applied sciences

MDPI

Article

Using Enhanced Representations to Predict Medical Procedures from Clinician Notes

Roberto Móstoles, Oscar Araque and Carlos Á. Iglesias *

Intelligent Systems Group, ETSI Telecomunicación, Universidad Politécnica de Madrid, Avda. Complutense 30, 28040 Madrid, Spain; ro.mostoles@upm.es (R.M.); o.araque@upm.es (O.A.)
* Correspondence: carlosangel.iglesias@upm.es; Tel.: +34-910671900

Abstract: Nowadays, most health professionals use electronic health records to keep track of patients. To properly use and share these data, the community has relied on medical classification standards to represent patient information. However, the coding process is tedious and time-consuming, often limiting its application. This paper proposes a novel feature representation method that considers the distinction between diagnoses and procedure codes, and applies this to the task of medical procedure code prediction. Diagnosis codes are combined with text annotations, and the result is then used as input to a downstream procedure code prediction task. Various diagnosis code representations are considered by exploiting a code hierarchy. Furthermore, different text representation strategies are also used, including embeddings from language models. Finally, the method was evaluated using the MIMIC-III database. Our experiments showed improved performance in procedure code prediction when exploiting the diagnosis codes, outperforming state-of-the-art models.

Keywords: healthcare; ICD prediction; deep learning; NLP; EHR; BERT; embeddings

1. Introduction

Critical reasoning is one of the most valuable skills for a healthcare professional when faced with a new medical case. A comprehensive view of the patient's history allows for a better assessment of their condition. Thus, having quick access to all relevant information about a patient promotes better, faster, and more cost-effective decision-making [1]. This idea is reflected in the upward trend in Electronic Health Records (EHR) adoption worldwide in recent years, which seeks to increase efficiency and improve quality of care by giving care providers access to comprehensive and accurate patient records [2]. However, as vast amounts of patient data become available, there is a growing motivation in the community to develop strategies to optimize the use of this type of information.

Currently, an essential source of information usually included in EHRs corresponds to unstructured data in the form of clinical texts acquired during medical checks and hospital admissions [3]. These annotations contain valuable information about each patient, but making practical use of this information is not trivial, due to the peculiarities of the texts, such as the presence of misspellings, typos, and ambiguous terms [4]. To deal with this lack of structure, clinicians assign specific medical codes to known conditions and events from their observations. To this end, medical professionals use various coding standards for different applications. Healthcare Common Procedure Coding System (HCPCS) [5], Current Procedural Terminology (CPT) [6], and Australian Classification of Health Interventions (ACHI) [7], for example, provide a collection of standardized codes for medical procedures, supplies, services, and interventions.

One of the most used standards is the International Classification of Disease (ICD) [8], developed and maintained by the World Health Organization (WHO) and recognized as one of the main bases for comparing health statistics related to morbidities and causes of death. The standard is periodically updated to reflect advances in medical knowledge and involves healthcare professionals worldwide [9]. In 2024, the standard spans 11 revisions,

check for updates

Citation: Móstoles, R.; Araque, O.; Iglesias, C.Á. Using Enhanced Representations to Predict Medical Procedures from Clinician Notes. *Appl. Sci.* **2024**, *14*, 6431. https://doi.org/10.3390/app14156431

Academic Editor: Luigi Portinale

Received: 19 June 2024
Revised: 11 July 2024
Accepted: 22 July 2024
Published: 24 July 2024

the latter of which, ICD-11, introduces a computable knowledge framework consisting of a large ontology of around 80,000 medical entities, as presented by Harrison et al. [8]. Each entity describes a health-related phenomenon and establishes relationships with other entities defined in the standard.

Given its wide adoption, this study is built around the ICD standard. To evaluate our results with comparable approaches to ICD procedure prediction in the literature [10,11], we used the Medical Information Mart for Intensive Care III (MIMIC-III) database [12], which is based on the ninth revision of the standard (ICD-9). The ICD-9 revision comprises a taxonomy of more than 15,000 codes for disease diagnoses and procedures.

Manually assigning medical codes still requires a medical expert with good knowledge to make the associations. This task is time-consuming and, most importantly, error-prone, particularly due to the large size of the ICD taxonomy [10] and the hindsight bias that can lead to incomplete or downright wrong clinical evaluations [13]. This coding standard is also widely used to assess medical costs and support billing processes with insurance companies, and consequently improper usage can have monetary consequences for institutions [14]. To mitigate these limitations, automatic ICD coding has been explored for decades using traditional Machine Learning (ML) tools [15–17] and, more recently, incorporating Deep Learning (DL) elements into the mix [3,10,18–20]. For an extended overview of coding schemes, relevant databases, and State-of-the-Art (SOTA) solutions for automatic ICD coding, the interested reader will find a wealth of information in the extensive review by Kaur et al. [21].

The ICD-9 coding standard integrates codes for medical diagnoses and procedures, which is a clinically relevant distinction, since they refer to different stages of a patient's care process. Following this observation, we propose a novel approach to the problem of automatic ICD coding by assuming that latent relationships between groups of diagnoses and procedures help identify the latter when there is some information about the former. We base our premise on the hypothesis that similar medical conditions often require comparable treatments, which is a relation that has been studied in the past [22,23] and still raises interest in various domains at the present time [24,25].

This paper proposes a novel feature representation method that uses information on known diagnoses combined with textual clinical annotations and uses both as input for the task of ICD procedure code prediction. To analyze how incorporating diagnosis codes affects the performance of the prediction task, we consider varying levels of detail in the definition of diagnoses by exploiting the hierarchical structure of ICD codes. Additionally, this paper tested this novel representation method using different strategies for transforming text annotations into structured vectors, including embeddings from language models. Finally, we evaluate the study using the discharge summaries of the MIMIC-III database and compare our results with other relevant studies of the SOTA.

The remainder of the paper is structured as follows. In Section 2, we present the most relevant state of the art on automatic ICD code prediction. An overview of the proposed architecture is presented in Section 3. The test methodology and experimental results are presented in Section 4. Finally, in Section 5, we summarize the conclusions of this work and discuss future lines of work.

2. Background

When applying ML-based strategies, automatic ICD coding is formulated as a multi-label classification problem. A wealth of work has explored this topic using different ML-based approaches, some using conventional ML models and others integrating DL elements into their solutions.

Perotte et al. [15] developed two prediction models based on Support Vector Machine (SVM), a flat classifier where each ICD code is predicted independently of any other, and another that exploits the hierarchical nature of the ICD codes, concluding that the latter yielded better coding results. Koopman et al. [16] applied an SVM-based solution to the specific task of classifying types of cancer from death certificates, allowing them to

reduce the label space to predict. Kavuluru et al. [17] also resorted to conventional ML techniques, evaluating the performance of various classification models based on SVM, Logistic Regression (LR), and Multinomial Naive Bayes (MNB), in the task of assigning ICD codes to extensive medical records. Other works have focused on the hierarchical structure of codes, such as in the work of Subotin and Davis [26], where they proposed a two-level hierarchical classification, a partial classification to first identify potential codes, followed by a second estimator that resolves the detailed codes.

With the appearance of DL, many strategies have been developed in recent years to improve the results of traditional ML approaches. Mullenbach et al., Nuthakki et al., Reys et al., and Huang et al. [3,10,11] proposed different DL-based approaches to automatic ICD coding and compared their results against conventional ML alternatives, collectively showing better results for the DL models. Mulllenbach et al. [10] developed a Convolutional Neural Network (CNN) model introducing per-label attention mechanisms, which boosted explainability by selecting the most relevant segments of the clinical notes. Nuthakki et al. [11] used transfer learning to adapt a Long Short-Term Memory (LSTM)-based architecture to predict diagnosis and procedure ICD codes. Reys et al. [18] adapted an attention-based CNN and applied it to medical records in Brazilian Portuguese, demonstrating good performance in non-English applications. Huang et al. [3] experimented with models based on LSTM and Gated Recurrent Unit (GRU), to capture long-term dependencies within the notes, some of which proved too long to retain useful information. They also considered a code hierarchy by simplifying the prediction task from specific codes to whole categories, which showed a significant improvement in accuracy at the expense of the level of detail of the final prediction.

More recently, other techniques based on transformer architectures have shown promising results using text representations that exploit context. Chen et al. [19] used word embeddings from different Bidirectional Encoder Representations from Transformers (BERT)-based models in a GRU classifier, showing improved performance compared to other embedding techniques. On the other hand, Pascual et al. [20] used a BERT-based model for automatic ICD coding and noted the limitations of the model when used with long pieces of text.

However, despite the benefits of using transformer-based architectures in Natural Language Processing (NLP), these do not currently hold the SOTA in ICD prediction. As noted in [20], CNN and Recurrent Neural Networkt (RNN) alternatives can process text of any length, while their BERT-based counterparts are inherently limited in the number of words they can evaluate. In [27], the authors arrived at a similar conclusion, stating that one of the main limitations of the transformer model in this area is its limited ability to handle longer text sequences compared to other alternatives. Still, there are positives to the transformer approach, such as its ability to handle long-range dependencies or its improved interpretability compared to traditional CNNs, which may prove helpful in some instances.

There has been a limited amount of research that specifically addressed the task of predicting ICD procedure codes compared to those that focused on predicting diagnosis codes. However, there are still studies that aimed at procedure prediction, such as the already presented study of Mullenbach et al. [10], where they considered the task of predicting diagnoses and procedures separately, or the work of Subotin et al. [26], which specifically focused on procedure prediction.

One of the premises of our study was to use some of the information already known about a patient's status to help predict appropriate treatments, i.e., using what we know about their diagnoses to predict relevant procedures. Some studies have explored label correlation in automatic ICD coding. Tsai et al. [28] developed a re-ranking module for ICD prediction by nesting a base estimator with a distribution estimator of the code set that captures the knowledge of the correlation between code groups. Yan et al. [29] proposed a multi-label large-margin classifier incorporating knowledge about code relationships, which showed an improvement in classification performance. Other studies resorted to graph-based architectures to integrate code co-occurrence. Mahdi et al. [30] explored

relationships between ICD codes by integrating a one-layer graph-based model with a code co-occurrence graph.

Lastly, some authors have also considered the explicit relationship between diagnoses and procedures in the task of ICD prediction, such as Haq et al. [31], where they developed an architecture to learn mappings between ICD diagnosis codes and CPT procedures. They exclusively considered mappings between diagnoses and procedures, without considering other sources of information, but their results showed a clear improvement over alternatives that did not exploit this relationship.

The correlation of codes has previously been studied in the context of ICD coding. However, the work in this area has either focused on analyzing the relation between codes that tend to appear together, without considering the clinical differences between diagnoses and procedures, or has not considered aggregating this information with other patient data.

3. Proposal

We propose an approach for predicting medical procedures from clinical notes by combining these with their known diagnoses as ICD codes. To our knowledge, no other study has exploited this relationship in predicting clinical procedures. Since procedures and diagnoses are often described by following the ICD coding standard, it is common in the SOTA to treat the joint prediction task of both groups, ignoring the clinical importance of each individually. Our method revolves around the premise that some diagnoses often trigger specific medical treatments; therefore, incorporating the information about this underlying relationship together with the clinical annotations can, in turn, produce better results in the task of procedure prediction compared to using either source of information alone.

The proposed process depicted in Figure 1 comprises three main stages: diagnosis ICD code detection, diagnosis ICD code expansion, and procedure ICD code detection.

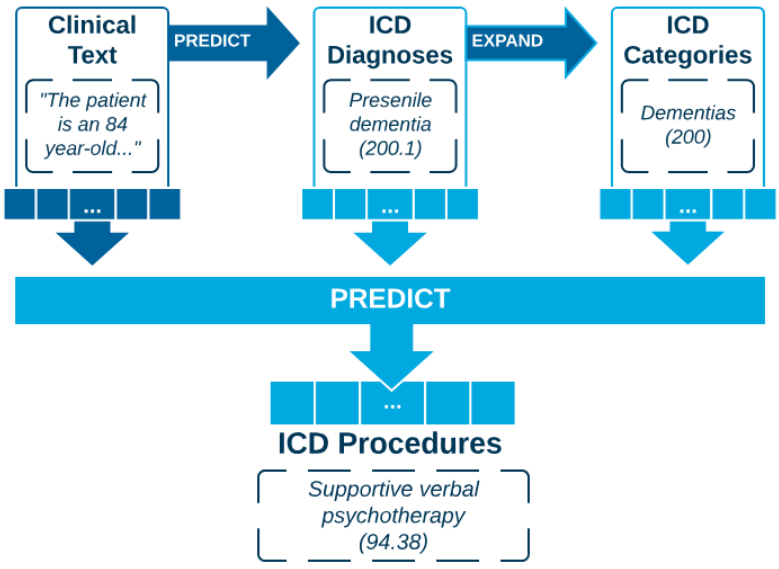

Figure 1. Approach to automatic procedure prediction combining text and diagnosis information.

The first step consists of automatically detecting ICD codes based on textual analysis of medical notes. Usually, this is the only task considered when detecting ICD codes. Our solution innovates by proposing an extended process.

The second step is optional and consists of the expansion of the codes detected in the previous step. This expansion consists of adding ICD codes from the hierarchy. For example, under the ICD-9 standard, the diagnosis "presenile dementia with delirium" belongs to the general category of "dementias", which are defined with the codes 290.11 and 290, respectively. During this step, we resolve and combine both definitions to use them later in

the procedure detection step. Some authors have also exploited the taxonomic structure of ICD codes to calculate patient similarity [32]. Our work proposes an expansion mechanism to improve the procedure of ICD code detection, which, in a nutshell, might improve the overall performance when predicting procedures by enriching the information about known diagnoses.

Finally, the third step is the detection of procedure ICD codes, taking the diagnosis ICD codes (potentially expanded) and the original text as input. This step aims to provide a new perspective for procedure ICD code prediction by considering it as a first-class task rather than a task included in general ICD code prediction. This consideration is also supported by the fact that ICD-10, a newer revision of the standard, already defines different classification schemas for diagnosis codes (ICD-10-CM) and procedure codes (ICD-10-PCS) [33].

Our experiment tested our assumptions regarding the hierarchy of ICD codes and treated procedure prediction as a separate task.

4. Implementation Design and Evaluation

Figure 2 provides a general overview of the implementation, which consists of four steps: data extraction from the MIMIC-III database, including diagnosis and procedure codes and medical annotations; text vectorization; diagnosis codification; and finally, procedure prediction using a combination of text vectors and diagnosis codes as input.

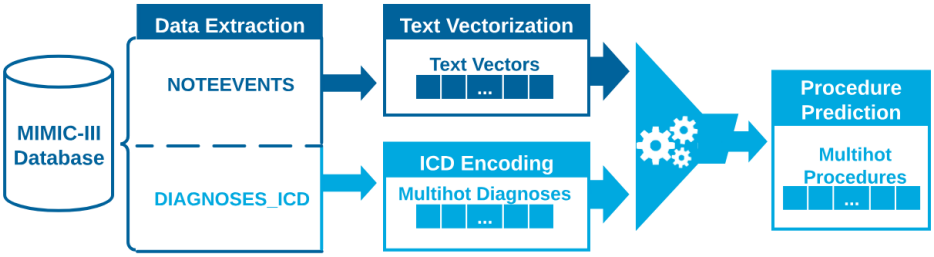

Figure 2. Implementation design overview. Diagnosis and procedure codes are described as multi-hot encodings.

The rest of this section details the implementation design and presents relevant materials, metrics, and results.

4.1. Data

To evaluate our approach, we used MIMIC-III [12], a publicly available database comprising a wealth of health-related data, including demographics, caregiver annotations, and related procedures. A complete description of the database can be consulted at [34]. For the present study, we used the text annotations in MIMIC-III associated with each medical record, specifically the discharge summaries, which contain information on admission and its cause, hospital stay, and relevant discharge reports. Each record includes a collection of known diagnoses and procedures described using the ICD-9 standard. In total, MIMIC-III contains 59,652 discharge summaries of 52,726 hospital admissions and 41,127 unique patients. The text annotations present an average of 1551 words per document, as shown in Figure 3, which shows a word count histogram across MIMIC-III's discharge summaries.

To support this study, we used the NOTEEVENTS, DIAGNOSES_ICD, and PROCEDURES_ICD tables, which contain text notes, diagnosis codes, and procedure codes, respectively. Each table describes each medical case with two unique identifiers: SUBJECT_ID for patients and HADM_ID for specific hospital admissions.

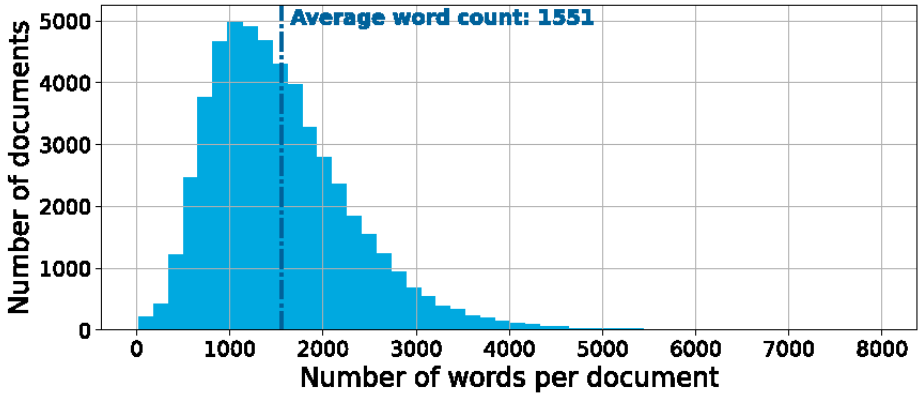

Figure 3. Word count distribution in MIMIC-III discharge summaries.

We used the same data splits as proposed by Mullenbach et al. [10], with a distribution of 47,719 records in the training set, 3372 in the test, and 1631 in validation. Entries without text annotations, diagnosis, or procedure codes were discarded. We performed minimal text preprocessing to remove special characters and any de-identified patient information, which is presented in the annotations using the special decorators "[** **]". In addition, explicit mentions of procedure codes or their names were removed from the text annotations to avoid bias.

In our study, we used the training set to fine-tune the language models used to generate the embeddings, and to train the downstream ML prediction models. The test set was used to tune the hyperparameters of the ML models. Finally, we evaluated the entire system's performance with the validation set.

4.2. Vectorized Annotations

Converting text into numeric vectors is essential in any ML task using unstructured text as input. There are multiple strategies for generating vectorized representations out of raw text, and in our approach, we cover two different methods: Term Frequency—Inverse Document Frequency (TF-IDF) features as a representation based on word importance and contextualized embeddings from transformers.

TF-IDF features encode the importance of a set of words in a specific document, which can be helpful in text classification [35]. We established a vocabulary size of 10,000 words, and fitted the implementation with the training set.

When using TF-IDF features, we extend the preprocessing outlined in Section 4.1. This process involves removing the stop words using the default English stop words from Natural Language Toolkit (NLTK) [36], then converting all words to lowercase, and finally removing words containing no alphabetic characters. This pre-processing is applied during training, and the resulting model is used to generate the vectorized representations.

Contextual embeddings from transformers are dense vector representations of words that are dynamic and influenced by the surrounding words [37]. Other works have previously used contextual embeddings in text classification tasks as a way to generate powerful representations to use in ML and DL algorithms [38–40], with some works reporting better results compared to traditional word embeddings, as shown in the study of Chanda et al. [41]. In our specific use case, we considered two different BERT-based models to extract contextualized embeddings, including DistilBERT [42] as a reference model trained on a non-specialized English corpus and ClinicalBERT [43]. This model was explicitly trained on notes from the MIMIC-III database.

We performed full fine-tuning of both models for the task of ICD procedure prediction using the texts in the training dataset, which, in our use case, showed significantly better results compared to only adjusting the top layers. We used a sequence length of 512 tokens, the maximum supported by both models, with padding and right-side truncation. Finally, we extracted the embedding from the last hidden state of the "[CLS]" token. The output of

the "[CLS]" is inferred by the rest of the words in the sentence, which is considered to be a good representation of the contents of the whole text. However, active research is still trying to find the optimal representation from contextual embeddings [44].

The vectorized representations obtained in this step are combined with the diagnosis codes to train a downstream classifier for the procedure prediction task.

4.3. ICD Encoding

Since each medical record usually has multiple ICD codes assigned to it, both for diagnoses and procedures, we use a multi-hot representation to describe the collection of codes related to any given case, with a single vector where each position relates to a specific code.

We perform set expansion for the diagnoses codes by considering the hierarchical structure of ICD codes, which first groups whole families of codes with similar characteristics and then, within each subsequent category in the hierarchy, describes specific codes with increasing detail. In the ICD-9 standard, the leading digits of a diagnosis code represent its main category, and the following encode its clinical details, as shown in Figure 4. To obtain the general category of any given code, only the leftmost part of its detailed code is taken. As a result, we consider three representations for each diagnosis code in the evaluation: its main category, the fully qualified code if available, and the expanded representation that consists of a combination of both, each represented as multi-hot vectors.

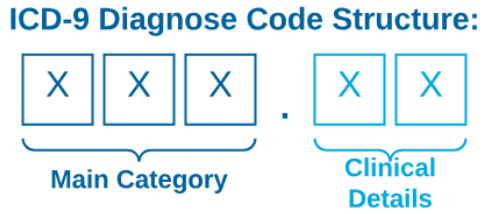

Figure 4. ICD diagnosis code hierarchy.

Another consideration is the frequency with which any code is assigned to a medical record. Some codes are more frequent and therefore likely to appear in more medical records compared to others that are assigned sparingly. Following this idea, some studies approached the task of automatic ICD coding by selecting a reduced subset of the most frequent codes seen across all medical records in their database of choice, suggesting that these codes cover most cases [3]. Using a reduced ICD code space ultimately improves prediction performance by omitting uncommon codes.

For a better comparison with the SOTA, we considered two subsets of ICD codes covering the 10 and 50 most frequent codes for both diagnoses and procedures, which correspond to the most popular configurations in the literature [3,10,11]. Nevertheless, it should also be noted that reducing the label space comes at a cost, as it can negatively impact the overall reliability, especially in the healthcare domain, where infrequent codes, either from diagnoses or procedures, tend to be clinically relevant, as noted by Yang et al. [45].

To visualize this phenomenon in actual data, we provide Figure 5, which shows the distribution of ICD codes (both diagnoses and procedures) per number of documents in which they appear in MIMIC-III. In the figure, the horizontal axis represents the distribution of codes for different subsets, sorted from least to most reoccurring, while the vertical axis shows the actual number of documents in which they appear. Bold-faced values represent the highest number of documents in which the most frequently reoccurring code appears. The distribution corroborates that most ICD codes are rarely assigned to many different records, while a small group of codes are the most prevalent across records. This behavior is also equally present for both diagnosis codes and procedure codes.

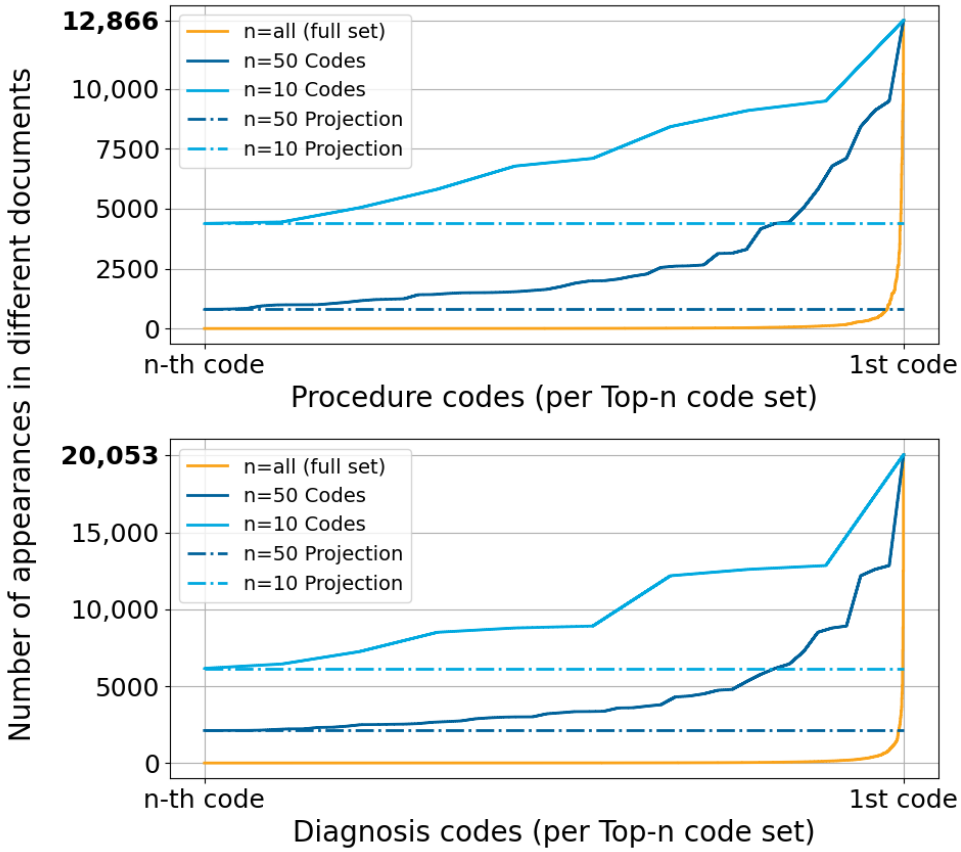

Figure 5. Histogram of ICD codes per number of documents they appear in across different code subsets in MIMIC-III.

When using subsets with the most frequent ICD codes, the representation is symmetric in terms of cardinality, which means that the number of procedure codes to predict is always the same as the number of diagnosis codes in the input. To select the codes of each subset, we first find the top diagnosis codes (*top-*n_D) for a given collection of records. We keep only the documents that contain at least one diagnosis code, and then repeat the same process to obtain the top procedure codes (*top-*n_P) using the remaining records. This process ensures that medical records always contain valid entries with at least one diagnosis and one associated procedure.

As a result of this step, we end up with a variety of representations, which leads to different inputs to consider in the prediction task, depending on the type of description used for the diagnosis code (main category, full code, or both), and the *top-*n_D and *top-*n_P subsets considered in each case. All of these were evaluated across different models in the experiments.

4.4. Prediction of Procedures

The inputs for the final prediction stage consist of different combinations of annotations and diagnosis representations, as described in Sections 4.2 and 4.3. Specifically, the collection of diagnosis codes related to each medical record, defined as multi-hot vectors, are combined with their corresponding vectorized annotations by concatenating both vectors. The resulting tensor is the input of the prediction task, and the target to predict is the collection of procedures for the medical case, also described as a multi-hot vector. The proposed experiment comprised two different classifiers in the prediction step, a simple LR model and an Multi-layer Perceptron (MLP) architecture, both implemented using Scikit-learn [46] and adjusted via grid search.

In the case of the LR classifier, the original multilabel classification problem was broken down into several binary classification problems, fitting one classifier per target, one for each procedure code to predict. During the fine-tuning step, we considered different optimizers and *L2* regularization with strengths ranging from 0.001 to 1 in multiples of 10. The adjusted LR classifier used a liblinear solver with a regularization strength of 0.1.

For the MLP fine-tuning, we considered a single hidden layer with 50, 100, 200, and 500 neurons and an *L2* regularization strength ranging from 0.0001 to 0.1 in multiples of 10. The adjusted MLP classifier used a single hidden layer with 200 neurons and an *L2* regularization strength of 0.0005. During training, we also used the Adam solver [47], adaptive learning rate to minimize training loss with an initial value of 0.005, and early stopping.

The result of this stage is the final prediction, which consists of the most relevant procedure codes for the given case represented as a multi-hot vector.

4.5. Metrics

The performance was measured by computing the *F1* score. We used micro-averaging to resolve the global performance in the multi-label prediction task, which provided a good representation of the global performance in predicting the most prevalent codes. Micro-averaged precision and recall are expressed as follows:

$$\text{Micro-Precision} = \frac{\sum_{c=1}^{C}\sum_{n=1}^{N} {y_n}^c \hat{y}_n^c}{\sum_{c=1}^{C}\sum_{n=1}^{N} \hat{y}_n^c}, \text{Micro-Recall} = \frac{\sum_{c=1}^{C}\sum_{n=1}^{N} {y_n}^c \hat{y}_n^c}{\sum_{c=1}^{C}\sum_{n=1}^{N} {y_n}^c} \quad (1)$$

where C denotes the number of classes (i.e., the number of ICD procedure codes) and N the number of samples (i.e., the number of medical records). ${y_n}^c$ and $\hat{y}_n^c$ are, respectively, the ground truth and the predicted outcome for the class c in sample n. The *F1* score is described as the harmonic mean between precision and recall.

4.6. Results and Discussion

In this section, we review the results achieved by the proposed approach to ICD procedure prediction, comparing with other works where applicable. Table 1 summarizes this evaluation, comparing the global micro *F1* score in the procedure prediction task for different *top-*n_P code subsets, models, and text features. Our models are denoted by "†", and the best overall results for each ICD code subset are denoted by "*". The considered input types are defined as T_V for text vectors and $ICD_{\{C,F\}}$ for the main categories of the diagnosis codes (C subscript) or the fully qualified codes (F subscript). Combined inputs are represented with combined symbols.

Analyzing the performance of the models, we observed that text representations (T_V) based on contextual embeddings from BERT provided slight improvements in performance over the TF-IDF alternatives. This is in itself an interesting outcome, as the TF-IDF-based representations used in the experiments were significantly higher-dimensional, with vectors of 10,000 features compared to the 768-dimensional vectors from BERT-based embeddings. Furthermore, the BERT models were limited to sequences of 512 tokens, far from the average word count per document of 1551 (Figure 3). These results suggest that there might be sequential or temporal relationships between words in the text annotations that the BERT-based embeddings were able to capture and encode into considerably smaller representations. Although a significant part of the original text was left out, the resulting vector did a better job of capturing the core semantic meaning of the whole medical annotation, using a fraction of the features.

Our results also show that there is indeed value in using diagnosis codes as part of the input ($ICD_{\{C,F\}}$), since there was a consistent boost in performance in every model when combining diagnosis codes with text vectors. Additionally, this boost increased with the level of detail of the diagnosis codes, and more interestingly, in some scenarios, it was still marginally improved when expanding the fully qualified code. We saw improvements in the *F1* score of up to 19% in our setup when using the most detailed diagnoses (ICD_{CF}).

Even when considering only the less detailed representations corresponding to the main categories (ICD_C), there was still a noticeable performance improvement compared to using only medical annotations as input. This result suggests that even with limited knowledge about a patient's diagnosis, the proposed method can potentially assist in predicting the ICD codes for the most relevant procedures.

Table 1. Comparison of global micro *F1* in procedure prediction. The proposed models are denoted by "†". The best overall results for each ICD subset are denoted by "*".

ICD Subset	Model	Text Features	Micro *F1* per Type of Input			
			T_V	$ICD_C T_V$	$ICD_F T_V$	$ICD_{CF} T_V$
Top-10	AWD-LSTM [11]	TF-IDF	0.690	-	-	-
	LR †	TF-IDF	0.549	0.665	0.670	0.675
		DistilBERT	0.561	0.647	0.667	0.670
		ClinicalBERT	0.582	0.658	0.679	0.680
	MLP †	TF-IDF	0.622	0.662	0.693	0.692
		DistilBERT	0.553	0.643	0.670	0.673
		ClinicalBERT	0.571	0.653	0.679	0.697 *
Top-50	LR [10]	Word2Vec	0.533	-	-	-
	CAML [10]	Word2Vec	0.614	-	-	-
	AWD-LSTM [11]	TF-IDF	0.480	-	-	-
	LR †	TF-IDF	0.519	0.551	0.560	0.564
		DistilBERT	0.562	0.588	0.584	0.593
		ClinicalBERT	0.592	0.614	0.620	0.623
	MLP †	TF-IDF	0.602	0.611	0.613	0.618
		DistilBERT	0.566	0.578	0.585	0.594
		ClinicalBERT	0.599	0.606	0.632 *	0.622

Compared with other works, we have to consider some caveats regarding the scope of the prediction task and how we compared our work with the SOTA. Most studies tackled the problem of ICD code prediction by considering either the diagnosis codes alone or diagnoses and procedures, but without making clear distinctions between both groups in the prediction task. This dramatically limits the studies we can reliably compare to, as predicting procedure codes is clinically and functionally different from predicting diagnoses. Each group of codes represents different clinical stages and contains a wildly different number of codes, which leads to inconsistent results when comparing the two prediction tasks, as Mullenbach et al. showed in their results [10].

Considering previous considerations, we only compared our results with studies from the SOTA that explicitly reported results for predicting ICD-9 procedures. Specifically, we considered the work of Mullenbach et al. [10], in which they proposed a model for ICD prediction, CAML, based on a CNN architecture with attention mechanisms, and Nuthakki et al. [11], where they used a fine-tuned version of an AWD-LSTM architecture.

Taking into account the best-performing models from the considered studies, the proposed method showed competitive results, outperforming the LSTM-based model in the top-10 prediction subtask with an *F1* score of 0.697 versus their 0.690, and also in the case of the top-50 subtask, showing an *F1* score of 0.632 versus the reported 0.614 of CAML. The best performing models are based on ClinicalBERT features combined with

an MLP classifier, showing marginally better results than the alternatives using an LR classifier and significantly better results when comparing ClinicalBERT with the other text representation alternatives.

The proposed method showed a significant boost in performance, even when using the least detailed diagnosis codes, as seen when combining the text vectors with only the diagnosis categories $ICD_C T_V$, where the model based on LR and embeddings from ClinicalBERT achieved comparable results to the CNN implementation from [10] with a comparable *F1* score of 0.614. When comparing similar setups, the proposed method also came out ahead. The model based on TF-IDF features and a LR classifier achieved an *F1* score of 0.564 when combining text annotations with diagnosis codes. These results outperformed those of the comparable LR with Word2Vec features from [10], which reported a score of 0.533.

The experiments showed that diagnosis codes and medical annotations can help predict relevant procedure codes, achieving competitive results that outperformed comparable methods in the SOTA. Our method also managed to achieve these results while relying on BERT-based models as a fundamental part of the text representation strategy, which, as other studies have pointed out [20,27], currently does not hold the SOTA in predicting ICD codes from medical annotations. However, resorting to solutions based on transformers could be particularly beneficial in healthcare, due to the potential insights into the decision-making process their attention mechanisms can deliver [48].

5. Conclusions and Future Work

This work proposed a novel approach to ICD procedure prediction supported by diagnoses. To achieve this, we developed a strategy to combine information about known diagnoses with textual annotations by concatenating their respective vectorized representations, and using the result as input to the prediction task. The hierarchical structure of the ICD scheme was also considered in this study. Diagnosis ICD codes are described using varying levels of detail. Specifically, we use the fully detailed code and its primary category, which captures its general characteristics. In addition, this work also performed set expansion over the resulting representations by aggregating both levels of detail into a single vector.

The evaluation was supported by F1 as the main performance metric, due to its popularity in the research community in this domain [3,10,11,49]. However, other metrics could be considered. Some authors [50,51], proposed using additional metrics that lead to more relevant results compared to the commonly used F1-based metric. Other authors [15,52] proposed metrics that evaluate the impact of mispredictions, which is especially important in healthcare, given the consequences of incorrect predictions [53,54]. An extended evaluation will be explored in the future to take these considerations into account.

The results show that incorporating the information about diagnosis produced better results in the prediction task than using medical annotations alone. The method proposed in this paper provided competitive results, outperforming comparable alternatives in the SOTA. We emphasized the use of language models to support text representation, due to their ability to generate highly relevant yet relatively small embedding vectors and because of the active research on explainability, an important quality in healthcare. Nevertheless, the proposed approach to ICD prediction is not limited to BERT-based architectures, as it can be implemented in other prediction pipelines.

One of this study's limitations is the reduced number of architectures used to test our assumptions. The work focused on validating the idea of implementing a procedure prediction task supported by diagnoses. Still, this does not give a complete picture of how the method would behave when combined with other DL architectures, such as CNNs, which currently hold the SOTA in ICD prediction. Future iterations of this work aim to address this gap by validating the results in a broader variety of architectures. Model optimization also needs to be addressed, given the community's interest in greener DL solutions [55] and the impact of computational cost in real-world use cases.

Finally, another limitation is that this work did not consider using rating scales [56,57], which are useful for patient assessment. Rating scales are especially important in the context of mental health [58], a topic that has gained relevance since the COVID-19 pandemic [59,60]. Future work aims to adopt newer revisions of the ICD standard to facilitate the analysis of these types of scales [61,62].

The representation method proposed in this article shows that it is possible to improve the performance in the ICD procedure prediction task by considering the clinical nuances between diagnosis and procedure codes. Furthermore, the approach presented in this study aims to reproduce how medical professionals draw their conclusions from observations, by establishing a diagnosis and then assigning relevant procedures to the case. Because of this, the approach can potentially improve the overall interpretability of the prediction task.

Author Contributions: Conceptualization: R.M., O.A. and C.Á.I.; methodology: R.M., O.A. and C.Á.I.; validation: R.M.; writing, review, and editing: R.M., O.A. and C.Á.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Spanish Ministry of Science and Innovation through the MIRATAR project (TED2021-132149B-C42) and by the European Union with NextGeneration EU funds.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study used data from the MIMIC-III database, accessible upon request via the PhysioNet data repository: https://physionet.org/content/mimiciii/1.4/, accessed on 21 July 2024. The code that supports this study is available at https://lab.gsi.upm.es/rmostoles/icd-prediction, accessed on 21 July 2024.

Acknowledgments: We want to thank all the MIRATAR (grant TED2021-132149B-C42 funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR) project participants and colleagues from CIBER (Centro de Investigación Biomédica en Red), UCLM (Universidad de Castilla - La Mancha), and UC3M (Universidad Carlos III de Madrid). We would also like to thank our colleague Matteo Leghissa for his feedback on the design of the illustrations.

Conflicts of Interest: The funding organization had no role in the design, validation, or writing of this study, nor did they have a deciding influence in the publication of the results.

References

1. King, J.; Patel, V.; Jamoom, E.W.; Furukawa, M.F. Clinical Benefits of Electronic Health Record Use: National Findings. *Health Serv. Res.* **2013**, *49*, 392–404. [CrossRef] [PubMed]
2. Upadhyay, S.; Hu, H.F. A Qualitative Analysis of the Impact of Electronic Health Records (EHR) on Healthcare Quality and Safety: Clinicians' Lived Experiences. *Health Serv. Insights* **2022**, *15*, 117863292110707. [CrossRef] [PubMed]
3. Huang, J.; Osorio, C.; Sy, L.W. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Comput. Methods Programs Biomed.* **2019**, *177*, 141–153. [CrossRef]
4. Baumel, T.; Nassour-Kassis, J.; Cohen, R.; Elhadad, M.; Elhadad, N. Multi-label classification of patient notes: Case study on ICD code assignment. In Proceedings of the Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
5. Nusgart, M. HCPCS Coding: An Integral Part of Your Reimbursement Strategy. *Adv. Wound Care* **2013**, *2*, 576–582. [CrossRef] [PubMed]
6. Dotson, P. CPT® Codes: What Are They, Why Are They Necessary, and How Are They Developed? *Adv. Wound Care* **2013**, *2*, 583–587. [CrossRef] [PubMed]
7. Australian Coding Standards for ICD-10-AM and ACHI (ICD-10-AM/ACHI/ACS Twelfth Edition). National Centre for Classification in Health: Sydney, Australia, 2010. Available online: https://www.ihacpa.gov.au/resources/icd-10-amachiacs-twelfth-edition (accessed on 21 July 2024).
8. Harrison, J.E.; Weber, S.; Jakob, R.; Chute, C.G. ICD-11: An international classification of diseases for the twenty-first century. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 206. [CrossRef] [PubMed]
9. Reed, G.M.; First, M.B.; Kogan, C.S.; Hyman, S.E.; Gureje, O.; Gaebel, W.; Maj, M.; Stein, D.J.; Maercker, A.; Tyrer, P.; et al. Innovations and changes in the ICD-11 classification of mental, behavioural and neurodevelopmental disorders. *World Psychiatry* **2019**, *18*, 3–19. [CrossRef] [PubMed]

10. Mullenbach, J.; Wiegreffe, S.; Duke, J.; Sun, J.; Eisenstein, J. Explainable Prediction of Medical Codes from Clinical Text. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018. [CrossRef]
11. Nuthakki, S.; Neela, S.; Gichoya, J.W.; Purkayastha, S. Natural language processing of MIMIC-III clinical notes for identifying diagnosis and procedures with neural networks. *arXiv* **2019**, arXiv:1912.12397.
12. Johnson, A.E.; Pollard, T.J.; Shen, L.; Lehman, L.w.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 160035. [CrossRef]
13. Banham-Hall, E.; Stevens, S. Hindsight bias critically impacts on clinicians' assessment of care quality in retrospective case note review. *Clin. Med.* **2019**, *19*, 16–21. [CrossRef]
14. Sanders, T.B.; Bowens, F.M.; Pierce, W.; Stasher-Booker, B.; Thompson, E.Q.; Jones, W.A. The road to ICD-10-CM/PCS implementation: Forecasting the transition for providers, payers, and other healthcare organizations. *Perspect. Health Inf. Manag. Am. Health Inf. Manag. Assoc.* **2012**, *9*, 1f.
15. Perotte, A.; Pivovarov, R.; Natarajan, K.; Weiskopf, N.; Wood, F.; Elhadad, N. Diagnosis code assignment: Models and evaluation metrics. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 231–237. [CrossRef] [PubMed]
16. Koopman, B.; Zuccon, G.; Nguyen, A.; Bergheim, A.; Grayson, N. Automatic ICD-10 classification of cancers from free-text death certificates. *Int. J. Med. Inform.* **2015**, *84*, 956–965. [CrossRef] [PubMed]
17. Kavuluru, R.; Rios, A.; Lu, Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artif. Intell. Med.* **2015**, *65*, 155–166. [CrossRef] [PubMed]
18. Reys, A.D.; Silva, D.; Severo, D.; Pedro, S.; de Sousa e Sá, M.M.; Salgado, G.A.C. Predicting Multiple ICD-10 Codes from Brazilian-Portuguese Clinical Notes. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, 20–23 October 2020*; Proceedings, Part I 9; Cerri, R., Prati, R.C., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 566–580.
19. Chen, P.F.; Wang, S.M.; Liao, W.C.; Kuo, L.C.; Chen, K.C.; Lin, Y.C.; Yang, C.Y.; Chiu, C.H.; Chang, S.C.; Lai, F. Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning. *JMIR Med. Inform.* **2021**, *9*, e23230. [CrossRef] [PubMed]
20. Pascual, D.; Luck, S.; Wattenhofer, R. Towards BERT-based Automatic ICD Coding: Limitations and Opportunities. *arXiv* **2021**, arXiv:2104.06709.
21. Kaur, R.; Ginige, J.A.; Obst, O. A Systematic Literature Review of Automated ICD Coding and Classification Systems using Discharge Summaries. *arXiv* **2021**, arXiv:2107.10652.
22. Dekker, J.; van Baar, M.E.; Curfs, E.C.; Kerssens, J.J. Diagnosis and Treatment in Physical Therapy: An Investigation of Their Relationship. *Phys. Ther.* **1993**, *73*, 568–577. [CrossRef]
23. Bannister, D.; Salmon, P.; Leiberman, D.M. Diagnosis-Treatment Relationships in Psychiatry: A Statistical Analysis. *Br. J. Psychiatry* **1964**, *110*, 726–732. [CrossRef]
24. Narendorf, S.C.; Cross, M.B.; Santa Maria, D.; Swank, P.R.; Bordnick, P.S. Relations between mental health diagnoses, mental health treatment, and substance use in homeless youth. *Drug Alcohol Depend.* **2017**, *175*, 1–8. [CrossRef]
25. van den Broek d'Obrenan, J.; Verheij, T.J.M.; Numans, M.E.; van der Velden, A.W. Antibiotic use in Dutch primary care: Relation between diagnosis, consultation and treatment. *J. Antimicrob. Chemother.* **2014**, *69*, 1701–1707. [CrossRef]
26. Subotin, M.; Davis, A. A System for Predicting ICD-10-PCS Codes from Electronic Health Records. In Proceedings of the BioNLP 2014, Baltimore, Maryland, 27–28 June 2014; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014. [CrossRef]
27. Chen, Y. Predicting ICD-9 Codes from Medical Notes—Does the Magic of BERT Applies Here? 2020. Available online: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1204/reports/custom/report25.pdf (accessed on 21 July 2024).
28. Tsai, S.C.; Huang, C.W.; Chen, Y.N. Modeling Diagnostic Label Correlation for Automatic ICD Coding. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021. [CrossRef]
29. Yan, Y.; Fung, G.; Dy, J.G.; Rosales, R. Medical coding classification by leveraging inter-code relationships. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 25–28 July 2010. [CrossRef]
30. Mahdi, S.S.; Papagiannopoulou, E.; Deligiannis, N.; Sahli, H. Co-Occurrence Graph-Enhanced Hierarchical Prediction of ICD Codes. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024. [CrossRef]
31. Haq, H.U.; Ahmad, R.; Hussain, S.U. Intelligent EHRs: Predicting Procedure Codes from Diagnosis Codes. *arXiv* **2017**, arXiv:1712.00481.
32. Jia, Z.; Lu, X.; Duan, H.; Li, H. Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 91. [CrossRef]
33. Watzlaf, V.; Alkarwi, Z.; Meyers, S.; Sheridan, P. Physicians' outlook on ICD-10-CM/PCS and its effect on their practice. *Perspect. Health Inf. Manag.* **2015**, *12*, 1b.
34. Goldberger, A.L.; Amaral, L.A.N.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **2000**, *101*, 23. [CrossRef]

35. Ramos, J. Using TF-IDF to Determine Word Relevance in Document Queries. *Proc. First Instr. Conf. Mach. Learn* **2003**, *242*, 29–48.
36. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python*; O'Reilly: Beijing, China, 2009; pp. 449–458.
37. Laskar, M.T.R.; Huang, J.X.; Hoque, E. Contextualized Embeddings based Transformer Encoder for Sentence Similarity Modeling in Answer Selection Task. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., et al., Eds.; European Language Resources Association: Paris, France, 2020; pp. 5505–5514.
38. Malik, P.; Aggrawal, A.; Vishwakarma, D.K. Toxic Speech Detection using Traditional Machine Learning Models and BERT and fastText Embedding with Deep Neural Networks. In Proceedings of the 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 8–10 April 2021. [CrossRef]
39. Tanaka, H.; Shinnou, H.; Cao, R.; Bai, J.; Ma, W. Document Classification by Word Embeddings of BERT. In *Computational Linguistics*; Springer: Singapore, 2020; pp. 145–154. [CrossRef]
40. Huang, H.; Jing, X.Y.; Wu, F.; Yao, Y.F.; Zhang, X.Y.; Dong, X.W. DCNN-BiGRU Text Classification Model Based on BERT Embedding. In Proceedings of the 2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS), Shenyang, China, 21–23 October 2019. [CrossRef]
41. Chanda, A.K. Efficacy of BERT embeddings on predicting disaster from Twitter data. *arXiv* **2021**, arXiv:2108.10698.
42. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
43. Wang, G.; Liu, X.; Ying, Z.; Yang, G.; Chen, Z.; Liu, Z.; Zhang, M.; Yan, H.; Lu, Y.; Gao, Y.; et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: A proof-of-concept trial. *Nat. Med.* **2023**, *29*, 2633–2642. [CrossRef] [PubMed]
44. Choi, H.; Kim, J.; Joe, S.; Gwon, Y. Evaluation of BERT and ALBERT Sentence Embedding Performance on Downstream NLP Tasks. *arXiv* **2021**, arXiv:2101.10642.
45. Yang, Z.; Kwon, S.; Yao, Z.; Yu, H. Multi-Label Few-Shot ICD Coding as Autoregressive Generation with Prompt. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 5366–5374. [CrossRef] [PubMed]
46. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
47. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
48. Clark, K.; Khandelwal, U.; Levy, O.; Manning, C.D. What Does BERT Look At? An Analysis of BERT's Attention. *arXiv* **2019**, arXiv:1906.04341.
49. Biswas, B.; Pham, T.H.; Zhang, P. TransICD: Transformer Based Code-Wise Attention Model for Explainable ICD Coding. In *Lecture Notes in Computer Science*; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 469–478. [CrossRef]
50. Wang, M.; Qiu, W.; Zeng, Y.; Fan, W.; Lian, X.; Shen, Y. IMP-ICDX: An injury mortality prediction based on ICD-10-CM codes. *World J. Emerg. Surg.* **2019**, *14*, 46. [CrossRef] [PubMed]
51. Marcou, Q.; Berti-Equille, L.; Novelli, N. Creating a computer assisted ICD coding system: Performance metric choice and use of the ICD hierarchy. *J. Biomed. Inform.* **2024**, *152*, 104617. [CrossRef] [PubMed]
52. Popescu, M.; Khalilia, M. Improving disease prediction using ICD-9 ontological features. In Proceedings of the 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011), Taipei, Taiwan, 27–30 June 2011. [CrossRef]
53. Nikolaj Blomberg, S.; Jensen, T.W.; Porsborg Andersen, M.; Folke, F.; Kjær Ersbøll, A.; Torp-Petersen, C.; Lippert, F.; Collatz Christensen, H. When the machine is wrong. Characteristics of true and false predictions of Out-of-Hospital Cardiac arrests in emergency calls using a machine-learning model. *Resuscitation* **2023**, *183*, 109689. [CrossRef]
54. Marotta, A. When AI Is Wrong: Addressing Liability Challenges in Women's Healthcare. *J. Comput. Inf. Syst.* **2022**, *62*, 1310–1319. [CrossRef]
55. Stojkovic, J.; Choukse, E.; Zhang, C.; Goiri, I.; Torrellas, J. Towards Greener LLMs: Bringing Energy-Efficiency to the Forefront of LLM Inference. *arXiv* **2024**, arXiv:2403.20306.
56. Koczkodaj, W.W.; Kakiashvili, T.; Szymańska, A.; Montero-Marin, J.; Araya, R.; Garcia-Campayo, J.; Rutkowski, K.; Strzałka, D. How to reduce the number of rating scale items without predictability loss? *Scientometrics* **2017**, *111*, 581–593. [CrossRef]
57. Miglietta, E.; Belessiotis-Richards, C.; Ruggeri, M.; Priebe, S. Scales for assessing patient satisfaction with mental health care: A systematic review. *J. Psychiatr. Res.* **2018**, *100*, 33–46. [CrossRef] [PubMed]
58. Ji, Q.; Zhang, L.; Xu, J.; Ji, P.; Song, M.; Chen, Y.; Guo, L. The relationship between stigma and quality of life in hospitalized middle-aged and elderly patients with chronic diseases: The mediating role of depression and the moderating role of psychological resilience. *Front. Psychiatry* **2024**, *15*, 1346881. [CrossRef] [PubMed]
59. Vadivel, R.; Shoib, S.; El Halabi, S.; El Hayek, S.; Essam, L.; Gashi Bytyçi, D.; Karaliuniene, R.; Schuh Teixeira, A.L.; Nagendrappa, S.; Ramalho, R.; et al. Mental health in the post-COVID-19 era: Challenges and the way forward. *Gen. Psychiatry* **2021**, *34*, e100424. [CrossRef] [PubMed]
60. Saqib, K.; Qureshi, A.S.; Butt, Z.A. COVID-19, Mental Health, and Chronic Illnesses: A Syndemic Perspective. *Int. J. Environ. Res. Public Health* **2023**, *20*, 3262. [CrossRef]

61. Paschke, K.; Napp, A.K.; Thomasius, R. Parents Rate Problematic Video Streaming in Adolescents: Conceptualization and External Assessment of a New Clinical Phenomenon Based on the ICD-11 Criteria of Gaming Disorder. *J. Clin. Med.* **2023**, *12*, 1010. [CrossRef]
62. Oldham, M.A. Describing the features of catatonia: A comparative phenotypic analysis. *Schizophr. Res.* **2024**, *263*, 82–92. [CrossRef]