Journal of Biomedical Informatics 139 (2023) 104323

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Explainable clinical coding with in-domain adapted transformers

Guillermo López-García [a,*], José M. Jerez [a,1], Nuria Ribelles [b,2], Emilio Alba [b,2], Francisco J. Veredas [a,1]

[a] Departamento de Lenguajes y Ciencias de la Computación & Research Institute of Multilingual Language Technologies, Universidad de Málaga, Málaga, Spain
[b] Unidad de Gestión Clínica Intercentros de Oncología, Instituto de Investigación Biomédica de Málaga (IBIMA), Hospitales Universitarios Regional y Virgen de la Victoria, Málaga, Spain

ARTICLE INFO

Keywords:
Clinical Coding
Explainable Artificial Intelligence
Transformers
Deep Learning
Natural Language Processing
Medical Entity Normalization

ABSTRACT

Background and Objective: Automatic clinical coding is a crucial task in the process of extracting relevant information from unstructured medical documents contained in Electronic Health Records (EHR). However, most of the existing computer-based methods for clinical coding act as "black boxes", without giving a detailed description of the reasons for the clinical-coding assignments, which greatly limits their applicability to real-world medical scenarios. The objective of this study is to use transformer-based models to effectively tackle explainable clinical-coding. In this way, we require the models to perform the assignments of clinical codes to medical cases, but also to provide the reference in the text that justifies each coding assignment.
Methods: We examine the performance of 3 transformer-based architectures on 3 different explainable clinical-coding tasks. For each transformer, we compare the performance of the original general-domain version with an in-domain version of the model adapted to the specificities of the medical domain. We address the explainable clinical-coding problem as a dual medical named entity recognition (MER) and medical named entity normalization (MEN) task. For this purpose, we have developed two different approaches, namely a multi-task and a hierarchical-task strategy.
Results: For each analyzed transformer, the clinical-domain version significantly outperforms the corresponding general domain model across the 3 explainable clinical-coding tasks analyzed in this study. Furthermore, the hierarchical-task approach yields a significantly superior performance than the multi-task strategy. Specifically, the combination of the hierarchical-task strategy with an ensemble approach leveraging the predictive capabilities of the 3 distinct clinical-domain transformers, yields the best obtained results, with f1-score, precision and recall of 0.852, 0.847 and 0.849 on the Cantemist-Norm task and 0.718, 0.566 and 0.633 on the CodiEsp-X task, respectively.
Conclusions: By separately addressing the MER and MEN tasks, as well as by following a context-aware text-classification approach to tackle the MEN task, the hierarchical-task approach effectively reduces the intrinsic complexity of explainable clinical-coding, leading the transformers to establish new SOTA performances for the predictive tasks considered in this study. In addition, the proposed methodology has the potential to be applied to other clinical tasks that require both the recognition and normalization of medical entities.

1. Introduction

Clinical coding of textual healthcare documents aims to assign standardized diagnosis and procedure codes to the different free-text sections that make up electronic health records (EHR): admission and discharge summaries, diagnosis test and pathology reports, nursing care reports, clinical notes, etc. [1]. These codes constitute a sort of summarized and objective information regarding patients' diseases and their associated clinical care and allow for the efficient and systematic accomplishment of subsequent research and medical audit tasks. Traditionally, clinical coding of EHRs in hospitals and healthcare centers has been carried out manually, resulting in a tedious task, with a high

* Corresponding author at: Campus de Teatinos, Bulevar Louis Pasteur, 35, 29071 Málaga, Spain.
E-mail address: guilopgar@uma.es (G. López-García).
[1] Address: Campus de Teatinos, Bulevar Louis Pasteur, 35, 29071 Málaga, Spain.
[2] Address: Campus de Teatinos s/n, 29010 Málaga, Spain.

https://doi.org/10.1016/j.jbi.2023.104323
Received 20 September 2022; Received in revised form 11 February 2023; Accepted 15 February 2023
Available online 20 February 2023

propensity for errors, which requires a huge human effort to be carried out. However, clinical coding can be automated [2,3], making it easier for professional coders to provide more accurate results.

Standardized clinical coding systems, such as ICD or SNOMED-CT, represent patient diagnoses, procedures, and other information using controlled clinical terminology. The ICD[3] (International Statistical Classification of Diseases and Related Health Problems) classification system establishes a standardized coding that allows statistical analysis of morbidity and mortality of patients in private or public health systems. The 10th edition of the ICD, ICD-10, is structured hierarchically into chapters that group codes of up to seven characters, thus allowing the coding of over 70,000 diagnoses and 72,000 different procedures, which gives an idea of the complexity underlying clinical text coding. Thus, the length of ICD codes ranges from 3 to 7 characters, depending on the degree of specificity needed for the disease or procedure to be coded. An added intrinsic difficulty for the development of automated clinical coding systems is that the distribution of ICD codes in the available annotated corpora is highly unbalanced. Therefore, the MIMIC-III corpus [4], which is one of the English corpus used as a reference for many of the natural language processing (NLP) tasks on clinical text, contains a large number of medical records with a few disease-specific ICD-9 codes. Thus, the three most frequent ICD-9 codes in MIMIC-III are 401.9 (unspecified essential hypertension), 428.0 (congestive heart failure, unspecified) and 427.31 (atrial fibrillation), which are present in 37.5 %, 23.8 % and 23.4 %, respectively, of the records. The hundredth most frequent code, V10.46 (personal history of malignant neoplasm of prostate) only appears in 2 % of MIMIC-III discharge reports [5].

A fundamental aspect regarding the applicability of automatic clinical coding is the explainability of the predictive models used. Most of the approaches used in clinical coding models act as "black boxes" [6], without giving a deeper and more detailed view of the reason for the selection of each label that the model automatically assigns to each text chunk. In general, eXplainable Artificial Intelligence (XAI) aims to understand why a certain predictive or classification algorithm obtains a specific result as output [7]. For the particular case of clinical coding, XAI aims to provide information that makes it possible to explain the assignment to a clinical text of certain ICD codes, to the detriment of others, in order to motivate the output proposed by the algorithm and thus support clinical decision-making. XAI is one of the priority lines of the DARPA projects [8], whose main objective is to create artificial intelligence (AI) systems whose models and decisions can be understood and thus trusted by final users. For its part, the European General Data Protection Regulation (GDPR)[4] promotes the explainability of the logic underlying automatic decision-making, considering "black box" AI models as an unfair and misleading business practice. Although deep learning (DL) algorithms are getting in general better performance results at most domains and tasks than traditional machine learning (ML) approaches, they are inherently less transparent than traditional methods due to their extreme complexity and the huge number of parameters that their network architectures of non-linear units contain [7].

In this article, we propose and evaluate different multilingual Transformer-based approaches, with in-domain adaptation, for explainable clinical coding. These models not only assign standardized disease and procedure codes to clinical texts, but also provide information related to the exact text spans that motivate the choice of each of these codes given as output. For this, the performance of two different multilingual transformers [9], such as XLM-RoBERTa [10] and mBERT [11], as well as a Spanish-based transformer model, called BETO [12], is analyzed. These pretrained models are fine-tuned on a corpus of a specific clinical domain, as is the case of a clinical oncology corpus in Spanish, and then trained and evaluated on downstream clinical coding tasks. In addition, we compare two different training strategies for explainable clinical coding: a hierarchical-task approach versus a multi-task approach. For the former, a first transformer that tackles a medical named entity recognition (MER) task is trained to identify clinical entities, i.e. text spans with diagnosis or procedure relevant information. The outputs of this MER transformer are subsequently used to train a second transformer that deals with a medical named entity normalization (MEN) task to assign ICD-10 labels to the clinical entities recognized by the first transformer. For the multi-task approach, the MER and MEN transformers are instead trained in parallel. As we will see in the following sections, our hierarchical-task MER + MEN approach for explainable clinical coding obtains better performance rates than our multi-task counterpart. Moreover, the performance of in-domain adapted transformers surpasses that of their non-adapted versions in all the scenarios analyzed herein. These different multilingual transformers and training approaches proposed in this study are evaluated on public corpora obtained from explainable clinical coding shared-tasks—namely CodiEsp-X [13], within the shared tasks of the e-Health CLEF 2020, and Cantemist-Norm [14], from the IberLEF 2020 conference—, obtaining for both cases results that exceed the current state-of-the-art (SOTA) for these shared tasks.

In the last few years, a new family of models has emerged capable of associating a contextual numerical representation of each word, taking into account the specific context in which the word appears within the text. These types of models are known as *contextual embeddings*. Some of them are based on semi-supervised sequence learning, as is the case with ELMo [15], ULMFit [16], Transformer [9], BERT [11] and more recently T5 [17] and XLNet [18]. The BERT model [11], based on the Transformer architecture [9], has been standing out among all those before it for allowing not only the extraction of contextual representations of words, but also the resolution of subsequent downstream tasks (such as text classification, NER, text summarization, information extraction, automatic translation, etc.). While those based on recurrent neural networks (usually bidirectional LSTM networks), such as ELMo [15] or ULMFit [16], present efficiency issues due to the sequential nature of these networks, models based on Transformer focus on the attentional mechanisms proposed in [9] to, among other advantages, increase computational efficiency by parallelizing much of the network architecture. Another peculiarity of the Transformer-based models (or *transformers*, for short) is that they can be pre-trained on a general domain corpus and later fine-tuned and adapted on a specific domain corpus to solve a specific NLP task. This technique, known as *transfer learning* [17,19], is commonly used to fit DL algorithms to small data sets. Most of the available transformers, such as BERT, have been trained using English corpora. This makes them less efficient at tackling NLP tasks in other languages, such as Spanish. In recent years a series of new transformers has emerged, such as XLM-RoBERTa (XLMR) [20,21] or Multilingual BERT (mBERT) [11], which have been pre-trained on multilingual corpora composed of texts in hundreds of languages and subsequently fine-tuned on monolingual corpora. These multilingual transformers have been proven to be effective models when tackling several NLP downstream tasks, such as named entity recognition (NER) or text classification [22,23].

2. Related work

In [6], that is one of the first systematic reviews on the SOTA of automatic clinical coding and classification systems, they analyze a total of 113 different studies, most of which apply rule-based strategies, regular expressions and grammars. These techniques suffer, in general, from not having enough generalization capacity. In recent years, the field of clinical coding has advanced remarkably thanks to the open publication of labeled clinical corpora of standardized clinical coding data. This is the case of the corpus published in [24], which contains labeled radiological reports or, more recently, the MIMIC-III corpus [4], with more than 50,000 discharge reports from the critical care units at

[3] https://www.who.int/classifications/classification-of-diseases.
[4] https://eur-lex.europa.eu/eli/reg/2016/679/oj.

the Beth Israel Deaconess Medical Center. With these datasets, various DL strategies have been used, which have generally shown higher performance rates than traditional ML methods. Thus, for example, [25] compares methods based on DL, namely convolutional neural networks (CNN), with various ML techniques, such as SVM, random forests and logistic regression, obtaining comparable or better results for DL strategies even without hyperparameter optimization. In [26] CNNs with attentional mechanisms were used for clinical coding, obtaining SOTA results when applied on the MIMIC-III dataset. Based on this latest work, in [27] a *per-label* attentional mechanism was incorporated to the convolutional architecture proposed in [26], which further improved the clinical coding performance of the models.

2.1. Clinical coding in Spanish

In the particular case of clinical coding in Spanish, not many works have been published, most of which use corpus of restricted use, not published for data protection reasons and, therefore, not available for analysis and the development of new competitive models. Thus, in [28] a non-public corpus in Spanish, from the hospital setting of the Basque Country health system, was used to train DL models for automatic clinical coding. For its part, in [29] an approach based on latent Dirichlet allocation (LDA) was used to perform multi-label classification of texts from EHRs obtained from the cardiology department of the Basque Country public health system, obtaining positive results for the 124 most frequent ICD codes present in that corpus. More recently, [30] compared algorithms based on binary outputs and extreme classification algorithms to assign ICD-10 codes to clinical texts in discharge reports from the Fundación Alcorcón University Hospital in Spain. This work concluded that the use of assembly methods based on the weighting of each code according to its frequency and its performance during training allows to obtain better results in extreme distributions, such as the one corresponding to the assignment of ICD codes to clinical texts. The authors of the current paper have developed automatic coding algorithms for clinical texts in Spanish, in which they have pre-trained, by using a non-public corpus of the clinical-oncology domain, several Transformer-based models to address downstream clinical coding tasks. These models were applied to both general clinical texts—namely, CodiEsp-D [31]—and oncology texts in particular—namely, Cantemist-Coding [32]—to obtain SOTA results [23].

2.2. Explainable clinical coding

The aforementioned corpora, MIMIC-III and the radiology reports corpus from [24], contain clinical documents in English, labeled with ICD-9 codes without any reference to the text segments (or spans of text) that give support to the assignment of those codes. Few studies have explored the explainability of ML and DL models for clinical coding. One of the most representative and cited works in the specific literature is [26], where the authors compared the ability of different models to identify in the text the *n*-grams related to each ICD code that is predicted. A manual and qualitative evaluation showed that a CNN architecture with attentional mechanisms could generate more significant and relevant *n*-grams for the clinical coding labels selected by the model. In [33] a Hierarchical Attention bi-directional Gated Recurrent Unit (HA-GRU) was proposed to produce a sentence-level explanation for each predicted code, instead of an *n*-gram-level explanation. Following this same line of proposing hierarchical architectures, [34] presented an approach to increase explainability that uses a Hierarchical Label-wise Attention Network (HLAN) that employs the weights both at word-level and at sentence-level to perform automatic coding. In [35] they used an architecture based on transformers to capture the interdependence between the tokens of a document, using an attentional mechanism for each different ICD code in order to learn specific representations of the entire document. To handle the frequency imbalance of the codes in the dataset, they used *label distribution aware margin* (LDAM) as the loss function. They applied this strategy to MIMIC-III discharge reports, resulting in a micro-AUC of 0.923. In [36] knowledge graphs and attentional mechanisms were used to train a multi-CNN architecture that, together with the use of adversarial learning to supply adversarial samples, allowed MIMIC-III discharge reports to be encoded using ICD-9 codes. They obtained micro-F1 results of 0.692 and allowed, through an analysis of the attentional weights associated with the models, that the predicted codes were explained by the spans of text highlighted as the most relevant.

Regarding automatic explainable clinical-coding in Spanish, recently in [27] CNNs with attentional mechanisms were used for clinical coding with ICD-10 codes on a non-public corpus in Spanish, achieving results that exceeded those obtained with other approaches that use CNN without attentional mechanisms. In that work, the authors carried out an a posteriori analysis of the explicability provided by the attentional mechanisms of the models, specifically that derived from the analysis of the attentional weights, as well as the chronological ordering of the ICD codes identified in the clinical histories for future predictive tasks.

In all these works cited so far, the absence of conveniently labeled information in the corpus —due to the lack of data on the text segments that motivate the assignment of each label— means that the explainability obtained by the models cannot be evaluated with quantifiable metrics that allow the objective comparison of the results, but in most cases only a subjective assessment of the explainability results obtained can be made.

In 2020, two independent initiatives were carried out to advance explainable automatic coding systems, within the competitive health tasks of the CLEF 2020 and the IberLEF 2020 international forums. As a result of these tasks, several corpora were made available for the NLP community. Two of those corpora supplied datasets of labeled samples of clinical texts in Spanish, with information on ICD-10 codes assigned to those samples. The organizers also provided information on the exact spans of text that motivated the assignment of those codes to each text chunk. Specifically, for the CodiEsp-X task [13] on explainable clinical coding in Spanish a corpus of 1000 clinical cases was provided, with 16,504 sentences classified with ICD-10 labels that included references to text segments that explained the selection of the labels. The best model presented in this shared task [37] proposed an approach based on BERT and semantic linking that obtained f1-score of 0.661, precision of 0.687 and recall of 0.562, and made use of data augmentation techniques to generate synthetic inputs to fine-tune pre-trained BERT models [11]. On the other hand, for the Cantemist-Norm shared task [14], texts from 1,301 clinical oncology cases labeled with ICD-O (oncology) codes were provided, with additional information about the text spans that motivated the choice of each code. The best model presented to this shared task obtained f1-score of 0.825, precision of 0.824 and recall of 0.826 [38], by using also BERT transformers. To the best of our knowledge, CodiEsp-X and Cantemist-Norm are the first two publicly available corpora, not only in Spanish but in any other language, for which information regarding the explainability of clinical coding is provided and, therefore, that can be used to objectively evaluate the performance of models oriented to explainable clinical coding.

There exist similar strategies in the literature that, like the one presented in this study, use models that join together MER and MEN approaches—either composed of transformers, recurrent neural networks or other NLP and ML models—to tackle explainable clinical-coding tasks. They can either follow a multi-task strategy [38–40] or a hierarchical-task (or "pipelined") strategy [37]. Furthermore, other groups of studies pose explainable clinical coding as a single MEN task [41–43]. However, for the particular case of models that follow these MER and/or MER approaches to tackle explainable clinical-coding on corpora that supply text spans labeled with ICD information, only three works deserve to be cited as noteworthy antecedents. On the one hand, in [38], which became the SOTA at the time Cantemist-Norm was held, the authors used a multi-task approach for clinical coding that used BERT as the core model to tackle MER as a machine comprehension task

and MEN as a sequence labeling task. On other hand, the SOTA for the CodiEsp-X shared task was reached, at the time the conference was held, by [37], that proposed a hierarchical-task approach that used a BERT model for MER, together with a semantic linking strategy for MEN. Additionally, in [44], CodiEsp-X was one of the different corpora used to address biomedical named entity linking tasks. The main contribution of this study was the creation of a tool for mapping identifiers between clinical ontologies and lexical resources. Besides, the authors also experimented with sequence labeling transformer-based models for detecting diagnosis and procedure concepts, although clinical coding tasks were not tackled in this study. In our work, we compare two end-to-end approaches—a multi-task versus a hierarchical-task approach—for both MER and MEN based on transformers adapted to the clinical-oncological domain in Spanish. In our multi-task approach both MER and MEN are tackled as sequence-labeling tasks. For its part, in our hierarchical-task strategy MER is considered as a sequence-labeling task, while MEN is carried out by following a text classification approach. The latter allows to reduce the intrinsic complexity of labeling the entities recognized by the MER module with labels that came from a highly imbalanced distribution of ICD codes.

Finally, for reproducibility purposes, all the data and code needed to replicate our work, is publicly available at https://github.com/guilopgar/TransformersExplClinicalCoding.

3. Paper contributions

The main lines of contribution of this work are the following:

- We systematically analyze the performance of transformer-based models for explainable clinical-coding. For this purpose, we compare the performance obtained by both clinical-domain and general-domain versions of the models when following multi-task and hierarchical-task approaches to tackle the problem of explainable clinical-coding.
- Our proposed hierarchical-task strategy leverages the context-aware predictive capabilities of transformers to achieve new SOTA performance on 3 different explainable clinical-coding tasks. Additionally, the developed methodology can be applied to other medical tasks involving both the detection and normalization of clinical entities.
- Finally, we deeply examine the differences in performance between the multi-task and the hierarchical-task strategies, identifying the crucial aspects of the hierarchical-task setting that lead to the observed increase in explainable clinical-coding performance.

4. Materials and methods

A schematic description of our developed methodology is shown in Fig. 1. In the next subsections, a more detailed description of the materials and methods is provided.

4.1. Corpora

In this section we describe the four corpora used in the different phases of this study. On the one hand, we characterize the corpus used for unsupervised pretraining and adaptation of several multilingual transformers to the specificities of the clinical domain. On the other hand, the corpora used for supervised training of the resulting transformers—to tackle three different explainable clinical-coding tasks—are also described in detail in the following subsections.

a) In-domain pre-training

With the goal of adapting transformers to the specificities of the medical domain, for this study we pretrain several multilingual transformers models on a private collection of real-world clinical cases retrieved from the Galén Oncology Information System [45]. The corpus contains 30.9 K oncology medical documents written in Spanish by physicians from the *Hospital Regional Universitario* and the *Hospital Universitario Virgen de la Victoria* in Málaga, Spain, comprising a total of 64.4 M words and 437.6 M characters.

b) Explainable clinical coding

We tackle three different explainable clinical-coding tasks, derived from two medical NLP shared tasks, namely the CodiEsp-X and the Cantemist-Norm tasks. On the one hand, the CodiEsp-X task is based on the CodiEsp corpus [13], a collection of 1 K medical cases in Spanish annotated with both ICD-10-CM[5] diagnosis and ICD-10-PCS[6] procedure codes. The CodiEsp-X task is separated into two different subtasks: CodiEsp-X-D and CodiEsp-X-P. The CodiEsp-X-D subtask focuses on diagnosis mentions, and it comprises the annotations from the CodiEsp-X task corresponding to ICD-10-CM codes. For its part, the CodiEsp-X-P subtask focuses on procedure mentions, comprising the ICD-10-PCS coding annotations from the CodiEsp-X task. On the other hand, the Cantemist-Norm task is based on the Cantemist corpus [14], a collection of 1.3 K oncology clinical cases in Spanish where tumor morphology mentions were annotated with ICD-O-3[7] codes. Both the CodiEsp and Cantemist corpora were split into training, development and test subsets.

In this way, the CodiEsp-X-D, CodiEsp-X-P and Cantemist-Norm tasks are each one centered on the prediction and normalization of a different type of medical concept. For all these three tasks, each available annotation assigns a particular clinical code to a medical document and additionally indicates the reference in the text that supports that coding assignment (see Fig. 2-A). In this way, each annotation available in the corpora comprises a textual mention (i.e., a reference to the text span bounding a mention) of a clinical concept and the corresponding standardized clinical code assigned to that mention. Table 1 summarizes the annotation distribution for the three tasks addressed in this study.

4.2. Transformer-based models

In this study, we tackle explainable clinical-coding problems using transformer-based models. Since the corpora considered in this work comprise clinical cases written in Spanish, we explore three transformers that support the Spanish language, namely, BETO [12], mBERT [11] and XLM-R [10].

- **BETO**: the Spanish BERT model, named BETO, employs a similar architecture to the BERT-Base model, with ∼ 110 M trainable parameters [12], and it uses a Spanish vocabulary of ∼ 31 K subwords.
- **mBERT**: it corresponds to the multilingual version of the BERT-Base model [11], pretrained on a collection of texts from 104 distinct languages. This model uses a multilingual WordPiece vocabulary of ∼ 110 K subwords, and it has ∼ 177 M trainable parameters.
- **XLM-R**: pretrained on a Common Crawl Corpus in 100 languages, the XLM-R model is the multilingual version of the RoBERTa-Base transformer [10], using a large multilingual SentencePiece vocabulary of ∼ 250 K subwords. The total number of trainable parameters of the model is ∼ 278 M.

4.3. Unsupervised in-domain pretraining

With the aim of adapting the aforementioned transformers to the specificities of a clinical domain, they are further pretrained on a corpus of unlabeled real-world oncology medical documents. We follow the same pretraining pipeline developed in [23], where in-domain adaptation of transformers was also performed. Specifically, mBERT and BETO are optimized on the basis of the Next Sentence Prediction (NSP) and the

[5] https://www.cms.gov/Medicare/Coding/ICD10/2018-ICD-10-CM-and-GEMs.

[6] https://www.cms.gov/Medicare/Coding/ICD10/2018-ICD-10-PCS-and-GEMs.

[7] https://seer.cancer.gov/icd-o-3/.

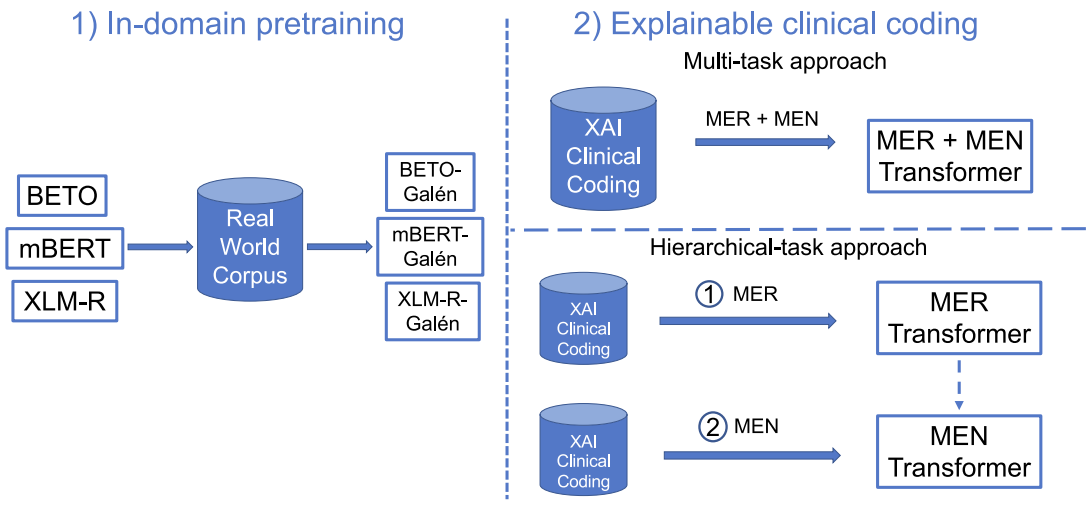

Fig. 1. Schematic representation of the methodology developed in this work for explainable clinical coding using transformers.

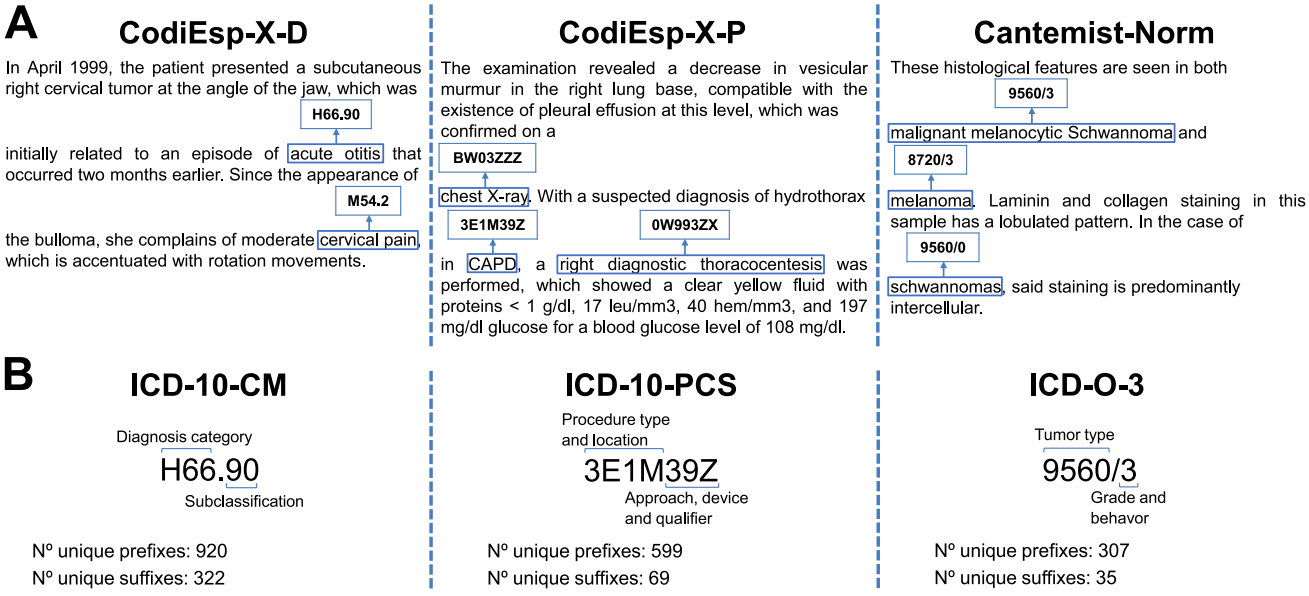

Fig. 2. **A.** Illustration of the explainable clinical-coding annotations format. **B.** Description of the code prefix and suffix pair obtained from each type of ICD code. The number of distinct prefixes and suffixes obtained from the ICD-10-CM, ICD-10-PCS and ICD-O-3 codes are calculated exclusively considering the clinical codes contained in the training and development subsets of the CodiEsp-X-D, CodiEsp-X-P and Cantemist-Norm corpora, respectively.

Table 1
Description of the number of annotations used in each of the 3 explainable clinical-coding tasks.

	CodiEsp-X-D			CodiEsp-X-P			Cantemist-Norm		
	Train.	**Devel.**	**Test**	**Train.**	**Devel.**	**Test**	**Train.**	**Devel.**	**Test**
Documents	500	250	250	435	222	224	501	499	300
Total annotations	7209	3431	3665	1972	1046	1112	6396	6001	3635
Unique ICD codes	1767	1158	1143	563	375	371	493	520	386
Unique unseen ICD codes	–	427	363	–	164	143	–	250	107

Masked Language Model (MLM) pre-training objectives [11]. The MLM objective with the dynamic masking modification [10] is used in this study to perform the pretraining of the XLM-R model.

4.4. Supervised fine-tuning for explainable clinical-coding

In this study, we address the explainable clinical-coding problem as a dual mER-MEN task. On the one hand, the MER subtask consists in the detection of the textual evidence of a coding assignment. Thus, from a MER task perspective, a single type of medical entity (i.e., diagnosis, procedure or tumor morphology) has to be recognized in each of the 3 different tasks tackled in this work. On the other hand, a medical entity must also be normalized. Hence, the MEN subtask corresponds to the assignment of a certain ICD code to each recognized entity. To perform this normalization procedure, following previous works [38,46], we leverage the hierarchical and multi-axial nature of the ICD coding

system to perform the clinical-coding assignment. As shown in Fig. 2-B, each ICD code can be split into two parts: a code prefix and a code suffix. The code prefix constitutes its basic part and describes general coding information, whereas the code suffix details more specific information and it may be absent. For instance, in the case of the ICD-10-PCS codes each code has a length of either 4 or 7 characters, with the first 4 symbols describing the type of procedure and its location, and the last 3 characters specifying the approach, device and qualifier information [47].

With the aim of addressing explainable clinical-coding with transformers, in this study we develop two distinct end-to-end approaches to tackle the dual mER-MEN task: a multi-task and a hierarchical-task strategy. In the following paragraphs, a detailed description of each end-to-end approach is given.

a) Multi-task approach

In the multi-task setting that we have followed for this study, both MER and MEN subtasks are addressed as sequence-labeling tasks, which are performed in parallel by a single transformer model. Specifically, the MER task corresponds to a multi-class sequence labeling problem, using the IOB2 [48] tagging scheme. On the other hand, for the MEN task, the goal is to assign an ICD code to each word that is part of a medical entity. However, instead of predicting the complete ICD code directly, we divide the problem into two distinct subtasks: one dedicated to the prediction of the prefix of the code, and the other one dedicated to the prediction of the code suffix. Given the highly imbalanced distribution of ICD codes, the rationale is to decrease the intrinsic complexity of the clinical coding assignment, since multiple codes prefixes and suffixes are shared across codes (see Fig. 2-B and Table 1). In this way, the MEN task has two classification objectives, each one corresponding to a multi-class sequence labeling subtask. Fig. 3 shows a schematic representation of the multi-task approach, and each of its six stages are described below.

Phase 1: Subword-level annotations. Firstly, word-level annotations are generated for both MER and MEN tasks. In the case of the MEN task, an additional "O" label is used to indicate the absence of a code prefix or suffix associated with a certain word. However, transformer-based models do not work at word-level. Alternatively, they obtain the input tokens by further splitting words into a sequence of subwords, each model using a particular tokenizer. For this reason, the original word-level annotations are converted to subword-level (i.e., token-level) by plainly assigning the same label to all subwords obtained from the same word.

Phase 2: Model fine-tuning. Using the resulting subword-level annotations, a single transformer is fine-tuned on both MER and MEN sequence-labeling tasks. In this way, the output representation encoded by the model for each token is fed into 3 independent classification feed-forward layers. The first classification layer addresses the MER classification objective, using R softmax units, with $R = 3$—representing the "I", "O" and "B" tags of the IOB2 scheme, respectively. To tackle the two distinct MEN classification objectives, the second and third final layers use $P + 1$ and $S + 1$ softmax units, respectively—an extra unit representing the additional "O" label was added to both layers—, with P as the number of codes prefixes and S as the number of codes suffixes to be predicted.[8] Since each of these three classification objectives corresponds to a multi-class sequence labeling task, categorical cross-entropy loss is adopted. Finally, to perform the supervised fine-tuning of the models, the sum of the three corresponding categorical cross-entropy losses is computed to guide the optimization of the models' parameters.

Phase 3: Subwords predictions. Thus, at inference time, given a sequence of n subwords (i.e., n tokens) as input to the model, three different matrices are outputted by the classification layers (see Fig. 3, after step 3). On the one hand, a $n \times R$ matrix corresponds to the predictions made by the MER classification layer, containing the probability of each token to be classified with a certain IOB2 label. On the other hand, the MEN classification layers output two matrices, a $n \times (P + 1)$ matrix and a $n \times (S + 1)$ matrix, which contain the probability of each input token to be associated with a particular code prefix and suffix, respectively.

Phase 4: Word-level predictions. As a result of the previous phase, three distinct subword-level probability matrices are obtained. Nevertheless, given that both MER and MEN annotations are created at word-level, the subword-level predictions are to be converted to their word-level counterparts. For this purpose, an average-based probability criterion is applied separately on each of the three subword-level probability matrices. Thus, for the predictions made for all subwords obtained from a single word, we perform the average of the predicted probabilities across the corresponding subwords, independently for each label. Therefore, the probability of a word to be associated with a certain label corresponds to the arithmetic mean of the probabilities of its subword components to be assigned that label.

Phase 5: Entity recognition. Subsequently, the medical entities recognized by the model are identified. In this way, considering the MER word-level probability predictions obtained as a consequence of the previous stage, the IOB2 label with the maximum probability is assigned to each word (e.g., the label B for the word "acute", or the label I for the word "otitis", in the example of Fig. 3, after step 5). The information provided by these assigned IOB2 tags allows us to identify the words forming each detected clinical entity. Secondly, with the aim of performing the normalization of each recognized entity, the MEN word-level predictions are converted to entity-level predictions. For this aim, an average-based probability criterion is independently applied on each of the two MEN word-level prediction matrices obtained from the previous phase. Hence, the probability of a particular coding label—either a code prefix or suffix—to be assigned to an entity is calculated as the arithmetic mean of the probabilities of its forming words to be associated with the corresponding label (e.g., the entity "acute otitis" obtained a probability of 0.77 for the prefix label H66, as a result of averaging the probabilities of words "acute" and "otitis": $(0.69 + 0.85) / 2$, as shown in the example of Fig. 3, after step 5).

Phase 6: Entity normalization. Finally, considering the MEN entity-level probability predictions obtained in the previous stage, both the code prefix and suffix with the maximum probability are assigned to each medical entity. As a result, this final combination of each recognized entity with its associated clinical code fully specifies a clinical-coding assignment, indicating both the predicted code and the textual reference that explains that coding assignment.

b) Hierarchical-task approach

In contrast to the multi-task setting, in the hierarchical-task approach the MER and MEN subtasks are not performed in parallel, but hierarchically. In this way, a first transformer model addresses the MER task, by following the same sequence labeling approach employed in the multi-task strategy (see Fig. 4-A). Afterward, the MEN task is subsequently performed by a second transformer (see Fig. 4-B). The principal objective of the hierarchical-task setting is to reduce the intrinsic difficulty of the MEN task. For this reason, instead of following a sequence labeling approach, we tackle the MEN task as a text classification problem. Hence, for each medical entity recognized by the MER transformer, the MEN transformer has to assign its corresponding clinical code.

As it can be seen from Fig. 4-B, considering a sequence of tokens given as input to the MEN model, two additional elements are used in the hierarchical-task setting to supply the MEN transformer with information of the presence of a medical entity that has to be normalized: the medical entity tokens and the medical entity embeddings. On the one hand, two special tokens are inserted into the input sequence: the $< M >$ and the $</M >$ medical entity tokens. The $< M >$ token is inserted right before the first subword of the medical entity, and it aims to

[8] The values of P and S vary for the CodiEsp-X-D, CodiEsp-X-P and Cantemist-Norm tasks. In each case, both values match the number of unique prefixes and suffixes, respectively, of the clinical codes contained in the training and development subsets of the corresponding corpus (see Fig. 2-B).

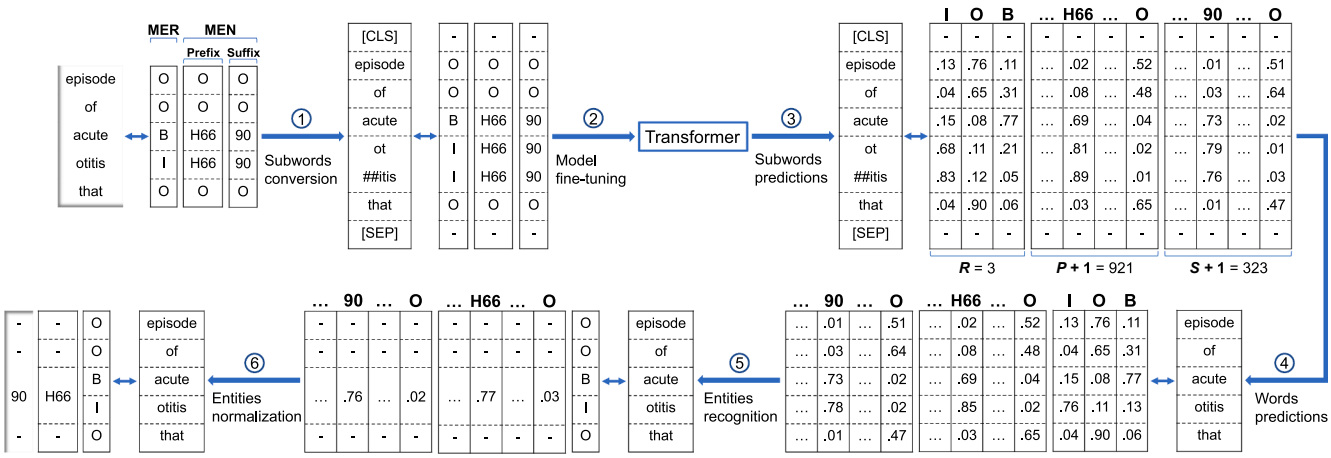

Fig. 3. Workflow of the six-phases multi-task approach developed to perform explainable clinical-coding using transformers. For illustration purposes, we use an annotated text fragment from the CodiEsp-X-D corpus as input to the model (see Fig. 2-A). Also, the WordPiece tokenizer of the mBERT model is employed to convert the input text fragment into a sequence of subwords.

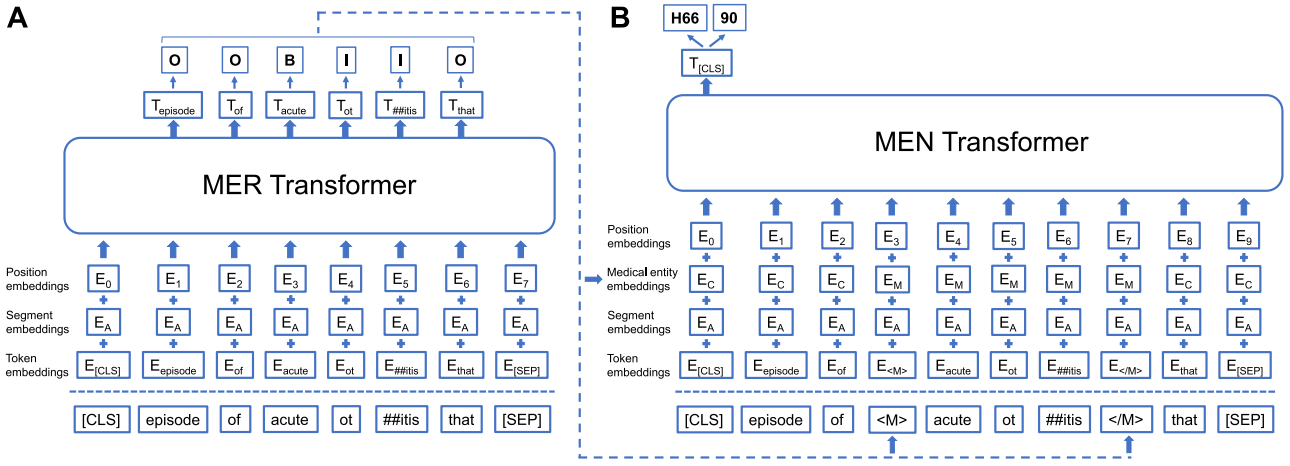

Fig. 4. Visual representation of the hierarchical-task approach. **A.** The sequence labeling MER transformer model. Given a sequence of subword tokens supplied as input to the model, the input representation of each token is constructed by summing the corresponding token, segment and position embeddings, as it was originally proposed by BERT [11]. **B.** The text-classification MEN transformer. Two additional elements are used to indicate the MEN model the clinical entity to be normalized: the medical entity tokens and embeddings. Consequently, the medical entity embedding is added to the sum of the token, segment and position embeddings to build the final input representation of each token.

identify the start of the entity, while the $< /M >$ token is inserted right after the last subword of the clinical entity, indicating the end of the entity. On the other hand, we use two additional medical entity embeddings: the E_C and the E_M embeddings. The E_C embedding is assigned to the tokens that are not part of the entity, i.e. those tokens that constitute the *context* part of the input sequence. For its part, the E_M embedding is assigned to each token belonging to the medical entity. Both the medical entity tokens and the medical entity embeddings are optimized during the fine-tuning of the MEN transformer model.

The usage of medical entity tokens and embeddings to supply the model with information of the presence of a clinical entity was inspired by the way BERT model itself is trained during the Next Sentence Prediction (NSP) task [11]. Thus, by means of special tokens and embeddings, BERT receives information to distinguish between the two sentences present in each sample used to train the model. For this purpose, BERT makes use of two special tokens—the *[CLS]* and *[SEP]* tokens—to delimit each sentence in the input, while segment embeddings are also used to give the model information of the tokens forming each of the two distinct sentences. Our hierarchical-task approach uses the same rationale to highlight the tokens of the input sequence forming the clinical entity to be normalized by the MEN transformer-based model.

Finally, in the next paragraphs, we give a description of the main differences between the fine-tuning and the prediction phases of the hierarchical-task approach.

Fine-tuning phase. At training time, two different transformers are independently fine-tuned on the MER and MEN tasks, respectively. On the one hand, we follow the same sequence labeling procedure used in the multi-task approach to fine-tune the model on the MER task, by using the IOB2 tagging scheme (see Phases 1–2 in Section 4.4.a and Fig. 4-A). On the other hand, regarding the MEN task, as it was formerly performed in the multi-task strategy, the problem is divided into two different multi-class classification subtasks to independently predict the prefix and the suffix of each ICD code. However, in contrast to the multi-task setting, in the hierarchical-task approach each of the two MEN classification objectives is addressed as a text classification task. For this reason, the output representation encoded by the model for the initial *beginning of sequence* (BOS) token—*[CLS]* token for the BERT-based models and $< s >$ token for the XLM-RoBERTa model—is fed into 2 different classification feed-forward layers with P and S softmax units, respectively (see Phase 3 in Section 4.4.a and Fig. 4-B). The categorical cross-entropy loss is adopted to tackle each multi-class classification

objective, and the sum of both losses is employed to guide the parameters optimization of the model. Lastly, regarding the training samples used to fine-tune the MEN transformer, we obtained them by using the available annotations from the corresponding corpus. In this way, for each text fragment inputted to the model, a single annotated clinical entity is considered to be normalized. Consequently, if a single fragment of text contains multiple annotated entities, one training sample is generated for each medical entity using the aforementioned medical entity tokens and embeddings (see Supplementary Fig. S1 for more details).

Prediction phase. At inference time, given the dependence of the MEN task on the entities detected as a result of the MER task, the predictions made by the MER transformer are firstly considered. Thus, following the same procedure developed in the multi-task strategy, the medical entities detected by the MER transformer model are to be firstly identified (see Phases 3–5 in Section 4.4.a). Subsequently, for each fragment of text containing a clinical entity previously recognized by the MER transformer, its corresponding clinical code is predicted by using the MEN transformer. Again, if a single fragment of text contains multiple detected entities, one inference sample is produced for each clinical entity using the medical entity tokens and embeddings (see Supplementary Fig. S1).

4.5. Experiments

During the unsupervised pre-training of the models, the same values for most of their hyperparameters are fixed (see Supplementary Table S1 for further details). When performing the supervised fine-tuning of the models, the RAdam [49] optimizer is employed with a learning rate of 3×10^{-5}, with a batch size of 16 and a number of epochs that is empirically estimated for each model using the development subset of each corresponding corpus, with an upper limit of 100 epochs. For all transformers analyzed in this work, we fix a maximum input sequence length of 128 subwords. However, the majority of the clinical documents from both the CodiEsp and Cantemist corpora has a sequence length clearly above 128. To overcome this limitation, we first split each document into sentences. Then, contiguous sentences are to be joined together in single fragments of text by using a greedy approach, so that the length of each fragment does not exceed that of 128 tokens. All models examined in this study employ the same fragments of text as input patterns. To tackle the two different classification objectives, both approaches use softmax output layers of $P + 1 = 308$ and $S + 1 = 36$ units for the Cantemist-Norm task, $P + 1 = 902$ and $S + 1 = 306$ units for the CodiEsp-X-D subtask, and $P + 1 = 447$ and $S + 1 = 64$ units for the CodiEsp-X-P subtask, respectively (see Section 4.4). Furthermore, given the considerable number of discontinuous textual references among the annotations, we employ an additional IOB2 prediction layer in the MER classification component (see Supplementary Fig. S2 for further details). This extra classification layer aims at recognizing the medical entities with discontinuous textual references, and it is only employed when tackling the CodiEsp-X-D and the CodiEsp-X-P tasks. Finally, regarding the hardware resources employed, all experiments are conducted on an exascale system of 4 NVIDIA DGX-A100 nodes.

5. Results

Tables 2, 3 and 4 show the performance of the three transformer-based models for explainable clinical-coding on the Cantemist-Norm, CodiEsp-X-D and CodiEsp-X-P tasks, respectively. For each transformer, we compare the original model pretrained on general domain corpora (see Section 4.2) with the corresponding version adapted to the clinical domain (see Section 4.3). Additionally, we also compare the predictive performance obtained by the models when they follow either the multi-task or the hierarchical-task approaches. The classification performance of the models is evaluated by using the micro-averaged precision, recall and f1-score metrics—i.e., the official evaluation metrics given for the analyzed tasks [13,14]. For each metric, the distribution of the values obtained from 5 distinct random fine-tuning instances of each model is described by reporting the mean, standard deviation and maximum values.

According to the f1-score, the three transformer models achieve the best performance rates when they have been adapted to the clinical-domain and follow the hierarchical-task approach. Among all models, mBERT-Galén obtains the highest average f1-scores by following the hierarchical-task strategy, with mean f1-scores of 0.826, 0.627 and 0.542 for the Cantemist-Norm, CodiEsp-X-D and CodiEsp-X-P tasks, respectively. For their part, the two other transformer-based models adapted to the medical domain achieve almost the same performance across the 3 tasks. Thus, by following the hierarchical-task strategy BETO-Galén obtains average f1-scores of 0.823, 0.625 and 0.533, while XLM-R-Galén achieves mean f1-scores of 0.823, 0.625 and 0.533 for the Cantemist-Norm, CodiEsp-X-D and CodiEsp-X-P tasks, respectively.

In this way, two distinct patterns emerge overall from Tables 2, 3 and 4. On the one hand, compared with the general-domain transformers, the clinical-domain version of the models improves their performance for the three distinct tasks analyzed herein. Thus, for each model, its "clinical version" achieves a higher average f1-score than the original general-domain model, when following both the multi-task approach and the hierarchical-task strategy. With the intention of verifying the statistical significance of the previous observation, in this study we perform non-parametric paired Wilcoxon signed-rank tests [50] to compare the distribution of the f1-score values obtained by the clinical-domain transformers with the f1-score values obtained by the general-domain models. In this way, for each specific task, the distribution of the 30 values of the f1-score metric achieved by the in-domain adapted transformers are compared with the 30 f1-score values obtained by the general-domain models. Significant p-values[9] are obtained for every task ($p = 2 \times 10^{-5}$ in the Cantemist-Norm task, $p = 2 \times 10^{-3}$ in the CodiEsp-X-D subatsk and $p = 3 \times 10^{-5}$ in the CodiEsp-X-P subatsk), hence accepting the alternative hypothesis—i.e., the performance of the clinical-domain transformers is significantly greater—in all cases.

On the other hand, the hierarchical-task strategy yields a superior performance across the three tasks. In this way, for every task, the mean f1-score achieved by each model when it follows the hierarchical-task approach is greater than the performance of the same model when it is trained with the multi-task strategy. Again, for each task, we perform the paired Wilcoxon signed-rank test to compare the distribution of the 30 f1-score values obtained by the models when following the hierarchical-task strategy with the corresponding 30 f1-score values achieved when using the multi-task approach. Significant differences can be observed for each task ($p = 1 \times 10^{-5}$ in the Cantemist-Norm, CodiEsp-X-D and CodiEsp-X-P tasks), thus proving that our transformer-based models achieve a significantly greater performance when they follow a hierarchical-task approach.

5.1. Ensemble

Additionally, we develop an ensemble approach to combine the different predictions made by the transformers. We adopt the ensemble strategy proposed in [23], adapting it to particularities of the methodology proposed herein to address the explainable clinical coding problem. Given the aforementioned superior performance of the hierarchical-task approach in comparison with the multi-task setting, we apply our ensemble approach to combine the predictions made by the models in this more favorable hierarchical scenario. Thus, the ensemble approach firstly combines the predictions made by the MER transformer at the word-level; then, based on the resulting predictions,

[9] All the p-values reported in this study have been corrected for multiple-tests using the Bonferroni procedure [50].

Table 2
Explainable clinical-coding performance of the transformer models on the Cantemist-Norm task. For the maximum values column of each metric, the best result obtained is bolded, while the second best is underlined.

Strategy	Model	Precision		Recall		F1-score	
		Mean ± Std	Max	Mean ± Std	Max	Mean ± Std	Max
	BETO	0.802 ± 0.011	0.820	0.797 ± 0.005	0.805	0.799 ± 0.004	0.804
	BETO-Galén	0.805 ± 0.005	0.810	0.806 ± 0.007	0.816	0.805 ± 0.004	0.813
	mBERT	0.805 ± 0.005	0.811	0.809 ± 0.006	0.816	0.807 ± 0.003	0.812
	mBERT-Galén	0.815 ± 0.011	0.830	0.814 ± 0.006	0.822	0.815 ± 0.004	0.819
	XLM-R	0.802 ± 0.007	0.814	0.806 ± 0.006	0.816	0.804 ± 0.005	0.810
Multi-task	XLM-R-Galén	0.812 ± 0.008	0.826	0.812 ± 0.003	0.817	0.812 ± 0.004	0.818
	BETO	0.818 ± 0.007	0.824	0.811 ± 0.004	0.816	0.814 ± 0.003	0.819
	BETO-Galén	0.825 ± 0.005	0.833	0.821 ± 0.005	0.826	0.823 ± 0.004	0.829
	mBERT	0.819 ± 0.009	0.832	0.818 ± 0.009	0.830	0.818 ± 0.002	0.820
	mBERT-Galén	0.828 ± 0.007	**0.839**	0.825 ± 0.004	0.830	0.826 ± 0.004	**0.832**
	XLM-R	0.814 ± 0.005	0.822	0.817 ± 0.004	0.822	0.815 ± 0.004	0.822
Hierarchical-task	XLM-R-Galén	0.824 ± 0.008	0.832	0.822 ± 0.008	**0.832**	0.823 ± 0.005	**0.832**

Table 3
Explainable clinical-coding performance of the transformer models on the CodiEsp-X-D task. For the maximum values column of each metric, the best result obtained is bolded, while the second best is underlined.

Strategy	Model	Precision		Recall		F1-score	
		Mean ± Std	Max	Mean ± Std	Max	Mean ± Std	Max
	BETO	0.672 ± 0.018	0.695	0.521 ± 0.014	0.533	0.586 ± 0.008	0.598
	BETO-Galén	0.664 ± 0.021	0.696	0.527 ± 0.005	0.534	0.587 ± 0.010	0.601
	mBERT	0.661 ± 0.017	0.684	0.540 ± 0.006	0.546	0.595 ± 0.004	0.598
	mBERT-Galén	0.671 ± 0.008	0.681	0.544 ± 0.002	0.546	0.601 ± 0.004	0.605
	XLM-R	0.643 ± 0.010	0.658	0.546 ± 0.005	0.550	0.591 ± 0.006	0.598
Multi-task	XLM-R-Galén	0.664 ± 0.010	0.677	0.543 ± 0.004	0.547	0.597 ± 0.007	0.605
	BETO	0.694 ± 0.005	0.702	0.559 ± 0.005	0.565	0.619 ± 0.001	0.622
	BETO-Galén	0.684 ± 0.009	0.695	0.576 ± 0.004	0.578	0.625 ± 0.004	0.631
	mBERT	0.694 ± 0.007	0.703	0.564 ± 0.003	0.568	0.622 ± 0.003	0.627
	mBERT-Galén	0.692 ± 0.006	**0.704**	0.574 ± 0.005	0.578	0.627 ± 0.004	**0.634**
	XLM-R	0.678 ± 0.012	0.691	0.564 ± 0.003	0.568	0.616 ± 0.004	0.622
Hierarchical-task	XLM-R-Galén	0.686 ± 0.010	0.695	0.575 ± 0.004	**0.582**	0.626 ± 0.004	0.629

Table 4
Explainable clinical-coding performance of the transformer models on the CodiEsp-X-P task. For the maximum values column of each metric, the best result obtained is bolded, while the second best is underlined.

Strategy	Model	Precision		Recall		F1-score	
		Mean ± Std	Max	Mean ± Std	Max	Mean ± Std	Max
	BETO	0.620 ± 0.008	0.634	0.401 ± 0.005	0.407	0.487 ± 0.005	0.493
	BETO-Galén	0.615 ± 0.028	0.647	0.423 ± 0.008	0.436	0.501 ± 0.007	0.510
	mBERT	0.607 ± 0.005	0.612	0.408 ± 0.009	0.416	0.488 ± 0.007	0.494
	mBERT-Galén	0.621 ± 0.014	0.642	0.418 ± 0.005	0.423	0.499 ± 0.007	0.510
	XLM-R	0.593 ± 0.018	0.624	0.413 ± 0.007	0.423	0.486 ± 0.006	0.496
Multi-task	XLM-R-Galén	0.613 ± 0.024	0.645	0.420 ± 0.008	0.427	0.498 ± 0.007	0.510
	BETO	0.635 ± 0.024	0.661	0.442 ± 0.007	0.449	0.521 ± 0.009	0.535
	BETO-Galén	0.636 ± 0.013	0.649	0.458 ± 0.012	**0.477**	0.533 ± 0.009	**0.548**
	mBERT	0.639 ± 0.023	**0.665**	0.441 ± 0.008	0.452	0.522 ± 0.007	0.533
	mBERT-Galén	0.647 ± 0.014	0.661	0.467 ± 0.004	0.471	0.542 ± 0.005	**0.548**
	XLM-R	0.618 ± 0.022	0.650	0.443 ± 0.010	0.454	0.516 ± 0.006	0.526
Hierarchical-task	XLM-R-Galén	0.625 ± 0.023	0.658	0.462 ± 0.011	0.474	0.531 ± 0.006	0.536

the clinical-coding probabilities outputted by the MEN transformer are also combined (see Section 4.4.b and Fig. 4).

Hence, in the first place, given a sequence of m words as input to the MER model, a $m \times R$ matrix of word-level predictions is produced (see Phase 4 in section 4.4.a). In fact, as a result of considering 5 distinct random executions for each fine-tuned model, 5 different $m \times R$ probability matrices are obtained from a single model. To combine the 5 distinct matrices into a single word-level probability matrix, our ensemble strategy plainly consists of summing the 5 probability matrices. Moreover, the ensemble approach could also be applied to combine the word-level predictions made by any number of different models by directly summing all the obtained probability matrices. Once a single $m \times R$ ensemble prediction matrix is generated, the recognized medical entities could be identified (see Phase 5 in Section 4.4.a).

Subsequently, the medical entities identified following the ensemble approach are normalized using the MEN transformer. In this way, for each recognized entity, the MEN model outputs two different probability vectors: a vector of length P and a vector of length S, which contain the probability of the entity to be assigned a certain code prefix and suffix, respectively (see Section 4.4.b). Since, as it was previously mentioned, 5 different random instances of each fine-tuned model are considered, 5 distinct vectors of length P and 5 vectors of length S are obtained from a single model. To combine the 5 different probability vectors into a single vector, again, our ensemble approach consists of summing the 5 distinct vectors. Also, the predictions made by any number of models for a particular recognized entity could be combined by plainly summing all

the probability vectors obtained. As a result of applying the ensemble approach, a single pair of *P*-length and *S*-length prediction vectors is produced for each identified medical entity. Finally, the label predicted with the maximum value in each vector corresponds to the clinical code assigned to the corresponding medical entity.

Table 5 describes the predictive performance of the ensemble strategy applied to combine both the predictions made by single models and the predictions outputted by multiple distinct models. With the aim of comparing the performance achieved by our ensemble approach with the current SOTA performance in the CodiEsp-X task [13], we join the results obtained by each ensemble model on the CodiEsp-X-D and CodiEsp-X-P subtasks to evaluate its overall performance on the CodiEsp-X task[10] (see Supplementary Table S2 for further details on the performance of the ensemble approach separately on both subtasks). In addition, we also compare the performance obtained by our ensemble strategy on the Cantemist-Norm task with the SOTA performance reported by the organizers of the shared task [14]. In this way, regarding the performance of the ensemble approach applied to single models, according to the f1-score metric, the mBERT-Galén ensemble obtains the best results on the Cantemist-Norm and CodiEsp-X tasks, with f1-score values of 0.842 and 0.622, respectively, thus surpassing the prior SOTA performance on both tasks. In relation to the ensemble approach applied to multiple models, the BETO-Galén + mBERT-Galén + XLM-R-Galén ensemble achieves the highest performance among all models analyzed in this work, with f1-scores of 0.849 and 0.633 on the Cantesmit-Norm and CodiEsp-X tasks, respectively, establishing a new SOTA performance for each of both explainable clinical-coding tasks.

6. Discussion

In this study, we have systematically analyzed transformer-based models for explainable clinical coding. With this aim, we have examined the performance of the models on the Cantemist-Norm, CodiEsp-X-D and CodiEsp-X-P tasks, by comparing the results obtained when following a multi-task approach with the performance achieved by following a hierarchical-task strategy. For the classification tasks explored in this work, the obtained experimental results demonstrate that transformer models achieve higher performance when following the hierarchical-task approach. In this section, we further examine the differences in performance between the multi-task and the hierarchical-task strategies, as well as identify the key aspects of the hierarchical-task setting that lead to the observed increase in performance.

Table 5
Explainable clinical-coding performance of the ensemble models on the Cantemist-Norm and CodiEsp-X tasks. For each metric, the best result obtained is bolded, while the second best is underlined.

Model	Cantemist-Norm			CodiEsp-X		
	P	R	F1	P	R	F1
BETO	0.836	0.828	0.832	0.708	0.542	0.614
BETO-Galén	0.838	0.834	0.836	0.695	0.557	0.619
mBERT	0.840	0.835	0.838	0.711	0.544	0.616
mBERT-Galén	0.843	0.840	0.842	0.707	0.556	0.622
XLM-R	0.835	0.833	0.834	0.693	0.546	0.610
XLM-R-Galén	0.843	0.838	0.840	0.696	0.560	0.620
BETO + BETO-Galén	0.845	0.835	0.840	0.712	0.556	0.624
mBERT + mBERT-Galén	0.846	0.841	0.844	0.722	0.552	0.626
XLM-R + XLM-R-Galén	0.845	0.841	0.843	0.712	0.559	0.626
BETO + mBERT + XLM-R	0.845	0.839	0.842	**0.724**	0.552	0.626
BETO-Galén + mBERT-Galén + XLM-R-Galén	**0.852**	**0.847**	**0.849**	0.718	**0.566**	**0.633**
Prior SOTA	0.824	0.826	0.825	0.687	0.562	0.611

[10] Although we have separately addressed the CodiEsp-X-D and CodiEsp-X-P subtasks, the organizers of the CodiEsp shared task reported the SOTA performance on the whole CodiEsp-X task [13].

6.1. Multi-task vs Hierarchical-task approaches

With the aim of performing a thorough comparison between both explainable clinical-coding approaches, we separately examine the performance of transformers on the MER and MEN components of our multi-task and hierarchical-task strategies. In Supplementary Tables S3, S4 and S5, the independent MER performance of the transformer-based models obtained by following each of the two strategies is described for the Cantemist-Norm, CodiEsp-X-D and CodiEsp-X-P tasks, respectively. According to the f1-score, each model achieves a significantly greater MER performance when following the hierarchical-task approach than when it is trained with the multi-task setting ($p = 1 \times 10^{-3}$ in the Cantemist-Norm task and $p = 1 \times 10^{-5}$ in both CodiEsp-X-D and CodiEsp-X-P tasks). Noticeably, the differences in MER performance are greater in the CodiEsp-X-D and CodiEsp-X-P tasks than in the Cantemist-Norm task. In this way, the differences in terms of average f1-score are 0.004, 0.011 and 0.021 in favor of the hierarchical-task approach for the Cantemist-Norm, CodiEsp-X-D and CodiEsp-X-P tasks, respectively. On the other hand, we perform an additional experiment to compare the performance of the MEN component of the multi-task and hierarchical-task settings in an isolated manner. In this way, we aim to prevent the performance of the MER components from influencing the performance of the MEN components—given the dependence of the MEN component on the clinical entities detected by the MER component. Hence, we compare the results obtained by the models in the multi-task setting (see Tables 2, 3 and 4), with the results achieved by the transformers in a modified version of the hierarchical-task approach in which the MEN transformer performs the normalization of the clinical entities previously detected by the MER component of the multi-task approach (see Tables 6 and 7). Consequently, the performance of the MEN components of the multi-task and hierarchical-task approaches are compared independently of their MER counterparts, since in both settings the same MER predictions are considered. When comparing the results shown in Tables 2, 3 and 4 with the results described in Tables 6 and 7, according to the f1-score, the MEN performance of every transformer is significantly higher in the hierarchical-task scenario ($p = 1 \times 10^{-5}$ in the Cantemist-Norm, CodiEsp-X-D and CodiEsp-X-P tasks). Also, the differences in MEN performance are higher in the CodiEsp-X-D and CodiEsp-X-P tasks than in the Cantemist-Norm task. In this way, according to the mean f1-score, an average improvement across all models of 0.010, 0.027 and 0.019 is observed in the Cantemist-Norm, CodiEsp-X-D and CodiEsp-X-P tasks, respectively.

Table 6
MEN performance of the hierarchical-task approach using the MER predictions of the multi-task setting on the Cantemist-Norm task. For the maximum values column of each metric, the best obtained result is bolded, while the second best is underlined.

Model	Precision		Recall		F1-score	
	Mean ± Std	Max	Mean ± Std	Max	Mean ± Std	Max
BETO	0.814 ± 0.013	0.834	0.809 ± 0.005	0.816	0.812 ± 0.005	0.818
BETO-Galén	0.819 ± 0.005	0.826	0.820 ± 0.005	0.825	0.819 ± 0.002	0.822
mBERT	0.813 ± 0.006	0.817	0.817 ± 0.004	0.821	0.815 ± 0.002	0.817
mBERT-Galén	0.825 ± 0.012	**0.843**	0.824 ± 0.004	**0.830**	0.825 ± 0.005	**0.832**
XLM-R	0.808 ± 0.003	0.812	0.812 ± 0.006	0.821	0.810 ± 0.002	0.813
XLM-R-Galén	0.821 ± 0.008	0.835	0.821 ± 0.003	0.825	0.821 ± 0.004	0.827

Table 7
MEN performance of the hierarchical-task approach using the MER predictions of the multi-task setting on both the CodiEsp-X-D and CodiEsp-X-P tasks. For the maximum values column of each metric, the best obtained result is bolded, while the second best is underlined.

Model	CodiEsp-X-D						CodiEsp-X-P					
	Precision		Recall		F1-score		Precision		Recall		F1-score	
	Mean ± Std	Max	Mean ± Std	Max	Mean ± Std	Max	Mean ± Std	Max	Mean ± Std	Max	Mean ± Std	Max
BETO	0.710 ± 0.019	**0.736**	0.544 ± 0.010	0.553	0.616 ± 0.005	0.623	0.651 ± 0.008	0.662	0.420 ± 0.007	0.428	0.511 ± 0.007	0.520
BETO-Galén	0.703 ± 0.019	0.735	0.561 ± 0.002	0.563	0.624 ± 0.008	**0.637**	0.645 ± 0.028	**0.683**	0.438 ± 0.013	**0.460**	0.521 ± 0.009	**0.531**
mBERT	0.688 ± 0.017	0.703	0.559 ± 0.007	0.564	0.617 ± 0.006	0.621	0.623 ± 0.006	0.630	0.424 ± 0.011	0.440	0.504 ± 0.009	0.518
mBERT-Galén	0.701 ± 0.005	0.707	0.566 ± 0.004	**0.572**	0.627 ± 0.003	0.630	0.643 ± 0.016	0.661	0.434 ± 0.008	0.442	0.518 ± 0.010	0.526
XLM-R	0.667 ± 0.015	0.688	0.564 ± 0.003	0.569	0.611 ± 0.007	0.620	0.613 ± 0.013	0.631	0.426 ± 0.011	0.439	0.502 ± 0.006	0.511
XLM-R-Galén	0.694 ± 0.009	0.704	0.565 ± 0.004	0.569	0.623 ± 0.006	0.628	0.636 ± 0.023	0.659	0.435 ± 0.010	0.447	0.517 ± 0.007	0.524

Recent works have shown the superior performance of multi-task approaches over hierarchical-task strategies in medical classification problems involving both the detection and normalization of clinical entities, such as medical concept normalization [39,40]. In the case of explainable clinical-coding, SOTA results were obtained for the Cantemist-Norm task at the time the shared-task was held by training BERT using a multi-task approach [38]. By jointly modeling the MER and MEN tasks, the multi-task setting leads the models to build shared representations between both tasks, which represents a potential advantage when the tasks are related, as it is the case for explainable clinical-coding. However, in contrast with the results obtained in the aforementioned works, we have experimentally demonstrated that, for the 3 explainable clinical-coding tasks tackled in this study, a hierarchical-task approach leads the transformer-based models to achieve a significantly higher performance than when following a multi-task strategy. Particularly, both MER and MEN tasks benefit from the hierarchical-task setting in terms of performance. In fact, we have observed a larger difference in performance between both approaches in the CodiEsp-X-D and CodiEsp-X-P tasks than in the Cantemist-Norm task for both MER and MEN. CodiEsp-X-D and CodiEsp-X-P tasks represent more complex classification problems than the Cantemist-Norm task, given the scarcity of annotated samples available for the CodiEsp corpus in comparison with the large number of distinct clinical codes to be predicted (see Table 1). Given the obtained results, we can identify-two key features of the hierarchical-task approach that have led to the observed increase in the performance of the models. On the one hand, the first crucial characteristic is to separately perform the MER and MEN tasks, in contrast with the multi-task approach in which both tasks are performed in parallel. This is supported by the fact that, by following the same sequence labeling approach, the MER component of the hierarchical-task approach obtains a higher performance than the MER part of the multi-task setting. Although by jointly performing MER and MEN the model can build shared representations for both tasks, this does not always represent an advantage, specially for complex classification problems where there is a considerable limitation on the number of available annotated samples. On the other hand, the second pivotal feature of the hierarchical-task setting is the text classification approach used to tackle the MEN task. In contrast with the multi-task strategy where the MEN problem is addressed as a sequence labeling task, the text classification procedure allows the MEN transformer of the hierarchical-task setting to focus on the normalization of a single clinical entity per input pattern to the model, hence alleviating the intrinsic complexity of the MEN task. In fact, when comparing the results obtained by the MEN components of the multi-task and hierarchical-task strategies, we observe the highest MEN performance gains of the hierarchical-task approach in both the CodiEsp-X-D and CodiEsp-X-P tasks, which present higher unbalanced distributions of ICD codes than the Cantemist-Norm task.

Finally, we further evaluate the predictive capabilities of the text classification MEN component of our hierarchical-task approach. In the context of medical entity linking, it is very common to observe a large overlap between the train and test subsets of the benchmark datasets derived from shared tasks—like the datasets employed in this work, both at the level of coding labels and mentions [51–53]. For this reason, following the evaluation strategies proposed in previous works [51,53], we have evaluated the generalization capabilities of the MEN transformer of the hierarchical task approach in three different setups, namely *zero-shot*, *few-shots* and *filtering* setups. Supplementary Table S6 and Table 8 contain the description of the data used to evaluate the model on each setting, as well as its classification performance, respectively. For the three tasks addressed in this study, although an expected degradation in performance is observed, the text classification MEN approach shows a robust performance in the configurations with limited overlap between train and test sets, both at the level of codes and mentions. Additionally, we compare the performance achieved by the MEN transformer with the results obtained by SapBERT [54], a transformer-based model that has shown to achieve SOTA performance using Cantemist-Norm and CodiEsp-X-D corpora for medical entity linking [53]. When considering all mentions from the test set (see *full* row in Table 8), our MEN text classification approach outperforms SapBERT in both Cantemist-Norm and CodiEsp-X-D tasks. On the other hand, when the overlapping mentions are removed from the test set (see *filtering* row in Table 8), the MEN transformer obtains a higher classification accuracy than SapBERT in the Cantemist-NORM task, while SapBERT outperforms our text classification strategy in the CodiEsp-X-D task. With this analysis we demonstrate that the MEN component of our hierarchical-task approach also achieves strong performance in settings that limit leakage between the training and evaluation sets, obtaining comparable results with SOTA models for medical entity linking. However, as it can be observed from Table 8, in terms of performance, the most challenging setup for our MEN strategy is the *zero-shot* scenario. This is an expected consequence of the text classification approach used by the hierarchical-task strategy (see Section 4.4.b), since only codes that have been seen during training can be predicted at inference time. In future works, we will focus our efforts on improving the performance of our hierarchical-task approach in the *zero-shot* setup, by combining an entity linking approach with the context-aware normalization of clinical entities performed by our proposed methodology.

6.2. Context-aware normalization of clinical entities

Another distinctive characteristic of the proposed hierarchical-task strategy is the ability of the MEN transformer to leverage the specific context in which a clinical entity appears within the text to perform its

Table 8
Performance of the MEN component of the hierarchical-task setting in the *full*, *filtering*, *few-shots* and *zero-shot* setups. Accuracy is used as the evaluation metric. The performance of the mBERT-Galén model is reported. With the aim of exclusively evaluating the performance of the MEN component, the gold-standard mentions of the test set are used as input to the MEN text classification transformer. Additionally, we also report the Acc@1 results obtained by the SapBERT model in both *full* and *filtering* setups [53]. Specifically, the performance of the *SapBERT + target* configuration is described, obtained by further fine-tuning the SapBERT model on the corresponding explainable clinical coding corpus. *SapBERT + target* represents the best performing model in both Cantemist-Norm and CodiEsp-X-D tasks among the approaches analyzed in [53].

Model	Experiment	Cantemist-NORM				CodiEsp-X-D				CodiEsp-X-P			
		All		Subset		All		Subset		All		Subset	
		Mean ± Std	Max	Mean ± Std	Max	Mean ± Std	Max	Mean ± Std	Max	Mean ± Std	Max	Mean ± Std	Max
mBERT-Galén	Full	0.894 ± 0.002	0.896	0.927 ± 0.002	0.929	0.715 ± 0.002	0.717	0.755 ± 0.003	0.758	0.546 ± 0.006	0.556	0.798 ± 0.009	0.809
	Filtering	0.640 ± 0.004	0.645	0.735 ± 0.005	0.742	0.307 ± 0.004	0.310	0.390 ± 0.005	0.395	0.278 ± 0.008	0.288	0.552 ± 0.020	0.578
	Few-Shots	0.544 ± 0.014	0.562	0.637 ± 0.016	0.658	0.405 ± 0.005	0.412	0.481 ± 0.006	0.489	0.285 ± 0.007	0.295	0.590 ± 0.014	0.610
	Zero-Shot	0.170 ± 0.013	0.182	0.304 ± 0.023	0.325	0.007 ± 0.002	0.010	0.011 ± 0.004	0.016	0.023 ± 0.003	0.027	0.227 ± 0.032	0.273
SapBERT + target	Full	–	0.795	–	–	–	0.672	–	–	–	–	–	–
	Filtering	–	0.533	–	–	–	0.476	–	–	–	–	–	–

normalization. In this way, given an input sequence of text, our hierarchical-task approach uses the medical entity tokens and embeddings to highlight the tokens of the input sequence forming the clinical entity, providing the model with the textual information of both the clinical entity to be normalized and the context part of the input sequence. Prior works have also tackled the normalization of medical entities as a text classification problem using transformers [41–43]. However, in these previous studies, the models were exclusively provided with the tokens corresponding to each clinical entity, and no contextual information was employed as input data. Transformers are based on the self-attention mechanism [9], which permits the models to extract contextual representations of the input tokens, taking into account the specific context where they appear within the input text. Providing no information about the context in which a particular medical entity emerges within a clinical text considerably limits the capacity of the transformer-based models to build effective contextual representations of the clinical entities to be normalized. Additionally, considering the particular textual context where medical entities occur may be beneficial to perform certain ICD coding assignments.

In fact, there are cases in which the context of a clinical entity is not only beneficial, but critical to perform the entity normalization. In this way, we have identified multiple annotations in which clinical entities with the same textual mention are assigned different ICD codes. For instance, in the *S0211-57352015000100011–2* clinical case of the CodiEsp-X-D corpus, the medical entity *cognitive impairment* is assigned the F09 code ("unspecified mental disorder due to known physiological condition"). However, in the *S1130-01082007001100012–1* document of the same corpus, a medical entity with the same textual mention is assigned the G31.84 code ("mild cognitive impairment"). Although both clinical entities have the same textual description, the textual context of the former entity describes a severe disorder with associated physiological conditions, while the context of the latter depicts a mild cognitive disorder. Thus, the two previous ICD coding assignments could only be performed by considering the context in which the entities appear within the clinical cases.

In order to show the ability of the MEN transformer of our hierarchical-task approach to utilize the textual context of the medical entities to perform their normalization, in Table 9, we describe some examples of annotations having the same textual mention but different ICD codes assigned. They were correctly predicted by our transformers when following the hierarchical-task setting. Hence, exclusively considering the tokens corresponding to each clinical entity is not sufficient to correctly normalize the type of entities described in Table 9,

Table 9
Examples of explainable clinical-coding annotations with the same textual mentions but different ICD codes assigned, that were correctly predicted by transformer-based models when following the hierarchical-task approach. For each normalized entity, its textual mention, part of the context in which it appears within the text and the assigned code is described. The first pair of annotations were correctly predicted by the BETO-Galén model from the test subset of the CodiEsp-X-D corpus, while the remaining annotations were correctly predicted by the mBERT-Galén model from the test split of the CodiEsp-X-P corpus.

Textual mention	Context	Code
right hypochondrium pain	… service for *right hypochondrium pain* for 6 months, accompanied by changes in bowel habits …	R10.11 ("right upper quadrant pain")
	… he developed persistent *right hypochondrium pain*, nausea and vomiting …	R10.31 ("right lower quadrant pain")
bladder catheterization	… conservative treatment was decided with *bladder catheterization* and flushing circuit with saline …	0T9B ("drainage of bladder")
	… bladder balloon obstruction in both renal fossae. Due to the suspicion of acute urinary retention, a *bladder catheterization* was performed, confirming a voiding cyst of 1,200 cc …	0T9B70Z ("drainage of bladder with drainage device, via natural or artificial opening")
cesarean	… treated with acetylsalicylic acid at a dose of 200 mg/day, with normal pregnancy and *cesarean* delivery …	10D0 ("extraction of products of conception")
	… pregnancy with normal course and no family history of interest. Elective *cesarean* delivery by breech presentation …	10D00Z1 ("extraction of products of conception, low, open approach")
hemodialysis	… was managed with hydroxyurea and *hemodialysis*. Biopsy of the renal graft was performed on the fifth day …	5A1D ("performance of urinary procedure")
	… returning to *hemodialysis* in November 2002 after the development of a glomerulonephritis nonproliferative secondary to HCV on renal graft …	5A1D00Z ("performance of urinary filtration, single")

and only by also attending to the context in which the entities appear within the clinical cases, their normalization can be accomplished.

Finally, we must also state that, although the developed methodology in this work was exclusively applied to tackle the problem of explainable clinical-coding, the exact same methodology can also be applied to address other prediction tasks from the clinical domain that involve both the detection and normalization of clinical entities. For instance, medical concept normalization [42,43]—a task in which the context where the clinical entities appear within the text can affect the results of normalization [55]—may potentially benefit from the context-aware normalization of the medical entities performed by our hierarchical-task approach. Additionally, the proposed methodology is language-independent. Consequently, it can also be applied to medical documents written in other languages distinct from Spanish, by straightforwardly substituting any transformer-based models supporting a specific language for the particular transformer-based models employed in this study (see Section 4.2).

7. Conclusion

In this work, we systematically examine the performance of transformers for explainable clinical-coding. Particularly, we compare the performance obtained by the general-domain version of 3 different transformer-based models with the results achieved by the clinical-domain version of the models obtained by further pretraining the architectures on a collection of real-world clinical cases, with the goal of adapting transformers to the specificities of the medical domain. We address the explainable clinical-coding problem as a dual mER-MEN task, in which each clinical entity has to be both detected and assigned a particular ICD code. With the intention of tackling explainable clinical-coding using transformers, we have developed two different approaches, a multi-task and a hierarchical-task strategy. For the 3 tasks considered in this study, the clinical-domain version of the models significantly outperforms the general-domain models. Moreover, the transformer-based models achieve a significantly higher performance when following the hierarchical-task approach than by following the multi-task strategy. In particular, in combination with an ensemble approach that leverages the predictive capabilities of the different models, the transformers following the hierarchical-task strategy set new SOTA performances for both the Cantemist-Norm [14] and CodiEsp-X [13] explainable clinical-coding tasks. By further examining the differences in performance between the multi-task and the hierarchical-task strategies, we identify-two critical features of the hierarchical-task approach that lead to the observed increase in performance, namely to separately perform the MER and MEN subtasks and the context-aware text-classification approach used to tackle the MEN task. Both features contribute to reducing the intrinsic complexity of the analyzed tasks. Finally, the proposed methodology can also be applied to address other prediction tasks from the clinical domain involving both the detection and normalization of clinical entities using transformer-based models.

Statement of Significance

Problem: Traditionally, clinical coding of Electronic Health Records (EHRs) in hospitals has been carried out manually. However, clinical coding can be automated, improving many medical and productivity aspects of the health professionals involved.

What is Already Known: Although many works have already tackled the problem of automatic clinical coding, most of the existing computer-based methods act as “black boxes”, which greatly limits their applicability to real-world clinical scenarios.

What This Paper Adds: This study aims to develop two different methodologies to effectively apply transformer-based models to the problem of explainable clinical coding, requiring the models to perform the assignments of clinical codes to medical cases, but also to provide the reference in the text that justifies each coding assignment. We demonstrate that our proposed hierarchical-task approach leads in-domain transformers to establish new state-of-the-art (SOTA) performances for three distinct explainable clinical-coding tasks. Additionally, the developed methodology can be potentially applied to address other clinical tasks that require both the recognition and normalization of clinical entities.

CRediT authorship contribution statement

Guillermo López-García: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **José M. Jerez:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Nuria Ribelles:** Funding acquisition, Resources, Writing – review & editing. **Emilio Alba:** Funding acquisition, Resources, Writing – review & editing. **Francisco J. Veredas:** Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Ministerio de Ciencia e Innovación, under Project PID2020-116898RB-I00, in part by the Ministerio de Economía y Empresa (MINECO), Plan Nacional de I + D + I, under Project TIN2017-88728-C2-1-R, and in part by the Junta de Andalucía, under Project PYC20-RE-046 UMA. Funding for open access charge: Universidad de Málaga / CBUA. The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Málaga.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104323.

References

[1] L.A. Baumann, J. Baker, A.G. Elshaug, The impact of electronic health record systems on clinical documentation times: A systematic review, Health Policy. 122 (2018) 827–836.
[2] J. Bronnert, Preparing for the CAC transition, J. AHIMA. 82 (2011) 60–1; quiz 62.
[3] G. Mujtaba, L. Shuib, N. Idris, W.L. Hoo, R.G. Raj, K. Khowaja, K. Shaikh, H. F. Nweke, Clinical text classification research trends: Systematic literature review and open issues, Expert Syst. Appl. 116 (2019) 494–520.
[4] A.E.W. Johnson, T.J. Pollard, L. Shen, L.-W.-H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci Data. 3 (2016), 160035.
[5] L. Virginio, J.C. dos Reis, Automated Coding of Medical Diagnostics from Free-Text: The Role of Parameters Optimization and Imbalanced Classes, in: Data integration in the Life Sciences, Springer International Publishing, 2019, pp. 122–134.
[6] M.H. Stanfill, M. Williams, S.H. Fenton, R.A. Jenders, W.R. Hersh, A systematic literature review of automated clinical coding and classification systems, J. Am. Med. Inform. Assoc. 17 (2010) 646–651.
[7] E. Tjoa, C. Guan, A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI, IEEE Trans Neural Netw Learn Syst. 32 (2021) 4793–4813.
[8] D. Gunning, D. Aha, DARPA's explainable artificial intelligence (XAI) program, AI Mag. 40 (2019) 44–58.
[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł.U. Kaiser, I. Polosukhin, Attention is All you Need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates Inc, 2017, pp. 5998–6008.

[10] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in, Association for Computational Linguistics, Stroudsburg, PA, USA, 2020.
[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv [cs.CL]. (2018). http://arxiv.org/abs/1810.04805.
[12] J. Canete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, Pml4dc at Iclr. (2020, 2020,) 2020.
[13] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, M. Krallinger, Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF eHealth 2020, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2020. http://ceur-ws.org/Vol-2696/paper_263.pdf.
[14] A. Miranda-Escalada, E. Farré-Maduell, M. Krallinger, Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results, in: M.Á. García Cumbreras, J. Gonzalo, E. Martínez Cámara, R. Martínez Unanue, P. Rosso, S. Jiménez Zafra, J.A. Ortiz-Zambrano, A. Miranda, J. Porta-Zamorano, Y. Guitiérrez, M. Rosá Aiala amd Montes-y-Gómez, M. García-Vega (Eds.), Iberian Languages Evaluation Forum (IberLEF 2020), Málaga, Spain, 2020: pp. 303–323.
[15] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, Stroudsburg, PA, USA, 2018. https://doi.org/10.18653/v1/n18-1202.
[16] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Stroudsburg, PA, USA, 2018. https://doi.org/10.18653/v1/p18-1031.
[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, arXiv [cs.LG]. (2020). http://arxiv.org/abs/1910.10683.
[18] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q.V. Le, in: XLNet: Generalized Autoregressive Pretraining for Language Understanding, in, Curran Associates Inc., Red Hook, NY, USA, 2019, pp. 5753–5763.
[19] M. Zeng, M. Li, Z. Fei, Y. Yu, Y. Pan, J. Wang, Automatic ICD-9 coding via deep transfer learning, Neurocomputing. 324 (2019) 43–50.
[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv [cs.CL]. (2019). http://arxiv.org/abs/1907.11692.
[21] G. López-García, J.M. Jerez, N. Ribelles, E. Alba, F.J. Veredas, Detection of Tumor Morphology Mentions in Clinical Reports in Spanish Using Transformers, in: I. Rojas, G. Joya, A. Catala (Eds.), Advances in Computational Intelligence, Springer International Publishing, Cham, 2021, pp. 24–35.
[22] M. Arkhipov, M. Trofimova, Y. Kuratov, A. Sorokin, Tuning multilingual transformers for language-specific named entity recognition, in, in: Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, 2019, pp. 89–93.
[23] G. Lopez-Garcia, J.M. Jerez, N. Ribelles, E. Alba, F.J. Veredas, Transformers for Clinical Coding in Spanish, IEEE Access. 9 (2021) 72387–72397, https://doi.org/10.1109/access.2021.3080085.
[24] J.P. Pestian, C. Brew, P. Matykiewicz, D.J. Hovermale, N. Johnson, K.B. Cohen, W. Duch, A shared task involving multi-label classification of clinical free text, in: Biological, Translational, and Clinical Language Processing, Prague, Czech Republic, Association for Computational Linguistics, 2007, pp. 97–104.
[25] S. Karimi, X. Dai, H. Hassanzadeh, A. Nguyen, Automatic diagnosis coding of radiology reports: a comparison of deep learning and conventional classification methods, in, BioNLP (2017, 2017,) 328–332.
[26] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein, Explainable Prediction of Medical Codes from Clinical Text, arXiv [cs.CL]. (2018). http://arxiv.org/abs/1802.05695.
[27] O. Trigueros, A. Blanco, N. Lebeña, A. Casillas, A. Pérez, Explainable ICD multi-label classification of EHRs in Spanish with convolutional attention, Int. J. Med. Inform. 157 (2021), 104615.
[28] A. Blanco, A. Casillas, A. Pérez, A. Diaz de Ilarraza, Multi-label clinical document classification: Impact of label-density, Expert Syst. Appl. 138 (2019), 112835.
[29] J. Pérez, A. Pérez, A. Casillas, K. Gojenola, Cardiology record multi-label classification using latent Dirichlet allocation, Comput. Methods Programs Biomed. 164 (2018) 111–119.
[30] M. Almagro, R.M. Unanue, V. Fresno, S. Montalvo, ICD-10 Coding of Spanish Electronic Discharge Summaries: An Extreme Classification Problem, IEEE Access. 8 (2020) 100073–100083, https://doi.org/10.1109/access.2020.2997241.
[31] G. López-Garcıa, J.M. Jerez, F.J. Veredas, ICB-UMA at CLEF e-Health 2020 Task 1: Automatic ICD-10 coding in Spanish with BERT, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2020. http://ceur-ws.org/Vol-2696/paper_101.pdf.
[32] G. López-García, J.M. Jerez, N. Ribelles, E. Alba, F.J. Veredas, ICB-UMA at CANTEMIST 2020: Automatic ICD-O Coding in Spanish with BERT, in: M.Á.G. Cumbreras, J. Gonzalo, E.M. Cámara, R.M. Unanue, P. Rosso, S.J. Zafra, J.A. Ortiz-Zambrano, A. Miranda, J. Porta-Zamorano, Y. Guitiérrez, A. Rosá, M. Montes-y-Gómez, M. García-Vega (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020: pp. 468–476.
[33] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, N. Elhadad, Multi-Label Classification of Patient Notes a Case Study on ICD Code Assignment, arXiv [cs.CL]. (2017). http://arxiv.org/abs/1709.09587.
[34] H. Dong, V. Suárez-Paniagua, W. Whiteley, H. Wu, Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation, J. Biomed. Inform. 116 (2021), 103728.
[35] B. Biswas, T.-H. Pham, P. Zhang, TransICD: Transformer Based Code-Wise Attention Model for Explainable ICD Coding, in: Artificial intelligence in Medicine, Springer International Publishing, 2021, pp. 469–478.
[36] F. Teng, W. Yang, L. Chen, L. Huang, Q. Xu, Explainable Prediction of Medical Codes With Knowledge Graphs, Front Bioeng Biotechnol. 8 (2020) 867.
[37] N. García-Santa, K. Cetina, FLE at CLEF eHealth 2020: Text Mining and Semantic Knowledge for Automated Clinical Encoding, in: CLEF (Working Notes), 2020. http://ceur-ws.org/Vol-2696/paper_111.pdf.
[38] Y. Xiong, Y. Huang, Q. Chen, X. Wang, Y. Nic, B. Tang, A Joint Model for Medical Named Entity Recognition and Normalization, in: M.Á. García Cumbreras, J. Gonzalo, E. Martínez Cámara, R. Martínez Unanue, P. Rosso, S. Jiménez Zafra, J.A. Ortiz-Zambrano, A. Miranda, J. Porta-Zamorano, Y. Guitiérrez, M. Rosá Aiala amd Montes-y-Gómez, M. García-Vega (Eds.), Iberian Languages Evaluation Forum (IberLEF 2020), Málaga, Spain, 2020: pp. 499–504.
[39] S. Zhao, T. Liu, S. Zhao, F. Wang, A Neural Multi-Task Learning Framework to Jointly Model Medical Named Entity Recognition and Normalization, AAAI. 33 (2019) 817–824.
[40] B. Zhou, X. Cai, Y. Zhang, X. Yuan, An End-to-End Progressive Multi-Task Learning Framework for Medical Named Entity Recognition and Normalization, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021: pp. 6214–6224.
[41] F. Li, Y. Jin, W. Liu, B.P.S. Rawat, P. Cai, H. Yu, Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)–Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study, JMIR Medical Informatics. 7 (2019) e14830.
[42] Q. Wang, Z. Ji, J. Wang, S. Wu, W. Lin, W. Li, L. Ke, G. Xiao, Q. Jiang, H. Xu, Y. Zhou, A study of entity-linking methods for normalizing Chinese diagnosis and procedure terms to ICD codes, J. Biomed. Inform. 105 (2020), 103418.
[43] Z. Ji, Q. Wei, H. Xu, BERT-based Ranking for Biomedical Entity Normalization, AMIA Jt Summits Transl Sci Proc. 2020 (2020) 269–277.
[44] E. Zotova, M. Cuadros, G. Rigau, ClinIDMap: Towards a clinical IDs mapping for data interoperability, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022: pp. 3661–3669.
[45] N. Ribelles, J.M. Jerez, D. Urda, J.L. Subirats, A. Márquez, C. Quero, L. Franco, Galén: Sistema de Información para la gestión y coordinación de procesos en un servicio de Oncología, RevistaeSalud. 6 (2010) 1–12.
[46] A. Blanco, A. Perez, A. Casillas, Exploiting ICD Hierarchy for Classification of EHRs in Spanish through multi-task Transformers, IEEE J Biomed Health Inform. PP (2021). https://doi.org/10.1109/JBHI.2021.3112130.
[47] Centers for Medicare and Medicaid Services (CMS), International Classification of Diseases, Tenth Revision, Procedure Coding System (ICD-10-PCS), 2021. https://www.cms.gov/medicare/coding/icd10 (accessed March 1, 2022).
[48] L.A. Ramshaw, M.P. Marcus, Text Chunking Using Transformation-Based Learning, in: S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, D. Yarowsky (Eds.), Natural Language Processing Using Very Large Corpora, Springer, Netherlands, Dordrecht, 1999, pp. 157–176.
[49] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, J. Han, On the Variance of the Adaptive Learning Rate and Beyond, arXiv [cs.LG]. (2019). http://arxiv.org/abs/1908.03265.
[50] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[51] E. Tutubalina, A. Kadurin, Z. Miftahutdinov, Fair evaluation in concept normalization: A large-scale comparative analysis for BERT-based models, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Stroudsburg, PA, USA, 2020: pp. 6710–6716.
[52] H. Kim, J. Kang, How Do Your Biomedical Named Entity Recognition Models Generalize to Novel Entities? IEEE Access. 10 (2022) 31513–31523.
[53] A. Alekseev, Z. Miftahutdinov, E. Tutubalina, A. Shelmanov, V. Ivanov, V. Kokh, A. Nesterov, M. Avetisian, A. Chertok, S. Nikolenko, Medical Crossing: a Cross-lingual Evaluation of Clinical Entity Linking, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022: pp. 4212–4220.
[54] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-Alignment Pretraining for Biomedical Entity Representations, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021: pp. 4228–4238.
[55] Y.-F. Luo, W. Sun, A. Rumshisky, MCN: A comprehensive corpus for medical concept normalization, J. Biomed. Inform. 92 (2019), 103132.