Overview of ELCardioCC Task on Clinical Coding in Cardiology at BioASQ 2026

BioASQ at CLEF 2026

Dimitris Dimitriadis[1,*], Vasiliki Patsiou[2], Eleonora Stoikopoulou[1,†], Achilleas Toumpas[1,†], Alexandra Bekiaridou[2], Athanasios Samaras[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

While automated clinical coding via standardized frameworks like ICD-10 has revolutionized electronic health record (EHR) analytics, the vast majority of progress remains confined to English-language narratives within general medicine, leaving cardiovascular and low-resource settings comparatively underexplored. To address this linguistic disparity in low-resource medical NLP, we present the second iteration of the ELCardioCC 2026 shared task, organized within the BioASQ challenge. Building upon the foundational insights of the first year 2025 task, where participants focused extensively on localized clinical text spans, the 2026 edition pivots to a macro-level objective, evaluating participating systems exclusively on document-level diagnostic code assignment for Greek cardiology hospital discharge letters. To support this expanded scope, we introduce a substantially larger dataset consisting of 2,500 de-identified cardiology discharge letters, marking a significant expansion over the previous year's corpus. The architectural design of the dataset provides multi-level complexity: a core subset of 1,453 documents features granular, dual-level annotations (both token-level mentions and document-wide diagnoses), while the remaining 1,047 documents provide document-level labels alone. This expanded framework challenges participants to leverage diverse methodologies, including Named Entity Recognition (NER), Entity Linking (EL), and multi-label classification, to map complex clinical narratives directly to standard global categories. Ultimately, ELCardioCC 2026 establishes a rigorous, scaled-up benchmark for evaluating modern clinical NLP architectures in a non-English setting, driving the development of generalizable, multilingual clinical technologies.

Keywords

Clinical coding, Document-level classification, Greek clinical NLP, BioASQ shared task, Named entity recognition, Entity linking, Benchmark dataset

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide and account for a substantial proportion of hospital admissions and healthcare utilization [1]. Management of patients with CVDs, who frequently present with multiple comorbidities, generates detailed discharge summaries describing diagnoses, procedures, medications, imaging findings, and other clinically relevant information. Although these narratives represent a rich source of clinical knowledge, they are predominantly recorded as unstructured free text, making their systematic analysis and secondary use challenging. Converting this information into standardized terminologies such as the *International Classification*

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*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); ampekiaridou@gmail.com (A. Bekiaridou); ath.samaras.as@gmail.com (A. Samaras); 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); 0000-0002-1143-2293 (A. Bekiaridou); 0000-0002-3404-2749 (A. Samaras); 0000-0001-7491-6319 (G. Giannakoulas); 0000-0002-7879-669X (G. Tsoumakas)

of Diseases, Tenth Revision (ICD-10) is essential for clinical documentation, hospital reimbursement, epidemiological surveillance, and cardiovascular research [2, 3]. However, accurately assigning ICD-10 codes from free-text discharge summaries remains a fundamental challenge for modern healthcare systems due to the complexity of clinical narratives and the variability inherent in manual coding practices [4].

Automated clinical coding has consequently emerged as a major research area within clinical natural language processing (NLP). Recent advances in deep learning, particularly transformer-based language models, have substantially improved the automatic assignment of ICD codes from clinical text [5, 6]. Nevertheless, automated clinical coding remains a challenging multi-label classification problem due to the complexity of medical terminology, lengthy clinical documents, highly imbalanced label distributions, and the hierarchical organization of ICD-10 [7]. Furthermore, existing progress is primarily driven by English-language corpora, which limits the applicability of these systems in multilingual and resource-constrained clinical settings, such as Greek healthcare.

To address these challenges, the *ELCardioCC* task was introduced as part of the BioASQ Challenge [8]. The task aims to advance research in automated clinical coding for Greek-language electronic health records, with a particular emphasis on the cardiovascular domain, where accurate coding is essential for clinical decision-making, epidemiological studies, and healthcare management.

This paper presents the second edition of the ELCardioCC challenge. While the previous edition featured two complementary subtasks—(i) assigning cardiology-related ICD-10 codes to discharge letters from Greek hospitals at the document level, and (ii) identifying and extracting text spans corresponding to ICD-10 concepts at the mention level, the current edition focuses exclusively on the document-level clinical coding task. By concentrating on this setting, participants are encouraged to develop robust models capable of inferring the complete set of relevant ICD-10 codes from the entire discharge letter, without relying on explicit mention annotations.

This focus on document-level coding reflects the setting in which clinical coding systems are most directly evaluated: assigning a complete set of relevant diagnosis codes from the full discharge letter. The task encourages participants to explore advanced approaches for long-document understanding, multi-label classification, and modeling of complex relationships among co-occurring diseases and clinical findings. By providing a standardized benchmark for Greek clinical coding, the second ELCardioCC challenge seeks to foster the development of accurate and generalizable methods, while facilitating fair and reproducible comparisons across competing systems.

The remainder of this paper is organized as follows: Section 2 provides an overview of the ELCardioCC task, including its objectives, the evaluation metrics, and baseline system. 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. We then outline the evaluation framework and, finally, present our baseline.

2.1. Description

Participants in the ELCardioCC 2026 shared task were tasked with developing automated clinical coding systems capable of assigning ICD-10 diagnosis codes to hospital discharge letters from the cardiology department of a Greek hospital. The discharge letters, written in Greek, contain detailed clinical information regarding patients' diagnoses, medical history, laboratory findings, treatments, and hospitalization outcomes. Each document is associated with one or more ICD-10 codes corresponding to the diagnoses assigned by expert clinical coders.

The provided data were divided into a development set and an unseen test set. The development set consisted of discharge letters accompanied by their gold-standard ICD-10 code annotations, enabling participants to train and validate their models. In contrast, the test set contained only the raw discharge letters, requiring participants to automatically predict the complete set of ICD-10 codes for each document without access to the reference annotations.

Participants were required to submit their predictions in JSON format, where each discharge letter was associated with the list of ICD-10 codes predicted by their system. Since a single discharge letter may correspond to multiple diagnoses, the task was formulated as a multi-label document classification problem. Teams were allowed to submit up to five runs, encouraging the exploration of different modeling strategies, architectures, and training configurations. System performance was evaluated by comparing the predicted ICD-10 codes against the gold-standard annotations of the unseen test set.

2.2. Evaluation

The task is formulated as a multi-label classification problem at the concept level (Figure 1). However, unlike standard multi-label settings, the mapping between concepts and ICD-10 codes is not strictly one-to-one. At the concept level, a single clinical concept may correspond to a set of valid alternative gold-standard codes, rather than a single label. Conversely, at the code level, the same ICD-10 code may appear in multiple gold code groups, meaning that a single predicted code can satisfy more than one concept. Given this structure, evaluation is performed at the concept level in order to better reflect clinically meaningful correctness, rather than exact code matching.

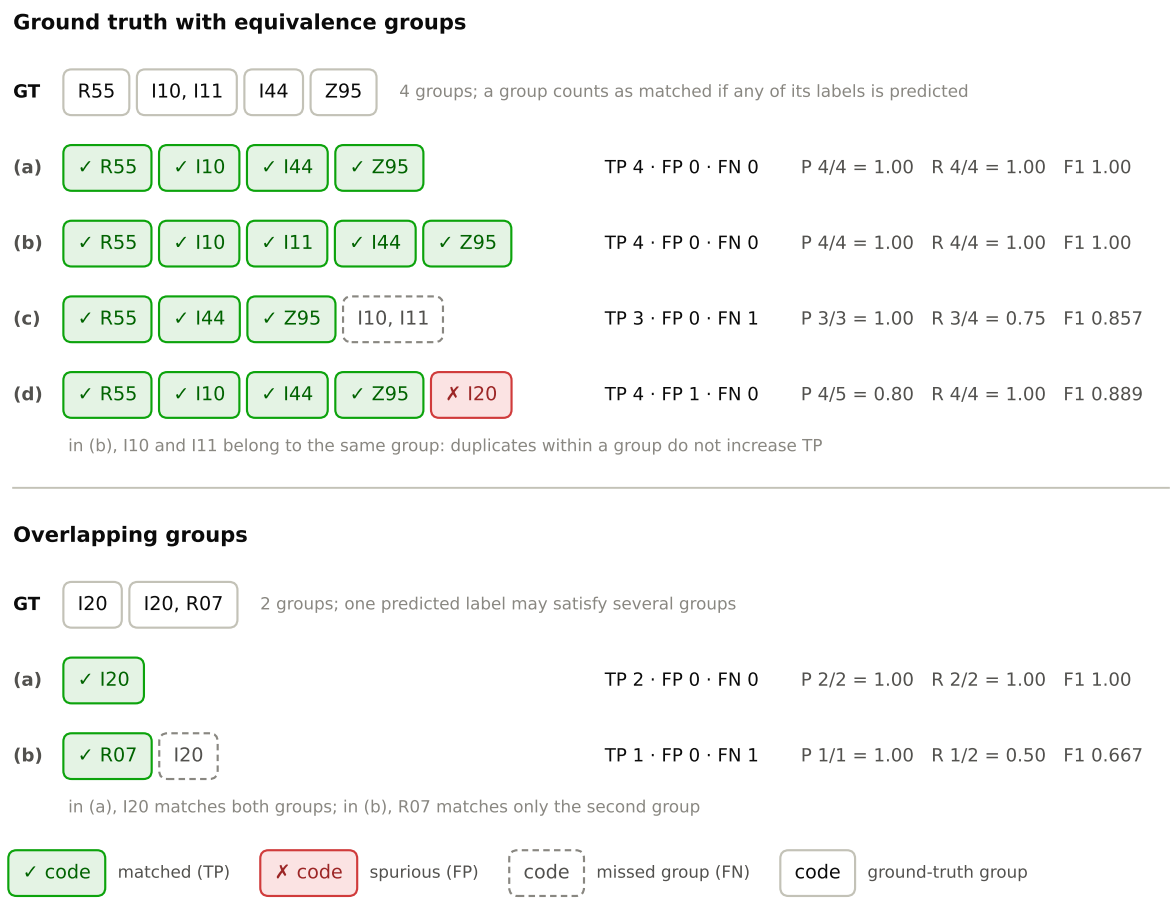


Figure 1: Worked examples of the group-level evaluation protocol. A ground-truth group is counted as a true positive if at least one of its labels is predicted; duplicate predictions within a group do not increase TP, and in overlapping groups a single predicted label may satisfy multiple groups.

A target concept is considered correctly identified if at least one predicted code belongs to its corresponding gold code group. No additional credit is assigned for predicting multiple codes within the same group; each gold group is therefore counted in a binary manner as either correctly matched or unmatched. Predicted codes that do not belong to any gold code group are treated as false positives.

Formally, let $\hat{Y}$ denote the set of predicted codes for a given clinical document, and let $G = \{G_1, \ldots, G_m\}$ denote the set of gold code groups, where each group $G_i = \{y_{i1}, \ldots, y_{ik_i}\}$ contains the set of acceptable ICD-10 codes corresponding to a single clinical concept, with $k_i \geq 1$.

A gold group G_i is considered correctly predicted if $\hat{Y} \cap G_i \neq \emptyset$. The number of true positives is then defined as:

$$TP = \sum_{i=1}^{m} \mathbb{1}_{\{\hat{Y} \cap G_i \neq \emptyset\}},$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, which takes the value 1 if the condition holds and 0 otherwise.

The number of false negatives is defined as:

$$FN = m - TP.$$

False positives correspond to predicted codes that do not belong to any gold code group:

$$FP = \left| \hat{Y} \setminus \bigcup_{i=1}^{m} G_i \right|.$$

Finally, precision, recall, and F1-score are computed as:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F1 = \frac{2PR}{P + R}.$$

2.3. Baseline

We frame Greek cardiology clinical coding as a multi-label document classification problem. Given a clinical document d and a fixed label set $\mathcal{L} = \{l_1, \ldots, l_K\}$ of K ICD codes, the model predicts a binary vector $\hat{y} \in \{0, 1\}^K$ indicating which codes apply to d. The gold annotations are provided at the document level as a collection of code *groups*, where each group enumerates a set of mutually acceptable (equivalent) codes for a single underlying clinical concept. During training we flatten this structure into a multi-hot target vector $y \in \{0, 1\}^K$ in which every code occurring in any group is marked as positive.

We build on GREEK-BERT [9] (`nlpaueb/bert-base-greek-uncased-v1`), a monolingual, uncased BERT-base encoder (12 Transformer layers, 768 hidden units, 12 attention heads, ≈110M parameters) pre-trained on the Greek portions of Wikipedia, the European Parliament Proceedings (Europarl), and OSCAR. We initialize the encoder from these pre-trained weights and add a randomly initialized linear classification head on top of the pooled `[CLS]` representation, producing one logit per label. The model is optimized for multi-label classification using a sigmoid activation with a binary cross-entropy (with logits) objective, summed over all K labels. Each document is tokenized with the GREEK-BERT WordPiece tokenizer and truncated or padded to a fixed length of 512 sub-word tokens, the maximum context window of the encoder.

We hold out a random 15% of the training documents as a validation split (85%/15%) and fix all random seeds to 42 for reproducibility. The model is fine-tuned with the AdamW optimizer using a peak learning rate of 2×10^{-5}, a linear warmup over the first 10% of optimization steps followed by linear decay, weight decay of 0.01, and a per-device batch size of 8. Training runs for up to 20 epochs with mixed-precision (FP16) computation. We evaluate on the validation split at the end of every epoch and apply early stopping with a patience of 3 epochs. The checkpoint achieving the highest validation F_1 is retained as the final model. At inference time, a label is predicted as positive when its sigmoid probability is at least 0.5.

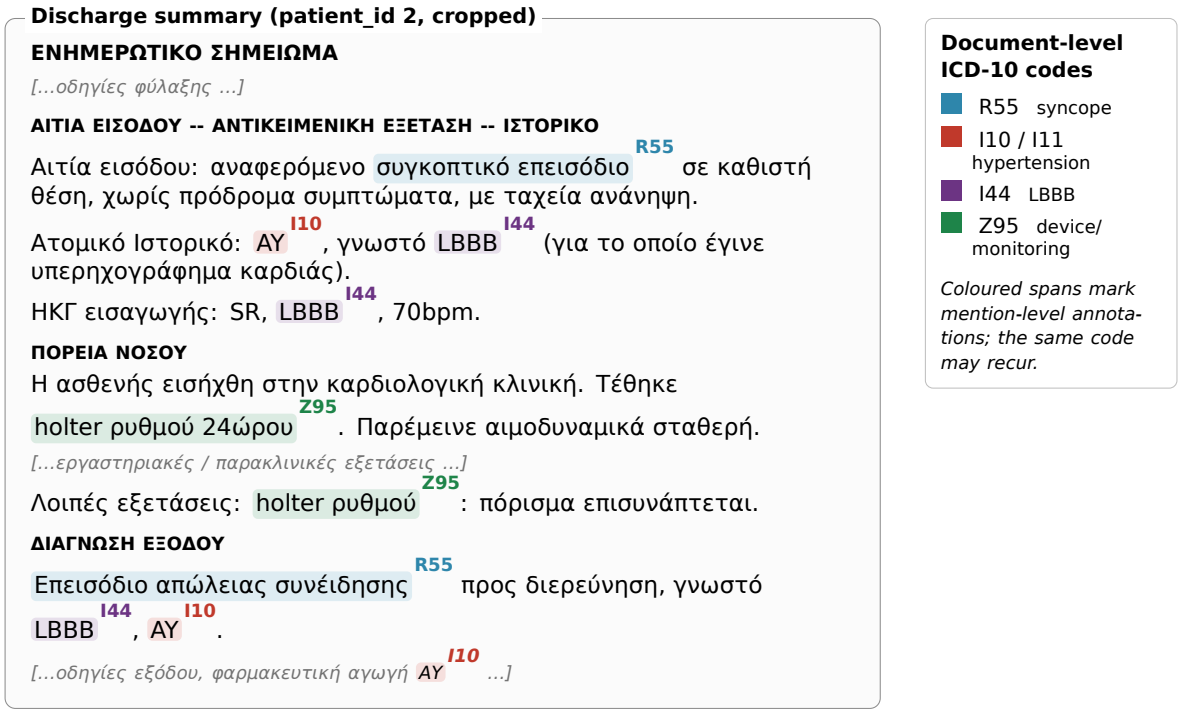


Figure 2: An annotated instance from the ELCardioCC corpus (patient 2, cropped). Coloured spans denote mention-level ICD-10 annotations; the same code recurs across the document, and the I10/I11 group illustrates multi-code assignment.

3. Dataset

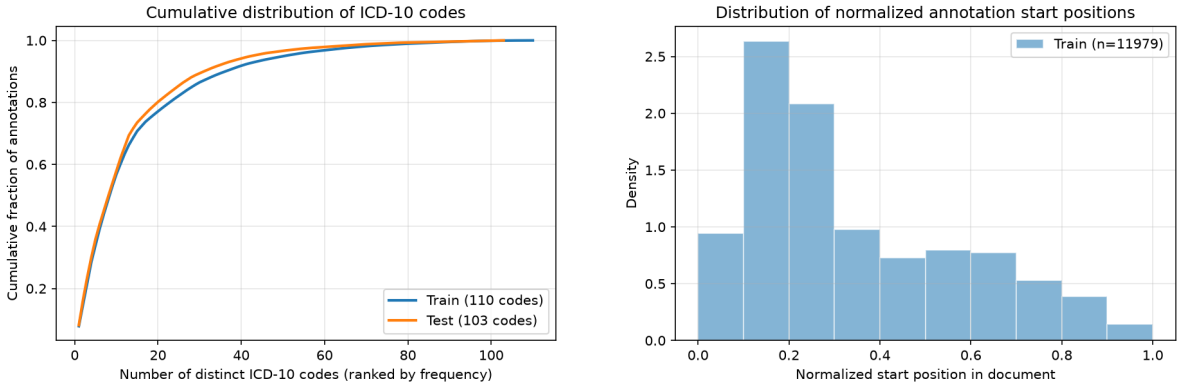
The ELCardioCC corpus consists of 3,000 de-identified Greek cardiology discharge summaries annotated with ICD-10 codes. An example of a cropped discharge letter for the second patient of the training set, along with its annotations, is presented in Figure 2. The documents are partitioned into a training set of 2,500 documents and a held-out test set of 500 documents, yielding an $83.3\%/16.7\%$ split. Training set annotations are provided at two granularities: at *document level*, where each clinical concept is recorded as a group of one or more candidate ICD-10 codes, and at *mention level*, where codes are additionally grounded to a text span in the document. Mention-level (span) annotations are available for 1,453 documents; the test set is associated with document-level codes only. Consequently, the positional and surface-form statistics reported below are computed on the training set, whereas code-frequency and label-coverage statistics are reported for both splits and for the corpus as a whole. Across the full dataset there are 17,105 annotation groups comprising 20,905 code occurrences, and every document carries at least one annotation.

At the document level, the corpus is densely annotated: a document is associated with 6.97 codes on average ($\sigma = 2.85$; median 7; range 1–20), distributed across 5.70 annotation groups per document on average. The training and test splits are closely matched in this respect (7.02 versus 6.70 codes per document), confirming that the partition is balanced and that the test set is representative of the training distribution. A non-negligible fraction of annotation groups (19.34% over the whole corpus) contain more than one candidate ICD-10 code, reflecting the inherent ambiguity of mapping free-text clinical mentions to a single normative code. At the mention level, the training set contains 11,979 grounded annotations, an average of 4.79 per document (range 0–27) across all training documents. The number of distinct surface mentions is considerably smaller (2,181 unique strings, 2,234 unique mention–code pairs), indicating that the same clinical expressions recur frequently both within and across documents.

Code distribution. The distribution of the labels is markedly skewed (Figure 3(a)). The ten most frequent codes account for 9,972 of 17,555 (56.80%) code occurrences in the training set and 1,940 of 3,350 (57.91%) in the test set; Table 1 reports their exact counts. This concentration is quantified by a Gini coefficient of 0.74 (train) and 0.75 (test) and a normalized Shannon entropy of 0.77 and 0.76, respectively, denoting that a small subset of cardiological concepts—most prominently prosthetic-device and graft status (Z95), hypertensive disease (I10/I11), chronic ischaemic heart disease (I25) and atrial fibrillation (I48)—dominates the corpus. The near-identical concentration profiles of the two splits (Figure 3(a)) further attest to the consistency of the partition.

Annotation position and length. Mentions are not uniformly distributed within documents (Figure 3(b)). When start offsets are normalised by document length, 56.62% of training annotations fall within the first 30% of the text, roughly three times the density observed over the remaining 70%. This reflects the structure of discharge summaries, in which diagnoses and clinical findings are stated early, while later sections contain discharge instructions, follow-up scheduling and raw laboratory results. Mentions are also short, averaging 13.92 characters (median 14; max 145) and 1.79 tokens, i.e. typically brief two- to three-word clinical expressions. On average, annotated mentions cover 3.29% of a document's characters, consistent with the sparse, targeted nature of clinical concept annotation.

Label set and split overlap. Annotations are based on a predefined set of labels of 115 ICD-10 categories, all of which are attested in the data. Of these, 110 codes occur in the training split and 103 in the test split. The two splits share 98 codes, while 12 are exclusive to training and 5 (I60, N02, N93, R04, R23) appear only in testing. This substantial overlap ensures that the test set is representative of the training distribution, while the small set of split-exclusive codes provides a controlled probe of model generalization to rare and unseen categories.



(a) Cumulative distribution of ICD-10 code occurrences in the train and test sets.

(b) Distribution of normalised annotation start positions in the training set.

Figure 3: Key statistics of the ELCardioCC dataset.

4. Methodologies

Several approaches have been proposed to address the ELCardioCC Shared Task. These approaches fundamentally diverge into two strategic frameworks: directly modeling the multi-label nature of document-level clinical coding, or decomposing the task by leveraging the token-level NER and EL annotations available in the training corpus.

A prominent strategy involves decomposing the task into a two-stage clinical pipeline. The **fernandogd97** team [10] implemented a pipeline consisting of mention detection followed by semantic linking. For the initial phase, they utilized a fine-tuned Greek-BERT model to identify clinical mentions. Their

Table 1
Top ten most frequently annotated ICD-10 codes and their occurrence counts in the train and test splits and the full corpus.

ICD-10 Code	Train	Test	Total
I25	1379	241	1620
I10	1207	275	1482
I11	1207	275	1482
Z95	1222	218	1440
R07	946	196	1142
I21	934	146	1080
I48	830	161	991
Y84	772	139	911
R06	763	139	902
E11	712	123	835

Table 2
Summary statistics of the ELCardioCC dataset by split and overall. Mention-level (span) statistics are available for the training set only.

Statistic	Train	Test	All
Documents	2,500	500	3,000
Annotation groups	14,347	2,758	17,105
Code occurrences	17,555	3,350	20,905
Multi-code groups (%)	19.18	20.16	19.34
Unique codes	110	103	115
Codes per document (mean)	7.02	6.70	6.97
Codes per document (std)	2.86	2.78	2.85
Mention-level annotations	11,979	—	11,979
Unique surface mentions	2,181	—	2,181
Mention length, chars (mean)	13.92	—	13.92
Mention length, tokens (mean)	1.79	—	1.79
Text annotated per doc. (%)	3.29	—	3.29
Gini coefficient	0.74	0.75	0.75
Normalised entropy	0.77	0.76	0.76

primary contribution focused on adapting the ClinLinker framework to the Greek language (Greek-ClinLinker). This adaptation was achieved by automatically translating English SNOMED CT concepts into Greek, incorporating ontological hierarchy information (parent and grandparent relations) via contrastive learning, and augmenting the training data with data from ELCardioCC 2025 and translated instances of the Spanish CodiEsp corpus.

Similarly, the **LSI_UNED** team [11] adopted a two-stage paradigm. In the first stage, they employ the W^2NER architecture to detect clinical entity mentions via a word-pair relation modeling approach. In the second stage, a transformer-based classifier assigns an ICD-10 code to each detected entity. Document-level labels are then derived from the set of entity-level predictions.

Conversely, other approaches treat the task as a direct semantic mapping or multi-label optimization problem. The **Georgios_1** team [12] developed MedicalModelWithDescriptions, an end-to-end architecture built upon an XLM-RoBERTa Large backbone. To explicitly align text with structured knowledge, they introduced a semantic label anchoring mechanism that projects document representations into a label-description space derived from Greek ICD-10 terminology. They mitigated extreme class imbalance through a knowledge-informed data augmentation strategy that combines frequency-aware sampling with stochastic synonym replacement using a curated medical lexicon. Optimization was robustly handled via Layer-wise Learning Rate Decay (LLRD) and Stochastic Weight Averaging (SWA), with final predictions refined through per-label dynamic thresholding and strict ontological hierarchy

enforcement.

Finally, the **stanimeros** team [13] proposed a highly diversified ensemble framework. Their architecture aggregates six heterogeneous components: a fine-tuned Greek BERT model, two variants of XLM-RoBERTa (Base and Large), a hybrid information retrieval module, a NER and EL integrated into the pipeline. The underlying neural networks were optimized using specialized multi-label loss functions—specifically Asymmetric Loss and ZLPR (Zero-Loss Positive Rate)—coupled with per-label threshold tuning. To synthesize the predictions, they employed a declarative metaheuristic ensemble framework that searches the decision space using random restarts, hill-climbing, and Variable Neighbourhood Search (VNS).

5. Results

Table 3 presents the official results of all submitted systems on the ELCardioCC Shared Task 2026 evaluation set, ranked according to the micro-averaged F1-score.

Team	Submission	Recall	Precision	F1
stanimeros	ensemble_metaheuristic_p4_winner_safer_thr_merge_and_weighted_correction	0.8510	0.8830	**0.8667**
Georgios_1	Baymax_submission_v1	0.8658	0.8618	0.8638
stanimeros	ensemble_metaheuristic_p4_winner_extended_merge_k2_weighted_per_label_correction	0.8687	0.8539	0.8613
stanimeros	ensemble_metaheuristic_p4_winner_safer_thr_merge_k2_weighted_per_label_correction	0.8767	0.8431	0.8596
stanimeros	mlc_greek_bert_100	0.8194	0.8811	0.8491
stanimeros	mlc_greek_bert	0.8310	0.8675	0.8489
LSI_UNED	test_set_NER_greekuncased_greekuncased	0.8187	0.8416	0.8300
LSI_UNED	test_set_NER_greekuncased_roberta	0.8176	0.8420	0.8297
LSI_UNED	test_set_NER_greekuncased_bertbase	0.8173	0.8417	0.8293
alikibliona	submission_A_reranker	0.8183	0.8246	0.8215
elcardiocc	baseline_bert_base_uncased_greek	0.6592	0.9036	0.7623
Pakakisd	max_micro_curriculum	0.8093	0.7200	0.7620
fernandogd97	Greek_ClinLinker-P_greek_bert_DA_codiesp	0.7915	0.7340	0.7617
fernandogd97	Greek_ClinLinker-GP_greek_bert_DA_codiesp	0.7864	0.7274	0.7557
fernandogd97	Greek_ClinLinker-P_greek_bert	0.7919	0.7142	0.7510
Pakakisd	submission_ensemble_weighted	0.8695	0.6547	0.7469
Pakakisd	max_macro_basic	0.8832	0.5573	0.6834
Pakakisd	ensembling_cascade	0.7937	0.5616	0.6578
fernandogd97	Greek_ClinLinker-P_greek_bert_DA_codiesp_enriched_all	0.9728	0.0570	0.1077
fernandogd97	Greek_ClinLinker-P_greek_bert_DA_codiesp_enriched_sm	0.9728	0.0570	0.1077

Table 3
Official results of the ELCardioCC Shared Task 2026 ranked by micro-averaged F1-score.

The competition was highly competitive, with the four best submissions achieving micro-F1 scores above 0.85. The best overall system was submitted by the **stanimeros** team, obtaining an F1-score of 0.8667 while maintaining a strong balance between precision (0.8830) and recall (0.8510). Their results demonstrate the effectiveness of combining multiple heterogeneous models through a metaheuristic ensemble, allowing complementary prediction patterns to be exploited more effectively than any individual component.

The **Georgios_1** system ranked second with an F1-score of 0.8638, only 0.29 percentage points below the winning submission. Unlike the winning approach, which relies on model diversity, this system adopts a single end-to-end architecture that leverages semantic ICD-10 label descriptions through a label-anchoring mechanism, together with knowledge-informed training strategies. Its nearly identical

precision and recall (0.8618 and 0.8658, respectively) indicate a well-balanced classifier capable of achieving competitive performance without requiring an ensemble of heterogeneous models.

Interestingly, the results also reveal the benefits of ensembling within the **stanimeros** submissions. While their individual Greek-BERT models achieved F1-scores around 0.849, the metaheuristic ensemble consistently improved performance by nearly two percentage points, highlighting the complementary nature of neural, retrieval-based, rule-based, and NER-EL components.

The two-stage NER+EL paradigm produced competitive but consistently lower performance. The three submissions from **LSI_UNED** achieved remarkably stable results around 0.83 F1 regardless of the language model used in the entity linking stage, suggesting that the overall performance is largely constrained by the quality of the initial entity recognition stage. Similarly, the **fernandogd97** systems obtained F1-scores between 0.75 and 0.76, demonstrating that adapting ClinLinker to Greek constitutes a viable solution, although still less effective than directly optimizing the multi-label document classification objective.

The official baseline achieved a relatively high precision (0.9036), the highest among all systems, but suffered from substantially lower recall (0.6592), resulting in an F1-score of 0.7623. This behavior indicates that the baseline adopts a conservative prediction strategy, correctly identifying only a subset of the relevant ICD-10 codes while missing many true labels.

The submissions from **Pakakisd** illustrate the inherent trade-off between recall and precision in multi-label clinical coding. Their highest-recall systems exceeded 0.88 recall but exhibited considerably lower precision, leading to lower overall F1 performance. Finally, the two enriched variants submitted by **fernandogd97** achieved extremely high recall (0.9728) at the expense of very low precision (0.0570), indicating severe over-prediction that substantially degraded the final F1-score.

Overall, the results suggest that direct optimization of the multi-label classification objective currently provides the strongest performance for ELCardioCC. In particular, architectures enriched with semantic knowledge and carefully designed optimization strategies, as well as heterogeneous ensemble methods, consistently outperform pipeline-based approaches that decompose the problem into sequential Named Entity Recognition and Entity Linking stages.

The comparison with the previous ELCardioCC Shared Task edition highlights a clear improvement in system performance and methodological sophistication. In the previous edition, the best MLC-X Phase A system achieved a micro-F1 score of 0.8472, while the winning system this year obtained 0.8667, improving the state of the art by 1.95 percentage points. Moreover, the precision-recall trade-off has become substantially better balanced: while previous high-performing systems often relied on conservative predictions with very high precision or aggressive strategies with high recall, this year's leading submissions achieved strong performance by jointly optimizing both aspects. The emergence of metaheuristic ensemble methods and semantic label-aware models represents a shift from the primarily encoder-based approaches observed in the previous edition, where systems such as droidlyx [14] and enigma [15] relied mainly on fine-tuned language models for NER, Entity Linking, and MLC-X.

6. Related Work

The Conference and Labs of the Evaluation Forum (CLEF) eHealth Lab has conducted annual challenges in clinical information retrieval, extraction, and management since 2013, aiming to advance research and innovation in the medical and biomedical domain. In its introductory year [16], participants were provided with English unannotated clinical reports gathered from the MIMIC II database [17], and tasked with automatically identifying the boundaries of disorder named entities in the text and mapping them, with SNOMED being selected as the knowledge base. Since then, the competition has involved tasks related to clinical coding in English, but large focus has been given to multilingual clinical datasets, in order to address linguistic disparity present in this field. For example, the CLEF eHealth Lab challenge in 2020 [18] introduced an Information Extraction (IE) task focused on automatic clinical coding in the Spanish language, with the 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 [19] introduced three sub-tasks centered on clinical procedures in Spanish texts. The first sub-task involved Named Entity Recognition (NER) in order to identify mentions of clinical procedures in the clinical text. The second sub-task, Clinical Procedure Normalization, required mapping these mentions to SNOMED CT codes through Entity Linking (EL). The third sub-task involved assigning SNOMED CT codes directly to full clinical reports for semantic indexing, independently of the other sub-tasks.

Similar recent efforts include the 2024 MultiCardioNER challenge [20], which focused on domain-specific clinical NER and coding, particularly in cardiology. This challenge 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 [21], DrugTEMIST, and the CardioCCC corpus, which contains cardiology-specific annotations.

In addition to these shared tasks, earlier iterations of the CLEF eHealth Lab challenges have also emphasized multilingual clinical coding, using the standardized ICD-10 framework as knowledge base. The 2018 CLEF eHealth Multilingual Information Extraction Task [22] 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 [23] included a multilingual coding task in a different setting, targeting 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.

Apart from the CLEF eHealth challenges, further independent studies have also expanded ICD-10 coding to non-English datasets. Reys et al. [24] 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. [25] 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 [26], 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. [27] 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 first version of the ELCardioCC shared task that took place in 2025 [8] followed the general direction and ambition of multilingual-focused related work and introduced two novel aspects: it involved a clinical dataset containing discharge letters in the Greek language, and incorporated both multi-label learning and explainable AI for automatic clinical coding. The current second iteration of the ElCardioCC shared task aims to build on these foundations and scale-up the evaluation of modern clinical NLP architectures in a non-English setting, by featuring a substantially larger corpus of cardiology hospital discharge letters in the Greek language.

7. Conclusion

This paper presented the second edition of the ELCardioCC Shared Task, which focused on automatic ICD-10 coding of Greek cardiology discharge summaries as a document-level multi-label classification problem. By concentrating on the end-to-end coding task, the challenge encouraged the development of systems capable of assigning complete sets of ICD-10 codes directly from clinical narratives, providing a realistic benchmark that closely resembles real-world clinical coding workflows.

The shared task attracted diverse methodological approaches, ranging from two-stage Named Entity Recognition and Entity Linking pipelines to end-to-end transformer-based multi-label classifiers and

heterogeneous ensemble methods. The results demonstrate that direct optimization of the document-level classification objective currently offers the best performance on this benchmark, with the top-performing systems effectively combining semantic representations, knowledge-informed training strategies, and advanced optimization techniques. At the same time, the competitive performance of pipeline-based approaches highlights the continued importance of structured entity extraction and ontology-aware modeling for clinical coding.

As a standardized benchmark for Greek clinical coding, ELCardioCC enables fair comparison between competing approaches and fosters the development of state-of-the-art models in low-resource clinical NLP. The release of the dataset, evaluation framework, and baseline system is expected to facilitate further advances in automated clinical coding for the Greek language.

Future editions of the shared task will explore more challenging settings, including larger and more diverse collections of clinical documents, additional medical specialties, and multilingual clinical coding scenarios. Furthermore, incorporating long-context language models, retrieval-augmented architectures, and explicit exploitation of the ICD-10 ontology represents promising directions for improving coding accuracy, particularly for rare diagnoses and highly imbalanced label distributions.

Declaration on Generative AI

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

References

[1] World Health Organization, Cardiovascular diseases (cvds), 2025. URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] A. E. W. Johnson, T. J. Pollard, L. Shen, et al., Mimic-iv, a freely accessible electronic health record dataset, Scientific Data 10 (2023) 1. doi:10.1038/s41597-022-01899-x.

[3] P. B. Jensen, L. J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nature Reviews Genetics 13 (2012) 395–405. doi:10.1038/nrg3208.

[4] M. Peng, C. Eastwood, A. Boxill, R. J. Jolley, L. Rutherford, K. Carlson, S. Dean, H. Quan, Coding reliability and agreement of international classification of disease, 10th revision (icd-10) codes in emergency department data, International Journal of Population Data Science 3 (2018) 445.

[5] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein, Explainable prediction of medical codes from clinical text, in: Proceedings of NAACL-HLT, 2018, pp. 1101–1111.

[6] F. Li, H. Yu, Icd coding from clinical text using multi-filter residual convolutional neural network, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 8180–8187.

[7] S. Ji, W. Sun, X. Li, H. Dong, A. Taalas, Y. Zhang, H. Wu, E. Pitkänen, P. Marttinen, A unified review of deep learning for automated medical coding, Patterns 3 (2022) 100477. doi:10.1016/j.patter.2021.100477.

[8] D. Dimitriadis, V. Patsiou, E. Stoikopoulou, A. Toumpas, A. Kipouros, D. Papadopoulos, A. Bekiaridou, K. Barmpagiannos, A. Vasilopoulou, A. Barmpagiannos, et al., Overview of ElCardioCC task on clinical coding in cardiology at BioASQ 2025, in: CLEF, 2025.

[9] J. Koutsikakis, I. Chalkidis, P. Malakasiotis, I. Androutsopoulos, GREEK-BERT: The greeks visiting sesame street, in: 11th Hellenic Conference on Artificial Intelligence (SETN 2020), Association for Computing Machinery, 2020, pp. 110–117. doi:10.1145/3411408.3411440.

[10] F. Gallego-Donoso, S. Lima-López, E. Farré-Maduell, L. Pratesi, M. Krallinger, Hierarchy-aware dense retrieval for greek cardiology clinical coding at elcardiocc 2026, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR-WS, 2026.

[11] A. Ramirez-Arrabe, A. Duque, J. Martinez-Romo, Lsi_uned at elcardiocc 2026: Leveraging clinical

entity extraction for icd-10 coding in greek cardiology reports, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR-WS, 2026.
[12] G. Chalkias, P. Amanatiadis, G. Tsavlidou, S. Malamas, Baymax at clef 2026: Leveraging xlm-roberta and semantic label embeddings for greek clinical coding, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR-WS, 2026.
[13] N. Strafiotis, P. Stanimeros, V. Katsara, G. Chalkias, G. Tsavlidou, S. Magalios, E. Nalmpanti, P. Stamatakis, A multi-component system for multi-label icd-10 classification of greek cardiology discharge summaries, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR-WS, 2026.
[14] 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.
[15] 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.
[16] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. Jones, et al., Overview of the share/clef ehealth evaluation lab 2013, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2013, pp. 212–231.
[17] J. Lee, D. J. Scott, M. Villarroel, G. D. Clifford, M. Saeed, R. G. Mark, Open-access mimic-ii database for intensive care research, in: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2011, pp. 8315–8318.
[18] 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.
[19] 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.
[20] 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).
[21] A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources., CLEF (Working Notes) 3180 (2022) 179–203.
[22] 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.
[23] 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.
[24] 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.
[25] 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.
[26] 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).
[27] 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.