and 106 in the test data. Among these, 95 codes are shared between both sets, while 49 are exclusive to training and 11 unique to testing. This distribution

Table 1
Top 10 most frequently annotated ICD-10 codes along with their respective counts in the Train and Test datasets

ICD-10 Code	Train Count	Test Count
Z95	1214	666
I48	835	412
I25	658	397
R07	529	268
I21	526	255
I10	525	269
I50	437	249
R06	421	263
Y84	412	289
E11	328	143
I34	280	187

indicates a substantial overlap, ensuring that the test set is representative, while also containing some unique codes to assess model generalization.

4. Methodologies

This section outlines the methods and models used by participating systems in the ELCardioCC task. For each team, we summarize their approaches to the sub-tasks they participated in. Further details can be found in their respective papers.

The droidlyx team from Fudan University [18] used a BERT-based sequence labeling model (bert-base-greek-uncased-v1) for NER, feeding token embeddings into a two-layer MLP classifier for BIO tagging. They fine-tuned the model and used a sliding window for inference. For EL, they translated text to English with LibreTranslate, applied SapBERT using the [CLS] token as the entity representation, and classified entities via a two-layer MLP into ICD-10 codes, with additional data enrichment. In MLC-X Phase A, they reused EL predictions for each letter.

The bhuang team from the University of Padova [19] used multilingual LLMs (Gemma-3, Phi-4, Gemini) with zero-shot prompting to extract clinical mentions from Greek discharge letters, translating them into English with descriptions. A BERT bi-encoder filtered irrelevant mentions. For EL, a two-stage retrieval approach was used: BM25 to narrow candidates, followed by a MedCPT cross-encoder to select the best ICD-10 code. The MLC-X task was addressed by aggregating codes from EL. Various ensemble strategies combined outputs from different prompts and processing methods to improve recall and capture long-tail/nested entities.

The enigma team from Sofia University and Graphwise [20] approached NER using fine-tuned BERT-based models (Greek BERT and XLM-RoBERTa) with BIO tagging. For EL, they first used a dictionary-based method, then a bi-encoder model (BGE-M3 variants) for semantic matching between mentions and ICD-10 codes, fine-tuned with ranking loss. A cross-encoder reranker was tested but not used. In MLC-X, they applied a simple multi-label classification using BGE-M3 to process full documents.

Finally, the pjmathematician team from Netaji Subhas University of Technology [21] used Qwen-based LLMs for all subtasks. For NER, they applied both base and LoRA fine-tuned Qwen models with prompts to translate Greek text and extract entities. For EL, they linked entities to ICD-10 codes using semantic similarity via a multilingual sentence transformer. In MLC-X, they used Qwen-72B to predict ICD-10 codes and, in one configuration, also extracted supporting Greek mentions. Inference was performed using LMDeploy across all tasks.

5. Results

We present the results of the participating systems alongside our baseline models across for NER (Table 2), EL (Table 3) and MLC-X (Table 4) sub-tasks. Notably, the droidlyx team achieved the highest F1 scores in each sub-task, indicating strong overall performance. Interestingly, although the enigma team also leveraged Greek variants of BERT, the differences in model, pre-processing, and sub-task implementation choices seem to have led to the superior performance by the droidlyx team. Another factor likely contributing to the strong performance of the droidlyx team is their fine-tuning strategy. By carefully fine-tuning the BERT model for the specific sub-tasks, they likely enhanced both adaptability and task-specific understanding, leading to their top-tier results.

A particularly surprising observation is the performance of the bhuang team, which utilized LLMs. Despite their potential, these models underperformed in the NER sub-task. However, in the MLC-X sub-task (Phase A), bhuang achieved the highest recall score (0.8576), suggesting that their system was effective at capturing relevant labels, albeit at the cost of precision.

The pjmathematician team consistently showed the lowest performance across all tasks, indicating potential issues in either their modeling approach or implementation. In contrast, our baseline models demonstrated competitive results, particularly in the MLC-X sub-task, where our system achieved the highest precision score (0.9531). This underscores the robustness of our baseline configuration, providing a solid benchmark for participants to attempt to surpass.

Table 2
Results of the NER sub-task for the participating teams along with our baseline system. The table is ranked based on F1 Score, and values in bold indicate the best performance in Recall and Precision.

Team Name	System	Recall	Precision	F1 Score
droidlyx	system1	0.7059	**0.7618**	0.7328
ELCardioCC_baseline	mbert_baseline	0.6959	0.7460	0.7201
enigma	greek-bert-exact-bge-m3	0.7012	0.7328	0.7167
enigma	greek-bert-statistical-bge-m3	0.7012	0.7328	0.7167
enigma	greek-bert-statistical-sap	0.7012	0.7328	0.7167
enigma	xlmr-exact-bge-m3	**0.7079**	0.7222	0.715
enigma	xlmr-statistical-bge-m3	**0.7079**	0.7222	0.715
bhuang	5nm	0.6448	0.5205	0.5761
bhuang	1nm	0.4914	0.5331	0.5114
bhuang	2nm	0.6338	0.4000	0.4905
bhuang	3nm	0.5460	0.4387	0.4865
bhuang	4nm	0.6575	0.3601	0.4653
pjmathematician	config1	0.2484	0.2586	0.2534
pjmathematician	config2	0.2892	0.2188	0.2491
pjmathematician	config3	0.3297	0.1732	0.2271

6. Related Work

The Conference and Labs of the Evaluation Forum (CLEF) eHealth Lab in 2020 [22] introduced an Information Extraction (IE) task focused on automatic clinical coding. This task aimed to assign ICD-10 diagnosis and procedure codes to Spanish clinical case documents, along with identifying relevant evidence text snippets supporting the coded information.

In a related effort, the 2023 MedProcNER Task [23] introduced three sub-tasks centered on clinical procedures in Spanish texts. The first was Clinical Procedure Recognition, a named entity recognition (NER) task for identifying mentions of clinical procedures. The second, Clinical Procedure Normalization, required mapping these mentions to SNOMED CT codes through entity linking (EL). The third sub-task, Clinical Procedure-based Document Indexing, involved assigning SNOMED CT codes directly to full clinical reports for semantic indexing, independently of the other sub-tasks.

Table 3
Results of the EL sub-task for the participating teams along with our baseline system. The table is ranked based on F1 Score, and values in bold indicate the best performance in Recall and Precision.

Team Name	System	Recall	Precision	F1 Score
droidlyx	system1	0.6529	**0.7046**	0.6778
ELCardioCC_baseline	EL_baseline	0.6476	0.6942	0.6701
enigma	greek-bert-exact-bge-m3	0.6548	0.6844	0.6693
enigma	greek-bert-statistical-sap	0.6548	0.6844	0.6693
enigma	greek-bert-statistical-bge-m3	0.654	0.6835	0.6684
enigma	xlmr-statistical-bge-m3	**0.6594**	0.6728	0.666
enigma	xlmr-exact-bge-m3	0.6585	0.6719	0.6651
bhuang	5nm	0.5927	0.4852	0.5336
bhuang	1nm	0.448	0.5	0.4726
bhuang	3nm	0.4963	0.4211	0.4556
bhuang	2nm	0.5734	0.3677	0.448
bhuang	4nm	0.5945	0.3374	0.4305
pjmathematician	config1	0.0616	0.0642	0.0629
pjmathematician	config2	0.0693	0.0525	0.0597
pjmathematician	config3	0.0212	0.0112	0.0146

Table 4
Results of the MLC-X sub-task for Phase A (Multi-Label Learning) and Phase B (Explainable AI) for the participating teams, including our baseline system. The table is ranked based on the F1 Score of Phase A. Values in bold indicate the best performance in Recall and Precision for Phase A, while underlined values highlight the best scores in Phase B.

Team Name	**System**	**SubTask 3a (MLC-X)**			**SubTask 3b (MLC-X)**		
		R	**P**	**F1**	**R**	**P**	**F1**
droidlyx	system1	0.8377	0.8569	0.8472	-	-	-
ELCardioCC_baseline	MLCX1_baseline	0.7422	0.9339	0.8271	-	-	-
bhuang	5nm	0.825	0.6947	0.7543	-	-	-
ELCardioCC_baseline	MLCX2_baseline	0.5864	**0.9531**	0.7261	0.4442	0.605	0.5122
bhuang	1nm	0.7676	0.6205	0.6863	-	-	-
bhuang	3nm	0.8237	0.5227	0.6395	-	-	-
bhuang	2nm	0.8355	0.5131	0.6358	-	-	-
bhuang	4nm	**0.8576**	0.4572	0.5965	-	-	-
pjmathematician	config5	0.2656	0.586	0.3655	-	-	-
pjmathematician	config4	0.2257	0.6056	0.3288	0.0932	0.2326	0.1331

Similarly, the 2024 MultiCardioNER challenge [24] focused on domain-specific clinical NER and coding, particularly in cardiology. It featured two sub-tracks: one for disease recognition in Spanish, and another for multilingual medication extraction across Spanish, English, and Italian. Participants utilized resources such as DisTEMIST, DrugTEMIST, and the CardioCCC corpus, which contains cardiology-specific annotations.

In addition to these shared tasks, earlier works have also emphasized multilingual ICD-10 coding challenges. The 2018 CLEF eHealth Multilingual Information Extraction Task [25] addressed ICD-10 coding of death certificates in French (11,932 records), Hungarian (21,176 records) and Italian (3,618 records). This task focused on mapping causes of death from medical narratives to ICD-10 codes, using datasets provided by French CépiDc, Hungarian KSH and Italian ISTAT. The CLEF eHealth Lab in 2019 [26] introduced another multilingual coding task, which targeted the multi-label classification of German non-technical summaries of animal experiments. Participants predicted ICD-10 codes for descriptions of benefits, harms, and pressures affecting animals in biomedical research projects, with data annotated using the German ICD-10 ontology.

Further independent studies have also expanded ICD-10 coding to non-English datasets. Reys et

al. [27] focused on Brazilian-Portuguese clinical notes, assigning diagnostic ICD-10 codes to 77,005 free-text hospital discharge summaries sourced from a Brazilian hospital. Sammani et al. [28] tackled multilabel ICD-10 coding for Dutch cardiology discharge letters, using a dataset of 10,637 records with domain-specific cardiology diagnoses. Another notable contribution is the MKE-Coder study [29], which addresses automatic ICD coding for Chinese electronic medical records. This study leverages a large-scale dataset of 87,797 records from multiple hospitals and emphasizes the tailored nature of Chinese clinical texts with shorter diagnostic descriptions and discharge summaries compared to English counterparts like MIMIC-III. Similarly, Tchouka et al. [30] investigated ICD-10 coding for 56,014 unstructured French clinical texts from the Nord Franche-Comté Hospital comprised from documents such as discharge letters, operating reports, and clinical notes with each record being associated with multiple ICD-10 codes.

In this context, the ELCardioCC task aligns with the general objectives of the above efforts but introduces two novel aspects: it is conducted in the Greek language, and incorporates both multi-label learning and explainable AI for automatic clinical coding.

7. Conclusions

This study presented ELCardioCC, a novel shared task developed under the BioASQ 2025 framework, targeting the automation of clinical coding in a low-resource language setting. The task focused on two primary objectives: (i) the assignment of ICD-10 codes to Greek cardiology discharge letters, and (ii) the extraction of specific mentions of ICD-10 codes from the discharge letters. With participation from five research teams, the challenge highlighted the effectiveness of transformer-based architectures and language-specific pretraining for both Named Entity Recognition (NER) and Entity Linking (EL).

Despite promising results, considerable performance variation across sub-tasks suggests that automated clinical coding remains far from a solved problem—particularly in linguistically constrained environments. The underrepresentation of submissions in the Multi-label Learning and Explainable AI sub-task further underscores the need to prioritize model interpretability alongside performance. This is especially crucial in clinical settings, where explainable decision pathways are essential for ensuring safety, trust, and regulatory compliance.

To advance the state of the field, future iterations of ELCardioCC should pursue several directions. First, expanding the dataset to include more diverse clinical specialties and larger volumes of annotated text will support deeper model generalization. Second, incorporating multilingual corpora and cross-lingual evaluation protocols may enable more robust transfer learning and better scalability to other under-resourced health systems. Third, the evaluation framework should evolve to include clinically meaningful error metrics, such as critical code omission and false attribution, which directly affect downstream decision-making. Finally, stronger emphasis should be placed on the development and benchmarking of explainable coding systems, including rationale extraction, clinician-facing justifications, and human-in-the-loop workflows.

ELCardioCC represents an initial yet important step toward equitable, interpretable, and language-inclusive clinical NLP. By enabling systematic evaluation and fostering open collaboration, it contributes a foundational initiative for advancing automated medical coding in real-world health information systems. We strongly encourage further research in this direction to advance transparency and interpretability in automatic clinical coding systems.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: (1) Grammar and spelling check, (2) Paraphrase and reword, and (3) Improve writing style.

References

[1] J. Sedlakova, P. Daniore, A. Horn Wintsch, et al., Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review, PLOS Digital Health 2 (2023) e0000347. URL: https://doi.org/10.1371/journal.pdig.0000347. doi:10.1371/journal.pdig.0000347.

[2] A. Roberts, D. Demner-Fushman, E. Tonkin, et al., A scoping review of automatic coding of clinical narratives with icd-10, Journal of the American Medical Informatics Association 28 (2021) 1000–1010. doi:10.1093/jamia/ocab028.

[3] H. Dong, M. Falis, W. Whiteley, B. Alex, J. Matterson, S. Ji, J. Chen, H. Wu, Automated clinical coding: what, why, and where we are?, NPJ digital medicine 5 (2022) 159.

[4] M. Turchioe, A. Volodarskiy, J. Pathak, D. N. Wright, J. E. Tcheng, D. Slotwiner, Systematic review of current natural language processing methods and applications in cardiology, Heart 108 (2022) 909–916. URL: https://doi.org/10.1136/heartjnl-2021-319769. doi:10.1136/heartjnl-2021-319769.

[5] K. He, et al., Autonomous international classification of diseases coding using pretrained language models and advanced prompt learning techniques, JMIR Medical Informatics 13 (2025) e63020. doi:10.2196/63020.

[6] A. Kang, et al., Medcoder: A generative ai assistant for medical coding, in: NAACL Industry Track, 2025.

[7] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum, Clinical natural language processing in languages other than english: opportunities and challenges, Journal of biomedical semantics 9 (2018) 1–13.

[8] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, V. Osmani, Natural language processing of clinical notes on chronic diseases: Systematic review, JMIR Medical Informatics 7 (2019) e12239. URL: https://medinform.jmir.org/2019/2/e12239. doi:10.2196/12239.

[9] A. Bracken, C. Reilly, A. Feeley, E. Sheehan, K. Merghani, I. Feeley, Artificial Intelligence (AI) - Powered Documentation Systems in Healthcare: A Systematic Review, Journal of Medical Systems 49 (2025) 28. doi:10.1007/s10916-025-02157-4.

[10] J. S. Boyle, A. Kascenas, P. Lok, M. Liakata, A. Q. O'Neil, Automated clinical coding using off-the-shelf large language models, arXiv preprint (2023).

[11] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. Maria Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[12] E. Stoikopoulou, Weakly Supervised NER for Cardiology Using Multilingual Transformers (2024).

[13] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual BERT?, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4996–5001. URL: https://aclanthology.org/P19-1493. doi:10.18653/v1/P19-1493.

[14] A. Kipouros, Investigating Entity Linking in Greek Electronic Health Records: Leveraging Hierarchical Structures and Bi-Encoder Architectures (2024).

[15] A. Toumpas, Transferring Labels from the Document Level to the Mention Level in Clinical Coding by Interpreting Transformer-based Classifiers (2024).

[16] J. Koutsikakis, I. Chalkidis, P. Malakasiotis, I. Androutsopoulos, Greek-bert: The greeks visiting sesame street, in: 11th Hellenic Conference on Artificial Intelligence, SETN 2020, ACM, 2020, p. 110–117. URL: http://dx.doi.org/10.1145/3411408.3411440. doi:10.1145/3411408.3411440.

[17] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, X. Liang, doccano: Text annotation tool for human, 2018. URL: https://github.com/doccano/doccano, software available from

https://github.com/doccano/doccano.
[18] Y. Liu, LYX_DMIIP_FDU at BioASQ 2025: Utilizing BERT embeddings for biomedical text mining, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[19] B. Huang, Clinical entity recognition and linking in greek discharge letters using multilingual-llm-based multi-stage system, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[20] B. Velichkov, A. Datseris, S. Vassileva, S. Boytcheva, Enigma @ ElCardioCC: Bridging NER and ICD-10 Entity Linking - A Hybrid Method for Greek Clinical Narratives, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[21] P. Vachharajani, Multilingual embedding and prompt-driven approaches for named entity recognition, entity linking, and clinical code prediction in greek discharge summaries, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.
[22] L. Goeuriot, H. Suominen, L. Kelly, A. Miranda-Escalada, M. Krallinger, Z. Liu, G. Pasi, G. Gonzalez Saez, M. Viviani, C. Xu, Overview of the clef ehealth evaluation lab 2020, in: International conference of the cross-language evaluation forum for European languages, Springer, 2020, pp. 255–271.
[23] S. Lima-López, E. Farré-Maduell, L. Gascó, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of medprocner task on medical procedure detection and entity linking at bioasq 2023., in: CLEF (Working Notes), 2023, pp. 1–18.
[24] S. Lima-López, E. Farré-Maduell, J. Rodríguez-Miret, M. Rodríguez-Ortega, L. Lilli, J. Lenkowicz, G. Ceroni, J. Kossoff, A. Shah, A. Nentidis, et al., Overview of multicardioner task at bioasq 2024 on medical speciality and language adaptation of clinical ner systems for spanish, english and italian, CLEF Working Notes (2024).
[25] A. Névéol, A. Robert, F. Grippo, C. Morgand, C. Orsi, L. Pelikan, L. Ramadier, G. Rey, P. Zweigenbaum, Clef ehealth 2018 multilingual information extraction task overview: Icd10 coding of death certificates in french, hungarian and italian., in: CLEF (Working Notes), CEUR-WS, 2018, pp. 1–18.
[26] M. Sänger, L. Weber, M. Kittner, U. Leser, Classifying german animal experiment summaries with multi-lingual bert at clef ehealth 2019 task 1., in: CLEF (Working Notes), 2019.
[27] A. D. Reys, D. Silva, D. Severo, S. Pedro, M. M. de Sousa e Sá, G. A. Salgado, Predicting multiple icd-10 codes from brazilian-portuguese clinical notes, in: Brazilian Conference on Intelligent Systems, Springer, 2020, pp. 566–580.
[28] A. Sammani, A. Bagheri, P. G. van der Heijden, A. S. Te Riele, A. F. Baas, C. Oosters, D. Oberski, F. W. Asselbergs, Automatic multilabel detection of icd10 codes in dutch cardiology discharge letters using neural networks, NPJ digital medicine 4 (2021) 37.
[29] X. You, X. Liu, X. Yang, Z. Wang, J. Wu, Mke-coder: Multi-axial knowledge with evidence verification in icd coding for chinese emrs, arXiv preprint arXiv:2502.14916 (2025).
[30] Y. Tchouka, J.-F. Couchot, D. Laiymani, P. Selles, A. Rahmani, Automatic icd-10 code association: A challenging task on french clinical texts, in: 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2023, pp. 91–96.