Journal of Digital Health
https://ojs.luminescience.cn/JDH

LUMINESCIENCE
https://luminescience.cn/

Original Research

Designing NLP applications to support ICD coding: an impactanalysis and guidelines to enhance baseline performance when processing patient discharge notes

Jessica Jha[1], Mario Almagro[2], Hegler Tissot[3*]

[1] Data Science, Drexel University Philadelphia, USA, E-mail: jcj67@drexel.edu;
[2] Computer Science, UNED Madrid, Spain, E-mail: malmagro@lsi.uned.es;
[3] Information Science, Drexel University Philadelphia, USA, Email: hegler.tissot@drexel.edu.

*Correspondence: Hegler Tissot, Email: hegler.tissot@drexel.edu.

Abstract: Financial costs are a major concern in the healthcare system, with medical billing and coding playing a key role in facilitating transactions and financing procedures. Billing involves filing claims with insurance companies and requires scrutiny of clinical summaries and electronic health records to correctly match diagnoses, prescriptions, and procedures to standardized codes. Accuracy in assigning International Classification of Diseases (ICD) codes is critical to proper reimbursement of care. Incorrect codes waste time and resources, and cause administrative and financial problems for hospitals, insurance companies and patients. Manual medical coding is a labor-intensive and error-prone process that creates additional administrative burden and inconvenience for hospitals, insurance companies, and patients. To simplify the process, clinical records are often processed to automatically identify and extract clinical concepts and corresponding ICD codes. Deep learning and natural language processing techniques have shown promise in a variety of tasks but applying them to medical coding has been challenging. Accurate coding requires a deep understanding of medical terminology, context, and guidelines that may be difficult to capture with traditional deep learning methods. Although deep learning shows promise in healthcare, its specific impact on ICD coding is not fully understood, and translating scalable deep learning methods into practical improvements in ICD coding remains a challenge. Evaluating deep learning models under the scenarios of real-world coding and comparing them to established practice is critical to determining their true effectiveness. In this work, we address the automation of ICD coding by highlighting pitfalls and contrasting different perspectives. We investigated automatic ICD coding using baseline machine learning models, with a focus on identifying ICD-9 codes in discharge notes from Medical Information Mart for Intensive Care (MIMIC) database. A thorough evaluation of different models and approaches is crucial to avoid over-reliance on any method. Our findings show that simpler methods can achieve comparable results to deep learning models while still requiring fewer computational resources.

Keywords: Clinical coding, Natural language processing, Machine learning, Baseline models, Concept extraction, Bag of words

Received: Aug.7, 2023; Revised: Oct.9, 2023; Accepted: Oct.12, 2023, Published: Oct.30, 2023

DOI: https://doi.org/10.55976/jdh. 22023119463-81.

1. Introduction

Much of the aggravation with the current healthcare system is the financial cost. The crucial role of medical billing and coding in the healthcare industry is sometimes overlooked or goes unnoticed, yet these are the behind-the-scenes processes that facilitate financial transactions and secure funding for medical procedures. Billing involves preparing and submitting claims to insurance carriers which requires close review of clinical summaries and electronic health records (EHRs) prepared by nurses and physicians [1]. Aiming to efficiently review thousands of records, medical coding in hospitals involves matching an exclusive numerical or alphanumerical index for diagnoses, prescriptions, and procedures from these free-form clinical notes to allow for a standardized way to communicate identification of health issues [2].

One of the most used classification systems for medical codes is the International Classification of Diseases (ICD) reporting and coding guidelines [3]. The people who conduct this work are usually called coders, who are experts at summarizing and abstracting long documents and advanced clinical terminology into corresponding codes used for data analysis, communication, records, and financial purposes. Based on these codes, insurance companies deny or accept claims. Therefore, it is imperative that codes are accurate the first time so that patient care is reimbursed correctly and timely. If the codes are incorrect or the problem is misidentified, the tedious reconciliation and resubmission wastes time and resources that could have gone into the hospitals' many other processes [4].

Manual medical coding is a laborious and resource-intensive process that demands expertise in abstraction and clinical domains [2]. However, due to the inherent complexity and subjectivity, it is error-prone, demanding additional administrative efforts to rectify these mistakes. Consequently, hospitals, insurance companies, and patients face the inconvenience of dealing with bureaucratic overhead and financial documentation issues caused by these inaccuracies. To make this medical coding process more streamlined, clinical notes need to be processed in a way that identifies and extracts clinical concepts and their ICD codes when applicable [5].

Despite significant advances in deep learning and natural language processing (NLP) techniques, it remains unclear whether these efforts have resulted in significant improvements in ICD coding. While deep learning models have shown promising results on various NLP tasks such as text classification and named entity recognition, their application to the complex and detailed domain of medical coding presents unique challenges. Accurate assignment of ICD codes requires a deep understanding of medical terminology, context, and clinical guidelines that may not be easily captured by traditional deep learning methods [6]. Furthermore, the interpretability and transparency of deep learning models in healthcare raises concerns about their reliability and trustworthiness [7]. Therefore, it is important to critically evaluate the impact of deep learning techniques on ICD coding and whether they have brought significant improvements in the field. Further research and evaluation are required to verify the effectiveness and practicality of deep learning methods and to improve the accuracy and efficiency of ICD coding [8].

Deep learning models have shown promise in healthcare applications, such as when predicting specific clinical risks [9], [10] or readmission rates [11], [12], while the specific impact of these models on the accuracy and efficiency of ICD coding is not yet fully understood. Scalable and accurate deep learning for electronic health records has been explored, but translating these approaches into practical improvements in ICD coding remains an ongoing challenge. It is important to carefully evaluate the performance of deep learning models in real-world ICD coding scenarios and compare them to established coding practices to determine their true effectiveness and potential to improve baseline performance when processing patient discharge reports [13].

In this paper, we provide an overview of the trends in automation of ICD coding by highlighting the shortcomings, while at the same time contrasting different perspectives to support ICD identification at a lower level, i.e., detecting related snippets of text and using them to train baseline machine learning models for automatic ICD coding. Our study also investigates the potential of baseline methods to gain insights into the effectiveness and utility of these methods for optimizing ICD coding in patient discharge reports. We explore the effectiveness of less complex baselines for solving the task of identifying ICD-9 codes using patient discharge notes from a publicly available dataset (MIMIC). Despite the widespread use of deep learning models, we demonstrate that simpler methods can achieve comparable results while requiring less computational resources. Our results highlight the importance of thoroughly evaluating the performance of various models and methods for different tasks to avoid over-reliance on a particular approach. By considering the importance of a vocabulary versus concept approach, researchers can improve the efficiency and practicality of their solutions while still achieving competitive results.

In Section 2, we introduce the three main components of health care data flow (terming, coding, and grouping), and contrast different clinical terminologies for each of the aggregation levels. Section 3 provides the background context that supports performing ICD coding in automatic or semi-automatic ways, and we outline the different approaches proposed for automatic ICD coding and the difficulties they entail. In Sections 4 and 5, we describe how we designed the baseline methods for automatic ICD coding and the corresponding results. Finally, in Section 6 we discuss some important baseline design aspects in NLP tasks. We also propose three distinct cumulative perspectives when designing more sophisticated machine learning applications to support ICD coders, which we

consider as the following progressive natural steps in automatic annotation, as terming practically consists of a lexical problem and coding requires greater abstraction. Each perspective focuses on dealing with different levels of complexity: semantic relatedness, contextual management, and reasoning. While the latter involves greater complications, the former is feasible given the current state of the art and more easily extended to the second level using contextual information.

2. Clinical terminologies

A report on inpatient care activity in England[1] shows the amount of Finished Consultant Episodes (FCEs) recorded in 2016-2017 increased by 2.5% from the previous year and by 33.4% from 2006-2007. The course of a single hospitalization can produce hundreds of pages of clinical information, digitally stored in the form of an Electronic Health Record (EHR). That increasing amount of patient-specific information could be theoretically used to enhance health-care services and expand research opportunities [14]. However, the vast amount of clinical data generated, stored, and shared by patient primary care centers and hospitals hinders accurate analysis, potentially detrimental to decision-making, management, and health policy in patient care.

To facilitate information management, modern health centers try to automatically capture structured data related to the patients' care, such as patient problems, procedures, socio-economic status, laboratory test results, and radiological imaging data. Nevertheless, a large amount of data related to diagnoses, medications, or patient history remains as unstructured data or mainly text. Such flexibility in the use of natural language is linked to greater variability, so that the text may contain typos, incorrect syntactical structures, synonyms, abbreviations, or ambiguities, making it difficult to process automatically. Thus, computer prompting is essential to use clinical terms as the prompt to trigger the appropriate clinical documentation procedures.

Over the past decade, there has been an increased interest in converting clinical text into structured data by coupling EHR systems with a core coded clinical thesaurus, which could be a vital component to efficiently facilitate communication between healthcare professionals and support clinical practice [15]. Thus, several countries are developing infrastructure for national health information by implementing standards, nomenclatures, codes, and vocabularies with the aim of producing open, standard, and interoperable EHR systems [16].

The Language of Health [17] was designed in the 1990s and since then has been used to describe the three main constituents that comprise the data flow required for direct and indirect care of patients by healthcare providers: (a) terming (or terminology), (b) coding (or classification), and (c) grouping. Figure 1 presents a pyramid dissected into three distinct sections, each representing a different level of granularity and specificity within the language of health, in which multiple terminologies are arranged to provide different levels of aggregation. The upper layers contain narratives with a higher level of abstraction. Therefore, there is a greater gain in comprehension and completeness at the expense of losing clinical details in the terminologies located in the upper layers.

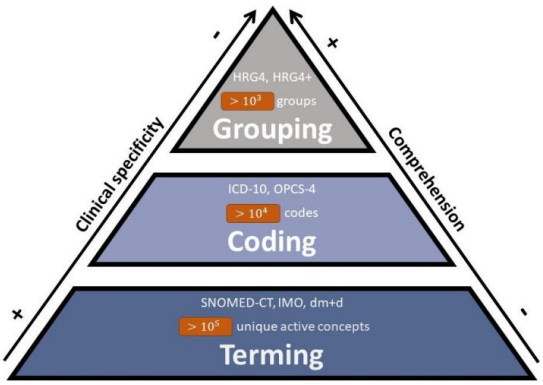

Figure 1. Pyramidal representation for the aggregation processes the flow of clinical information goes through – adapted from [17]

Clinical terminologies are key components for standardizing terms for clinical concepts. They are particularly useful for supporting many processes such as (a) the development of clinical guidelines focused on the treatment of specific conditions, (b) the retrieval of relevant data for local and national comparisons in patient care, clinical audits, and outcome studies, and (c) clinical and decision support systems [17]. Such a variety of purposes implies the design of terminologies with different granularities, which hampers a potential interoperability between them. Hence, efficient automatic retrieval of medical information is required to correctly search and find references for diagnoses, surgeries, and procedures in each patient record [18].

Differences in the magnitude of each terminology led to the development of different standards to ensure each classification is consistently applied: terming has a magnitude of hundreds of thousands of clinical terms, focused on clinical records and guidelines, audit, and decision support systems; coding includes tens of thousands of categories in order to support local service planning, contracting, epidemiology, and national assessment; and grouping uses a thousands of high-level groups to manage resources and support service planning and assessment [17]. Although it is theoretically possible to use the terms of the lower layers and external knowledge to figure out those of the upper layers (as described in the next sections), terming, coding, and grouping are mutually complementary processes.

Clinical terminologies are designed according to different criteria for different purposes. Nevertheless, there are mappings between terminologies that exploit this flow of clinical information specificity to establish equivalences between lower-layer term sets and higher-layer concepts.

1.http://digital.nhs.uk/pubs/apc1617

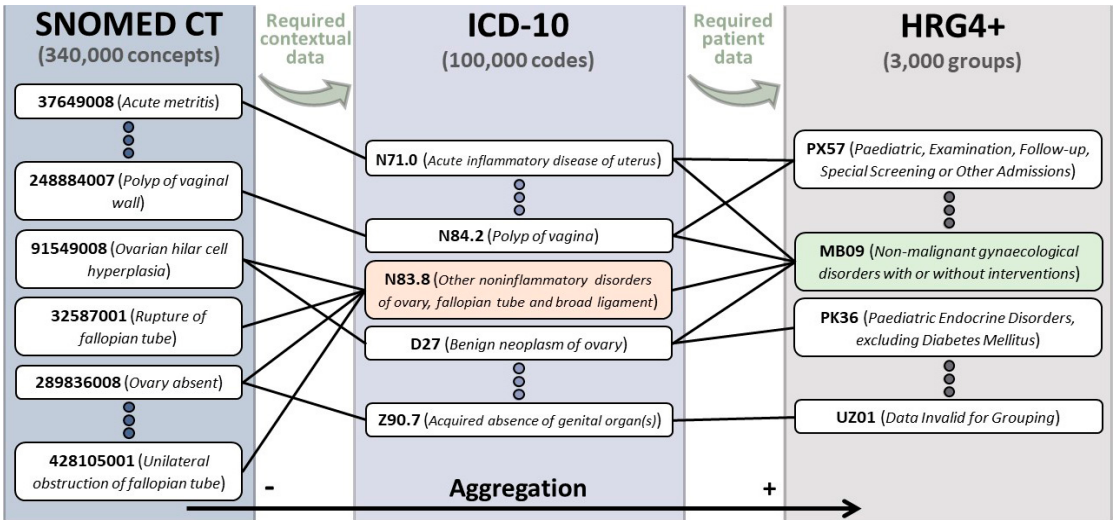

Figure 2. Aggregation process using SNOMED CT to ICD-10 and ICD-10 to HRG4+ mappings: (a) the mapping process between SNOMED CT to ICD-10 uses contextual data to translate detailed clinical information into standardized diagnostic codes; (b) then patient data is used to aggregate ICD-10 codes into HRG4+ categories based on various factors such as clinical similarity, resource utilization, and cost patterns

These mappings are unidirectional (Figure 2), since the aggregation process leads to the loss of clinical information, which does not allow the original meanings to be restored. Although meanings of a lower layer can be aggregated into those of a higher layer, external information, not contained in the terminology, is still needed (e.g., the mapping between the terminology and coding layers may require contextual information or non-clinical data, while mapping from coding to grouping, requires patient data). Their chain of converting detailed clinical information from terms captured in the EHR system into classification assignments is of the fundamental importance, for example, to support national commissioning dataset returns [19], and as the unit currency to support the local commissioning and contracting process [20].

The required infrastructure to meet the increasing demands for improved quality, greater volume of services and more effective use of resources, includes shared information flow, national standards to enable inter-computer communication, and carrying information aligned to the purposes of the healthcare industry. At this point, the language components become relevant, with: (a) terming being the most fundamental building block for any set of clinical data; (b) coding offering an intermediate level of aggregation which is useful for statistical analysis of incidence and trend, and service management; and (c) grouping as a higher level of aggregation for planning, contracting, and commissioning purposes [17].

2.1 Terming

Terming is designed for clinicians, and serves as a front-end interface for naming standard data to capture clinical concepts in the form of medical terms with appropriate granularity and precision. On the back-end, various clinical descriptions are mapped to codes that form an ontology, enabling the EHR system to automatically recognize a unique concept ID for each term. This ensures standardization across different platforms, organizations, and even countries when exchanging health data.

In the UK, the National Health Service (NHS) uses SNOMED CT [21] as a systematized nomenclature of medicine. SNOMED CT is a comprehensive standard reference and interface terminology that supports both general and very specific concepts. Each concept is defined by a set of attribute-value pairs (relations) that distinguish it from all other concepts. SNOMED CT supports a model that specifies correct attributes and value sets for each domain of meaning, comprising one code per meaning, one meaning per code.

In [22], different sources of controlled clinical terminology are compared in terms of the attributes of completeness, clinical taxonomy, administrative mapping, term definitions, and clarity, by assembling 1,929 source concept records from a variety of clinical information. SNOMED CT was considered richer as a clinical taxonomy due to its compositional nature and was found to be much more complete in identifying clinical details suitable for terming. However, there are many more duplicates of code assignments on SNOMED CT, with a loss of clarity, due to a lack of syntax and evolutionary changes in the coding scheme. The Unified Medical Language System (UMLS) [23] was pointed as a rich lexical resource, with mappings to many source vocabularies, although it still has limitations in clinical representation in an EHR perspective, mainly due to the different granularities and purposes of its source schemes.

2. https://www.gov.uk/government/publications/personalised-health-and-care-2020

With a vision to support patient care, by enabling different platforms to share data without ambiguity, the Personalized Health and Care 2020[2] white paper stated that by 2020 all health records should be digital real-time and interoperable, and by April 2020 the entire health care system should adopt SNOMED CT as the clinical data standard in the UK, as agreed by NHS Digital.

2.2 Coding

The intermediate layer in Figure 1 is referred to as coding. According to the World Health Organization (WHO)[3], medical coding consists of characterizing health reports with standardized sets of codes that represent diagnoses, thereby achieving to report diseases and health conditions as a foundation for identifying health trends and statistics.

The most widely used classification systems are the International Classification of Diseases and Related Health Problems (ICD) [3] and the Classification of Interventions and Procedures (OPCS) [24]. Both are typically used for statistical purposes, with a higher level of abstraction requiring more complete information processing about meanings, contexts, and interactions. The design of these standards focused on the ease of storage, retrieval, and analysis of health information for data comparisons that can provide evidenced-based decision-making. These criteria promote easy statistical monitoring of the incidence and prevalence of diseases, injuries, symptoms, reasons for occurrence, and other factors that influence health status. In this way, such monitoring can be properly used for service, billing, planning, research, and education, so that the better quality and detail of coding, the better quality and detail of information available, leading to improved patient care [18].

For example, in the UK, the current coding system for diagnoses and healthcare related problems in the NHS is the 10th revision of the ICD (ICD-10). The operating environment requires the Patient Administration System (PAS) to provide monthly commissioning datasets comprising ICD codes for each hospital patient admission when discharged as well as hospital outpatient visits. ICD codes are manually assigned by highly trained health specialists (coders), who analyze the case notes (normally in the free text format) as well as use available computerized systems and abstract the clinical information, such as diagnoses, comorbidities, procedures, complications and any other issues related to healthcare. The information that coders currently access is normally in the free text format. Consequently, this task entails great expenses. In addition, although ICD is designed for non-clinical use, this standard requires the workforce who interact with the systems to possess sufficient clinical knowledge, so that they are able to understand the subtlety among different health conditions described in the patient record.

Coding terminologies are designed at a much more aggregated level, and include many meanings per code. ICD uses a hierarchical structure to define the universe of diseases, which allows coders to use residual categories (other specified or unspecified) to capture the leftover equivalents from the lower terming level that may not fit any of the specific categories. In this way, the granularity and purpose of ICD is different from that of SNOMED CT. The latter consists of 400,000 concepts to be used for clinical terming, whereas ICD comprise approximately 70,000 concepts to be used along the clinical coding process. In theory, it should be possible to automatically link ICD to SNOMED CT as part of the back-end transition between terming and coding – UMLS provides mapping resources between multiple clinical terminologies.

However, the focus of coding is statistical, as opposed to using lexical resources - the more general and vaguer the coding descriptions are, the less granular the mapped clinical terms become, which encompass multiple, sometimes nonclinical, meanings. In ICD, the less common diseases are grouped into general categories, which lead to a loss of information during coding: (a) not otherwise specified (NOS) codes, which are used in cases with insufficient information for more specific codes; and (b) not elsewhere classified (NEC) codes, comprising cases with more specific information but are not covered by existing ones. An example of lossy compression is ICD-10 code N83.8 (Other noninflammatory disorders of ovary, fallopian tube, and broad ligament) shown in red in Figure 2. For coders, it means that there are certain conditions for which they are not able to find a more specific code along the ICD standard. However, the description of N83.8 may not be fully clear to clinicians, a lack of meaning that SNOMED CT supports by providing over 30 unique concepts that can be matched to N83.8.

2.3 Grouping

Finally, the top layer in Figure 1 is called grouping, and it is designed for administrative staff and can be roughly translated as how the Trust gets paid, as the way of supporting and managing reimbursement for provided health services.

Diagnosis Related Group (DRG) [25] is a health classification system used in several European countries to standardize prospective payment for hospitals and to promote cost containment initiatives. DRG covers all charges associated with an inpatient stay from the time of admission to discharge, including services performed by an outside provider. In addition to patient's information (e.g., age, gender, admission method), trusts can also be paid differently based on whether the patient stay is an elective admission or an emergency admission.

Within the NHS in the UK, for example, Healthcare Resource Group (HRG) [26] is the way of materializing the grouping process. For example, sets of ICD codes

3. https://www.who.int

are mapped to HRG codes using additional patient information such as gender and age to disambiguate them. HRG is the classification system at the most aggregated level and consists of patient events that have been judged to consume a similar level of resource. For example, different knee-related procedures requiring similar levels of resources may be assigned to the same HRG[4]. An example of this level of aggregation from coding can be illustrated in Figure 2 by the ICD-10 code N83.8, which is mapped as the HRG code MB09 (Non-malignant gynecological disorders with or without interventions), highlighted in green color - , there are 264 ICD-10 codes that can be grouped into the same HGR.

3. Related work

The integration of automatic tools for clinical textual processing in health systems is estimated to be a major step forward on embracing more focused and personalized actions [27], which ultimately means increasing the productivity of health professionals. However, there remain some unresolved big challenges to be overcome, such as automatic ICD coding. Assigning ICD codes to EHRs directly affects a multitude of processes, from the calculation of morbidity and mortality statistics to the estimation of medical expenditure, and it involves many complexities for both coders and automatic systems. Thus, most of the proposed approaches do not achieve satisfactory performance and cannot be applied to real systems. In this section, we outline some attempts to fully solve this task.

The drawbacks of ICD coding can be divided into those that are inherent to the field, and others that are specific to each task. Regarding the former, the biomedical area involves a high degree of linguistic variability due to the large and specific vocabulary and the abundance of synonyms and lexical variants. Especially in the clinical field, typographical errors, an abundance of abbreviations and acronyms, and a lack of grammatical structures are common due to the heavy workload of clinicians. Therefore, automatic ICD coding is a high-level task that is even more complex than linking other clinical concepts to EHRs. It requires to understand the context and semantics and tends to follow unbalanced distributions.

Different ways of dealing with automatic ICD coding can be distinguished, and the proposals can be grouped into three main groups. Unsupervised approaches can be used to find similarities between ICD descriptions and clinical concepts in the text. However, given the complexity of working with more abstract meanings, supervised approaches are the most widely used algorithms [2], [28]. Finally, using more specific clinical terminologies to identify the most relevant concepts and subsequently to exploit the equivalences made by experts with ICD is explored by mapping approaches. The strengths and weaknesses of each approach are discussed below.

3.1 Unsupervised approaches

Unsupervised approaches link ICD-provided codes to EHRs. Medical knowledge bases and ontologies are normally used to identify health concepts in text, and further find correspondences with the concepts based on the ICD descriptions. However, health authorities frequently use terminologies different from those reflected in ICD code descriptions, mainly because the former represents more specific instances, and the latter is associated with a higher level of abstraction (that makes it difficult to directly look for similarities). Besides that, the textual pieces required to identify a diagnosis do not necessarily have to be found sequentially in the clinical notes - on the contrary, the pieces of text with significant meanings are often scattered throughout the EHRs.

Contextualizing meanings is key to determining what information should be considered. The meanings of the concepts are highly influenced by co-occurring concepts or modifiers [29]. Therefore, various limitations must be considered when collecting the scattered data to exclude irrelevant information such as denied, non-patient-related, temporal- and clinical-suspicion-related concepts. In addition, there is the fact that this is a goal-oriented task, meaning that only medical information related to the cause of hospitalization is coded. Thus, during the clinical coding process, different constraints must be considered to exclude irrelevant information such as negation statements, suspected condition, or information that does not pertain to the patient. However, a negation such as "the patient has no personal history of myocardial infarction" is equally important to the clinical assessment but need not be coded. Coding is not required for either (a) a documented condition that does not affect the patient's care treatment at the time of this admission, or (b) conditions that were previously treated and no longer exist. For example, if the patient is admitted for pneumonia, the fact that the person had a wrist fracture ten years ago has no bearing on the current admission. Nevertheless, clinicians note these conditions when they write down the patient’s medical and surgical history.

Despite all these difficulties, an unsupervised coding method can have a wide coverage, avoid the frequent biases in this task, and be suitable for any data collection. This motivates numerous studies based on word co-occurrence [30]-[32] to explore the generation of queries from documents and their expansion through medical knowledge databases. In contrast, semantic similarity is explored to match words between ICD descriptions and text by applying the Longest Common Subsequence (LCS) method [33] or using ontologies to estimate similarities [34].

3.2 Supervised approaches

4. NHS is currently using HRG4+, which is derived from the clinical coder’s coded information, as a unit of currency to underpin the financial scheme PBR (Payment by Results).

ICD coding can be considered as a multi-label classification task using supervised learning. To this end, labelled data is required, though strong data protection regulations make it a hard task. Moreover, the data generated by healthcare centers are extremely scattered, biased, and unbalanced.

Both available medical services and local factors determine what percentage of ICD codes are reported by the healthcare center. In addition, the nature of diagnoses provides a clear tendency to concentrate on a small number of codes, while the vast majority appear very infrequently. In this way, a supervised model will not be able to model the missing or infrequent codes during training, thus only a very small group of ICD codes that consist of the most likely diagnoses in the center can be predicted.

The strength of supervised models comes from their performance in encoding similar data, which often far exceeds that of other approaches. Therefore, supervised learning is useful when the distribution of training data to which the model is to be applied is not unbalanced and overlaps the distribution of training data. Some authors have opted for this way to achieve better classification results by exploiting the hierarchical ICD structure [35], [36]. However, given that such balanced distribution is impractical, most authors tend to use external sources of knowledge to enrich the learning step, which includes (though not restricted to): (a) adding medical terminologies into the representation of documents [37]; (b) learning patterns directly from dictionaries [38]; (c) combining supervised learning with information retrieval techniques [39], [40]; (d) combining dictionaries, other corpora and synonyms [41]-[43]; (e) enriching small corpora with equivalence mappings [44]; and (f) utilizing word embeddings trained from medical documents [45].

The application of deep learning to ICD code identification offers several advantages. Approaches to ICD code identification have changed significantly in recent years, with deep learning emerging as the standard implementation reference for this type of problem. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have made progress in extracting complex patterns and representations from patient discharge notes [46]. These models have revolutionized the field by automatically learning hierarchical features, which enables them to capture intricate relationships within textual data. By leveraging large-scale annotated datasets, deep learning algorithms can learn robust representations that adapt to different clinical settings and coding practices.

Deep learning models must adapt to handling diverse and evolving medical terminologies and language variations. Results are promising, but it is essential to consider the potential limitations and challenges they pose. Deep learning models often require substantial computational resources and extensive training data to achieve optimal performance. The training process can be computationally expensive and time-consuming, especially when working with large volumes of patient discharge notes. Additionally, deep learning models may be prone to over-fitting, especially when dealing with limited or imbalanced data sets. It is crucial to address these challenges and explore alternative approaches to ensure the practicality and efficiency of ICD code identification systems.

A study in 2019 [47] aimed to assess the performance of deep-learning-based systems in automatically mapping clinical notes to ICD-9 medical codes. The research focused on end-to-end learning methods without manually defined rules. Traditional machine learning algorithms and state-of-the-art deep learning methods, including Recurrent Neural Networks and Convolutional Neural Networks, were applied to the MIMIC-III dataset. The results demonstrated that the deep learning-based methods outperformed conventional machine learning methods. The best models achieved an average F1 score of 0.6957 and an accuracy of 0.8967 in predicting the top-10 most common ICD-9 codes in MIMIC. However, this is only true when all scores for each of the top 10 labels were averaged together. For the class-based results, the study shared the Average Precision Score which ranged from 0.4 to 0.9. Based on the assessment using standard metrics, the study concluded that the deep learning-based systems showed superior performance in assigning ICD-9 codes on the MIMIC-III dataset.

A recent study [48] reproduces, compares, and analyzes the state-of-the-art and automated medical coding machine learning models, and it shows that several models underperform due to weak configurations, poorly sampled train-test splits, and insufficient evaluation. In addition, the analysis confirms that models struggle with rare codes, while long documents only have a negligible impact.

In [49], the authors conducted a study on supervised multi-label text classification for assigning ICD-9-CM diagnosis codes to electronic medical records (EMRs) from three datasets of varied scale. They compared their approaches, which included problem transformation techniques, feature selection, training data selection, label calibration, and learning-to-rank, with basic approaches to evaluate the impact of these additional learning components on diagnosis code assignment. Results showed that the classifier chains performed comparably to the state of the art on a gold standard dataset with short reports. They also examined a large dataset (UKLarge) and a subset (UKSmall) and found that feature selection, data selection, and label calibration significantly improved performance on UKSmall but did not have the same effect on UKLarge.

Finally, the HiLAT model [50] uses a hierarchical label-wise attention mechanism and a pretrained transformer language model, called ClinicalplusXLNet, for automatic ICD coding from discharge reports. HiLAT extracts specific text representations for each ICD code and maps them accordingly. The model's label-wise attention

weights are utilized to highlight the keywords that contribute to code predictions. The authors suggest that HiLAT, along with ClinicalplusXLNet, can achieve state-of-the-art results in multi-label text classification tasks, particularly in the clinical health domain. They propose deploying HiLAT to enhance and streamline manual processes in clinical coding, focusing on predicting the 50 most frequent ICD codes. Additionally, the authors highlight potential applications of HiLAT in automated patient identification for clinical trials and identification of specific clinical endpoints in real-world evidence studies.

3.3 Mapping approaches

Oppositely to directly encode documents by using ICD, structured data can be extracted from EHRs in the form of specific clinical concepts and then used to derive ICD codes through an expert-knowledge mapping. Different authors highlight that maintaining more specific terms during encoding EHRs is a prerequisite to avoid the loss of clinical information in the upper layers [15], [51]. Hence, mappings are pointed out as a promising way to transform collected meanings into less granular categories.

SNOMED CT is the most complete clinical terminology. ICD equivalent mapping is available and under continuous development by the International Health Terminology Standards Development Organization (IHTSDO) and WHO [52]. Granularity of ICD codes leads to the aggregation of SNOMED CT codes. The cardinality is zero-to-many. An ICD code may include many SNOMED CT concepts, so there are several ways to group the SNOMED CT codes extracted from a document. For this reason, the available mapping is a rule-based mapping that suggests multiple candidates. The idea is to present all ICD codes that contain the meaning elements of the collected SNOMED CT concept. SNOMED CT itself is not sufficient to find equivalences, so additional patient information, such as age or gender is required.

The mapping from SNOMED CT to ICD might facilitate an encoding that preserves the detailed clinical information. 45% of SNOMED CT concepts with ICD equivalencies can be directly associated with a code, without ambiguity; the remaining concepts depend on both other codes and additional contextual information. Although this seems to be the right way to solve the automatic coding, there are several problems that hinder its unsupervised application.

Given the differences in terms of granularity, SNOMED CT to ICD mapping is not exhaustive – more than 6% of codes are not covered by equivalencies (e.g., concepts with laterality or episode of care information). Moreover, the ambiguity occurs in different forms. Currently, there are 724 ambiguous concepts that cannot be assigned until clarified, and more than 25,000 have multiple targets. For example, the concept 68449006 (Coxitis) is associated with 14 different ICD-10 codes. There are also contradictions such as synonyms corresponding to different ICD codes due to the structure of the two terminologies. Finally, although contextual information is required to perform a full ICD coding task, SNOMED CT to ICD mapping is based on context-free assumptions. The meanings of SNOMED CT per se do not provide all the supporting information. Thus, the mapping is not designed to automatically assign ICD codes, but to suggest a list of equivalent ICD codes from a given SNOMED CT concept, and to delegate the task of choosing the appropriate codes to a person.

In January 2022, WHO released the 11th revision of the International Classification of Diseases (ICD-11), which has evolved beyond its original purpose in epidemiology to encompass billing, quality and safety, and research. However, the transition to ICD-11 is expected to take an effort for 4-5 years, and requires guidance and testing from WHO. Successful implementation within the healthcare system demands ongoing investment and planning, as well as the evaluation of the impact of ICD-11 on existing processes. Sharing this knowledge within organizations, along with testing and implementing solutions, will streamline the transition and ensure that ICD-11 effectively meets the diverse needs of its users [53].

4. Materials and methods

The field of deep learning has witnessed a surge in popularity over the last decade, with many researchers using these models to tackle various tasks. However, it is often unclear how efficient these methods are, compared to less complex baselines. While deep learning models have achieved remarkable success in several domains, their efficiency is sometimes questionable, which has prompted researchers to discuss the complexities, limitations, and challenges associated with deep learning models, including their interpretability, efficiency, and generalization performance, as well as to investigate alternative, less complex approaches to solving some tasks [6].

While deep learning has attracted a lot of attention in natural language processing (NLP) and healthcare, baseline methods play a crucial role in the field of machine learning as they provide a point of reference and comparison for evaluating the performance of advanced models [54], [55]. Baseline models can also effectively capture contextual information and dependencies between words and phrases in patient discharge notes, though with expected less accuracy. However, this holistic understanding allows models to grasp the subtle nuances and semantics necessary for establishing baseline performance in ICD code assignment.

For our experimental evaluation, clinical records were obtained from the Medical Information Mart for Intensive Care, MIMIC [56], an open-source database for de-identified health-related data from patients in critical care

units. We used version 3 of MIMIC, which covers the years 2008-2019. Of the available health-related data, only discharge summaries were used - a total of 59,652 discharge summaries are available in MIMIC-III. We then identified and removed discharge notes corresponding to cases with more than one discharge summary per hospital admission. The result was that 47,006 discharge notes were randomly split, approximately 60% for training, 10% for tuning, 30% for testing.

We aim to establish stronger baselines in text classification in the healthcare domain. Therefore, we explore the use of the Logistic Regression [57] and XGBoost (eXtreme Gradient Boosting) [58] algorithms as baseline methods to evaluate the performance of vocabulary- based versus concept-based datasets when identifying ICD-9 codes in patient discharge notes gathered from MIMIC-III dataset.

XGBoost is a powerful algorithm that can be used for both regression and classification tasks. Its ability to handle complex, high-dimensional datasets and to provide highly accurate predictions make it a popular choice among data scientists and machine learning practitioners [59]. It has gained significant attention and achieved a level of success in various machine learning competitions and real-world applications due to its efficiency, scalability, and high-quality performance. In a recent study, researchers attempted to use XGBoost to develop mortality prediction models and compared their performance with logistic regression, Injury Severity Score (ISS), and Trauma Mortality Prediction Model based on International Classification of Diseases (TMPM-ICD10), using a dataset that uses ICD-10 codes. The findings suggest that machine learning models utilizing XGBoost outperform logistic regression, ISS, and TMPM-ICD10 in predicting mortality [60].

4.1 Feature extraction

Feature engineering refers to the process of extracting relevant features from raw text data to enhance the performance of machine learning models. This crucial step reduces irrelevant information and redundancies from raw text data that can negatively impact the performance of models. MetaMap [61] is one of the available tools that facilitates information extraction, text mining, and information retrieval by identifying relevant concepts in clinical text.

MetaMap uses a combination of lexical and syntactic features to identify and annotate medical concepts and entities in clinical text, such as diseases, treatments, and laboratory results. These features include, but are not limited to, keywords, part-of-speech tags, dependency relationships, and semantic types. MetaMap also employs domain-specific knowledge sources, such as the Unified Medical Language System (UMLS) [23], to enhance its feature extraction capabilities. By utilizing a combination of features and knowledge sources, MetaMap can accurately identify and extract relevant information from clinical text, which is essential for various biomedical applications such as clinical decision support systems.

We evaluate different feature engineering methods for predicting ICD codes. First, we processed discharge notes through MetaMap to extract unique concept identifiers (CUIs) corresponding to SNOMED-CT terms. In addition, we used spaCy [62] to capture the vocabulary from each discharge note in the form of lemmatized tokens. Table 1 outlines the number of discharge notes, and the total number of CUIs and words (lemmas) identified in each training, tuning, and testing subsets.

We also tried to use stemming, a heuristic that reduces words to their base or stem form by removing suffixes, reducing reliance on dictionaries, and potentially making the use of non-lexical words and technical terms more robust. However, compared to lemmatization, stemming is a less linguistically sound approach that may lead to less accurate root representations. Indeed, our experimental results show that XGBoost trained on both lemmas and stems have equivalent F1 scores in the test set, with differences observed in the third decimal place, mostly in favor of those models trained with lemmas [63].

Table 1. MIMIC-III discharge notes in each subset after pre-processing and removing patient admissions with more than one discharge note

Set	Discharge notes	CUIs	Lemmas
Training (60%)	28,204	43,041	10,722,340
Training (10%)	4,701	26,882	1,904,405
Training (30%)	14,101	36,296	5,450,021
Total	47,006		

The full set of CUIs and lemmas was then divided into three distinct experimental datasets composed by binary features:

1)BoW: the Bag-Of-Words dataset contains binary features (0 = absent, 1 = present) that indicate whether a lemma occurs in each discharge note. We only consider lemmas that occur in at least 68 discharge notes. This is the statistical sample required to estimate the characteristics of the entire population at a confidence level of 90% within a ±10% margin of error, adding up to 8,933 different lemmas.

2)CUI: In the second experimental dataset, each binary feature corresponds to the occurrence of CUIs in each discharge note. 8,944 different CUIs occurring in at least 68 notes were considered.

3)DSyn: Finally, the third dataset is a subset from the CUI dataset, only considering 946 CUIs labeled by MetaMap as belonging to the semantic type DSyn (Disease or Symptom), as a way to evaluate how much direct mentions those concepts can be correlated to the resulting ICD code for each patient.

Finally, the target labels for the classification task were

the top-10 most common ICD-9 codes in MIMIC-III database. Table 2 depicts the number of positive cases for each ICD-9 code among the 47,006 considered discharge notes. We used the top-10 ICD codes from MIMIC-III because they offer less imbalanced distributions of labels which reflects in less bias towards the predominant negative class, which makes the experiments more reliable in terms of how resulting performance is reported and facilitates comparison against previous research work.

Table 2. Top-10 MIMIC-III ICD-9 codes with the corresponding percentage of positive occurrences within the 47,006 patients with unique discharge notes

ICD-9	Description	Positive cases
4019	Hypertension	38.31%
4280	Congestive heart failure	23.81%
42731	Atrial fibrillation	23.27%
41401	Coronary atherosclerosis	22.83%
5849	Acute kidney failure	17.00%
25000	Diabetes Type II	16.96%
2724	Hyperlipidemia	16.56%
51881	Acute respiratory failure	13.89%
5990	Urinary tract infection	11.97%
53081	Esophageal reflux	12.04%

4.2 Baseline models

Although ICD coding can be viewed as a multi-label classification task, multi-label classifiers face challenges in handling complex relationships between multiple classes. The presence of multiple labels can lead to dependencies and correlations between classes, making the evaluation of performance metrics more nuanced and often requiring specialized metrics such as Hamming loss or subset accuracy [64]. Moreover, they can become computationally expensive and memory-intensive, especially when dealing with a large number of classes and complex relationships between them. The scalability of multi-label classifiers may be a concern when dealing with high-dimensional data or large-scale multi-label problems [65]. Conversely, binary classifiers can focus on distinguishing the minority positive class on imbalanced datasets, leading to potentially higher precision and recall for each individual class [66]. Finally, binary classifiers are generally more scalable since only two classifiers (positive and negative) need to be trained, and the training time and computational resources required are generally less compared to multi-label classifiers [67]. We train binary classification models to predict class probabilities on the training set and then extract the predicted probabilities for the positive class. The predicted probabilities and the true labels of the tuning set are used to calculate precision, recall, and thresholds for the precision-recall curve analysis.

Both Logistic Regression and XGBoost models are trained on the training set. In Logistic Regression, the maximum number of iterations used for each model was 10,000, with no convergence warnings reported. Due to the imbalanced nature of the proposed task, we then use the tuning set to obtain the threshold that maximizes the F1 score based on the Area Under the Precision-Recall Curve (AUPRC, also known as Average Precision). This threshold value is considered the optimal point that balances precision and recall, and it is used to decide the predicted class in the test set. We set the model prediction as a positive class (1) when the resulting probability is≥ threshold (as opposed to using 0.5 as the standard threshold for balanced datasets), otherwise setting it as a negative class (0) - in imbalanced datasets, the tuned threshold tends to be lower than 0.5.

For XGBoost, one additional step must be taken to determine the optimal depth (hyperparameter), tested from 4 to 12 in the tuning set. During the tuning process of the XGBoost models, we found that the optimal depths varied depending on the specific data sets and labels. For the BoW and CUI datasets, average depths were 9.9 and 10.4 respectively, ranging 5-13 for the BoW dataset, and 7-15 for CUI. For the DSyn dataset, the average depth was 12.0, as expected by the fact that DSyn is a subset of CUI dataset and might impose more decisions to figure out predictions due to the loss of information resulting from SNOMED CT concepts not considered in this UMLS semantic type.

Finally, the F1 score is calculated using the test set. The area under the curve and the average precision scores were also recorded to be compared with the previous literature. Each label has an independent model that went through training and tuning before applying the test set for label identification.

5. Results

The baseline scores resulting from the Logistic Regression and XGBoost models are presented in Table 3. In experiments with three different datasets, the results consistently show that XGBoost outperforms logistic regression in all precision, recall, and F1 score.

Logistic regression is a linear model that has difficulty in handling nonlinear patterns and high-dimensional data effectively. It performs better overall on all metrics in the DSyn dataset. However, DSyn is a subset of CUI dataset, which means that the first dataset contains only a portion of the data present in the second. Although a loss of information compared to the original dataset from which it was derived is expected, logistic regression still performs better in this scenario.

In contrast, XGBoost demonstrates superior performance by effectively capturing complex relationships and interactions in the data. Its ability to cluster weak learners and optimize decision boundaries can lead to better prediction accuracy and more robust generalization. This

Table 3. Baseline Precision, Recall, and F1 scores resulting from Logistic Regression and XGBoost models tested in three different structured representations for features extracted from MIMIC discharge notes: (a) Bag-of-words (BoW); (b) SNOMED CT concepts resulting pre-processing notes with MetaMap (CUI); and (c) the Diagnosis and Symptoms semantic type subset extracted from MetaMap (DSyn)

ICD-9		Precision						Recall						F1					
		Logistic Regression			XGBoost			Logistic Regression			XGBoost			Logistic Regression			XGBoost		
Code	Description	BoW	CUI	DSyn	BoW	CUI	DSyn	BoW	CUI	DSyn	BoW	CUI	DSyn	BoW	CUI	DSyn	BoW	CUI	DSyn
4019	Hypertension	0.602	0.601	0.634	0.703	**0.730**	0.626	0.798	0.802	0.785	0.868	**0.878**	0.815	0.686	0.687	0.701	0.777	**0.797**	0.708
4280	Congestive heart failure	0.637	0.619	0.730	**0.749**	0.737	0.720	0.709	0.750	0.789	0.815	**0.829**	0.814	0.671	0.678	0.758	**0.781**	0.780	0.764
42731	Atrial fibrillation	0.863	0.809	0.891	0.887	0.890	**0.905**	0.782	0.801	0.752	**0.922**	0.900	0.745	0.821	0.805	0.816	**0.904**	0.895	0.817
41401	Coronary atherosclerosis	0.665	0.719	0.685	0.793	**0.799**	0.696	0.759	0.677	0.688	**0.801**	0.790	0.697	0.709	0.697	0.687	**0.797**	0.795	0.697
5849	Acute kidney failure	0.503	0.575	0.660	0.665	**0.724**	0.672	0.603	0.517	0.607	**0.679**	0.661	0.650	0.548	0.544	0.632	0.671	**0.691**	0.661
25000	Diabetes Type II	0.519	0.510	0.484	**0.664**	0.621	0.477	0.612	0.577	0.649	**0.782**	0.782	0.733	0.562	0.541	0.544	**0.718**	0.692	0.578
2724	Hyperlipidemia	0.505	0.547	0.624	0.641	**0.650**	0.607	0.697	0.560	0.774	**0.855**	0.785	0.792	0.586	0.553	0.691	**0.733**	0.711	0.687
51881	Acute respiratory failure	0.444	0.444	0.482	**0.590**	0.577	0.573	0.627	0.635	0.632	0.694	**0.736**	0.548	0.520	0.523	0.547	0.638	**0.647**	0.560
5990	Urinary tract infection	0.583	0.585	0.630	0.639	**0.649**	0.636	0.551	0.530	0.704	0.733	**0.744**	0.718	0.567	0.556	0.665	0.682	**0.693**	0.675
53081	Esophageal reflux	0.620	0.592	0.679	0.680	**0.680**	0.671	0.564	0.657	0.826	**0.840**	0.830	0.837	0.591	0.623	0.745	**0.751**	0.748	0.745
													Macro F1	0.626	0.621	0.680	**0.745**	0.745	0.689

Best score in each subset highlighted in gray; best precision, recall and F1 score in each ICD code underscored in bold

suggests that XGBoost is a more reliable and robust choice for designing baseline models for classification tasks on these datasets, which mostly show better performance between the BoW and CUI datasets, but rarely outperform the DSyn subset. XGBoost Macro F1 is still slightly better in BoW (0.7453) when compared to the same score resulting from CUI dataset (0.7449).

Despite the differences in feature representation methods, models trained on bag-of-words features were as effective as models trained after pre-processing clinical records and extracting SNOMED CT concepts. This means that capturing the frequency of individual words in the text may be as effective as capturing the semantics and relationships between medical concepts represented by the corresponding SNOMED CT codes. However, it should be noted that the effectiveness of these models highly depends on the nature of the task and the complexity of the dataset. Bag-of-Word features are sufficient for tasks involving simple relationships between words, while clinical record preprocessing and SNOMED CT concept extraction can provide significant benefits when dealing with complex medical language and interrelated medical concepts. In addition, the computational cost and data pre-processing effort associated with SNOMED CT concept extraction should also be considered. In some cases, the added complexity of the pre-processing step may not justify a small performance gain, making bag-of-words a more practical and efficient choice. In general, both approaches can produce effective underlying models, but decisions should be made considering the specific requirements of the task, the complexity of the data, and the trade-off between performance and computational resources.

Table 4 compares different approaches in terms of how the results are reported. Although most approaches present F1 scores in the test set, many do not disclose the Precision and Recall scores used to achieve F1. However, when reported, Precision and Recall can be used to explain how the model was tuned in favor of one of the composite F1 metrics, such as in [49], in which we clearly see that the model favors Precision, with an evident imbalance compared to Recall, making the F1 score drops significantly.

Table 4. Comparison of different ICD coding approaches based on the macro average of different scores: Precision (P), Recall (R), F1 score (F1), Area Under the Receiver Operating Characteristic Curve (AUC), and Area Under the Precision-Recall Curve (AUPRC)

Auto ICD Coding Approach	P	R	F1	AUC	AUPRC
Our Baseline - XGBoost BoW (10 codes)	0.701	0.799	0.745	0.934	0.778
[49] CMC dataset	0.540	0.440	0.470		
[49] UKLarge LR + L2R + NERC	0.822	0.218	0.211		
[50] HiLAT + ClinicalplusXLNet (50 codes)	0.627	0.710	0.690	0.927	
[68] BERT-ICD (50 codes)				0.845	
[69] MultiResCNN (50 codes)			0.606		
[70] KEPTLongformer (50 codes)	0.673*		0.689	0.926	
[71] Text-TF-IDF-CNN + LS + DR + TD (50 codes)			0.687	0.934	
[49] UKLarge (≥2%; 92 codes)	0.959	0.187	0.167		
[69] MultiResCNN (full codes)			0.085		
[72] Zero-Shot ICD Coding	0.317	0.281	0.298	0.941	
[73] eFastText-UMLS (full codes)	0.479	0.629	0.544		

* Precision@5

The area under the curve is supposed to be calculated during tuning and model optimization. Our approach uses tuning to find the optimal threshold for separating positive and negative classes and to maximizing F1 score while maintaining a balance between Precision and Recall. Although we use the tuning set to perform such optimization, the overall model performance still remains equally accurate when using the test set for evaluation, with minor and expected decreases in the reported F1 score. However, in previous research work, it is frequently unclear whether AUC and AUPRC scores are calculated in the validation/tuning set or incorrectly using the test set, which is not appropriate for such metrics. Finally, AUC is mostly used, though it is not the best solution for evaluating imbalanced datasets, as when using AUPRC.

The differences between AUC and AUPRC are evident in Figure 3. Although each of the top-10 ICD-9 codes in MIMIC poses different levels of complexity regarding the ability of each model to predict them correctly, mainly due to the imbalanced aspect of each label (varying from 38% positive in the top-1 ICD-9 code to 12% positive in the top-10 ICD-9 code as described in Table 2), the average AUC tends to be flat and consistently reports scores in the range of 0.9, whereas the average AUPRC correlates more strongly with the resulting F1 scores, better reflecting the complexity of each ICD-9 label.

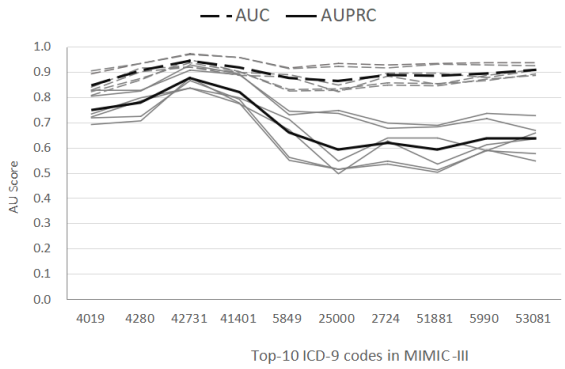

Figure 3. AUC and AUPRC scores across the top-10 most frequent ICD-9 codes in MIMIC-III. The average AUC and PRC across all labels is highlighted

We calculated the derivative of the linear function that aligns the percent distribution of the positive class against AUC and AUPRC. From AUPRC, the derivative resulted -0.00794, which is consistent – the more imbalanced, the more difficult the classification task and the lower the AUPRC. In contrast, from AUC, the derivative resulted 0.00085, which is almost a flat line, but still increasing, with AUC score slightly better, even though the imbalanced class makes the task harder.

6. Discussion

Natural Language Processing (NLP) techniques are key components for unlocking clinical evidence from EHR notes. Clinical NLP tools aim to enhance productivity, provide accessibility, and ultimately improve health systems by extracting clinical concepts from free text using distinct approaches, including lexical and semantic matching, machine learning, and rule-based systems. Those clinical concepts are defined and standardized in terminologies.

Our primary goal in this research was to figure out the differences of performance between the baseline methods and the more sophisticated approaches from the literature. It was quite a surprise that we achieved competitive results when working with these baseline methods. We had expected to understand what the actual contribution the so-called language models would make compared to the simple baseline models, but we found that the results reported from more sophisticated models were at least questionable.

6.1 Vocabulary importance

The vocabulary used in establishing baselines for machine learning approaches to identify ICD codes in the healthcare domain holds significant importance. ICD codes represent a standardized system for classifying and categorizing medical diagnoses, procedures, and conditions. To accurately identify these codes, machine learning models rely on the vocabulary of the set of terms and concepts present in the data. A comprehensive and domain-specific vocabulary is crucial for achieving accurate and reliable results. It ensures that the models can effectively capture the nuances and intricacies of clinical terminology, medical conditions, and procedures. By encompassing a wide range of relevant terms, the vocabulary enables the baseline models to make meaningful associations and predictions. Oppositely, insufficient or incomplete vocabulary can lead to suboptimal performance, as models may have difficulty in recognizing and interpreting specific medical concepts or terms. This can lead to classification errors, inaccurate predictions, and reduce overall effectiveness of machine learning approaches.

Therefore, when establishing baselines for machine learning approaches to ICD code identification, it is vital to develop and utilize a robust and comprehensive vocabulary that encompasses the diverse terminology used in the healthcare domain. A well-constructed vocabulary enhances models' understanding of medical texts, facilitates accurate code assignment, and contributes to the overall reliability and effectiveness of the machine learning-based identification process.

6.2 Word sparsity

The BoW (Bag-of-Words), CUI (Concept Unique Identifier), and DSyn (Disease and Symptoms) datasets each consist of nine subsections. This yields a total

number of 27 datasets, resulting in a combined number of 81 subsections for all three sets. The initial subsection in each set has the highest term density, while subsequent subsections become increasingly sparse, with the ninth subsection having the lowest term density. The density of terms in these datasets is determined by the frequency of words or concepts present to them.

To assess the impact of term density on the accuracy of ICD-9 code prediction, the BoW dataset was divided into three subsets, with each subset grouping lemmas according to their frequency, i.e., by the number of discharge notes in which each lemma is observed.

The goal was to systematically evaluate the importance of different feature frequencies when designing the baselines models, and to investigate whether the models would achieve higher or lower accuracy in predicting ICD codes when confronted with data sets that contain higher or lower term frequencies.

It was recognized that stop words are insignificant in establishing clinical labels, and it was observed that the subset composed with less frequent lemmas, despite being sparser, contains more meaningful clinical terms and has less noise from irrelevant data. However, it was found that relying solely on the frequency subsets did not yield satisfactory model performance. Therefore, we found that employing the full feature set produced the most favorable outcomes.

6.3 Area under 'what '?

AUC measures the ability of the model to distinguish between positive and negative samples, while AUPRC (area under the precision-recall curve) emphasizes the model's ability to rank positive samples higher than negative samples, which is crucial when dealing with imbalanced datasets.

AUC assesses the model's discriminative power, measuring the probability that a randomly selected positive instance has a higher predicted probability than a randomly selected negative instance. Its interpretation is intuitive and ranges from 0 to 1, with higher values indicating better performance and enabling easy model comparisons [74]. In addition, AUC is less affected by skewed class distributions and allows a comprehensive evaluation of the model's performance at all thresholds. Several studies highlight the effectiveness of AUC in imbalanced datasets [75], [76].

Conversely, AUPRC is a performance metric that focuses on the trade-off between precision and recall, which is particularly relevant for imbalanced datasets. When positive classes are rare, AUPRC is sensitive to changes in callbacks and puts more weight on the performance of positive classes. This becomes critical when the cost of false negatives is high, such as in medical diagnosis, fraud detection, or rare disease prediction [77], [78]. Therefore, AUPRC provides a more meaningful measure of the model's performance in distinguishing between the positive and negative classes, especially when the prevalence of the positive class is low.

In summary, both AUC and AUPRC are valuable evaluation metrics for imbalanced datasets, each providing unique insights into model performance. If the goal is to focus on overall classification performance, especially in scenarios where negative classes dominate, AUC is an appropriate choice. However, in cases where positive classes are few and correct identification is critical, more attention should be paid to AUPRC due to its recall sensitivity.

There are several other studies [79]-[81] that explore the benefits of using AUPRC in cases of highly imbalanced datasets and advocate its adoption to overcome the limitations of AUROC/AUC in such cases, highlighting the importance of considering the trade-off between precision and recall, and the sensitivity of AUPRC to rare positive cases, which makes it a more informative and appropriate metric for evaluating the performance of models on imbalanced data. We agree with those claims, and also recommend using AUPRC to better understand the performance of the model when making informed decisions.

6.4 Supporting ICD coding

Terming, coding, and grouping terminologies were designed for different purposes and audiences, which entails differing structures. From terming to grouping, the degree of clinical specificity gradually decreases, abstracting the meaning of concepts and requiring increasingly complex analysis from lexicon to semantics, context, and reasoning. For this reason, although terming to coding mappings has been deemed as an important resource for retaining clinical detail after coding - as essential meanings are retained - the loss of context precludes the application of an automatic approach. Mapping can be considered as another useful source of information to support systems based on other perspectives that focus on terming only when a high level of clinical detail is needed. An effective coding system could analyze all the necessary textual information and in turn seize the information coming from terming.

Regarding the base layer in Figure 1, lexical-based tools can automatize the terming task by identifying terminology concepts from EHR documents, such as Bio-Yodie [82] and cTAKES (clinical Text Analysis and Knowledge Extraction System) [83]. A lexical NLP approach is overall able to highlight predefined dictionary-based keywords and expressions in an exact matching way, not necessarily aware of either context and semantics, or correlations between the annotated more specific coding terminology. The output cardinality of a lexical approaches is usually token- or keyword-centered.

Within the intermediate layer, coding raises as a more challenging task. Corresponding terminologies are presented in a hierarchical structure and using more

complex descriptions, including abstract concepts that encompass multiple meanings. Thus, we propose three different perspectives when designing more sophisticated ML applications focused on recognizing ICD codes in text, each one playing a cumulative role in supporting the coding process: a) semantic relatedness, b) contextualized, and c) advisory perspectives.

6.4.1 Semantic relatedness perspective

This is a partially sentence-centered approach that aims to identify ICD references taking the meanings of words into account to get higher-level annotations than those produced by purely lexical-based approaches. A semantic relatedness perspective is not focused on the final coding result, and does not consider context when providing annotations. This task works like a robot, as it is not allowed to make any clinical assumptions.

The proposed constraint for a semantically related approach includes the proper choice of the related ICD code for each given piece of text (multi-word expression), considering the surrounding words within the same sentence when reasoning about either the sense or meaning of terms. Thus, NLP applications following this role can identify concepts related to ICD codes based on the semantic relatedness of textual expressions. The output cardinality is expression-centered with possibly multiple alternative ICD codes presented to the same expression. No disambiguation is performed in this role, i.e., each diagnosis is linked to a specific part of a document, and the same expression may have multiple ICD correspondences.

A semantic relatedness approach focuses on generating low-level features to support either (a) clinicians in validation, (b) a subsequent high-level approach, or even (c) coders by capturing and presenting clinical concepts that may be related to the final ICD diagnoses. For example, some sentences or expressions associated with code R50.9 (Fever, unspecified) could be presented as follows:

(E1) the patient is admitted with a high fever
(E2) patient with pyrexia
(E3) three episodes of fever during pregnancy
(E4) patient does not exhibit hyperthermia
(E5) patient had constant fevers last year

Although this perspective may have some resemblance to the mapping approach in Section 3.3, the results differ. The mapping approach is based on collecting ICD candidates by combining only essential clinical meanings, whereas an approach based on the semantic relatedness perspective also deals with non-clinical meanings.

6.4.2 Contextualized perspective

This is a document-centered approach that takes multiple contextual components from text into account, including: a) negations, b) temporality, and c) experiencers. Context was explored using lexical-based approaches. However, the complexity of how ICD references can be found in text could possibly drive this context-based approach to be more difficult. An ICD code consists of pieces of meaning, so if one isolated piece of text produces a code, maybe another piece of text in a different location contributes to changing the whole meaning and producing a different code. Similar to semantic relatedness, this perspective does not focus on the final coding result. However, semantic assumptions are made to determine whether an ICD reference is contextualized in terms of (a) a positive or negative mention, (b) a current, historical, or hypothetical temporal reference, or (c) a patient or relative experience. Context plays a role in this annotation perspective that can help coders identify the final diagnosis for a patient.

This approach is constrained by the selection of those text that correspond to ICD codes most closely related to the current diagnosis (or symptoms). It is allowed to make assumptions in terms of context to filter out those references that are not directly related to the patient, given in a negative expression (negations), and related to past or future (hypothetical) references in time. NLP design for this perspective is required to collect contextualized evidence in the text that can support the diagnosis. The output cardinality is document-centered with potentially multiple diagnoses per document. However, each identified possible diagnosis can be linked to a specific part of a document (sentence, partial sentence, multi-word phrases or even a single meaningful keyword).

Considering the same examples presented in the semantic relatedness perspective, E1 and E2 could be associated with ICD-10 code R50.9. However, for example E3, more contextual data are needed to determine whether it is a current pregnancy or it refers to the patient's prenatal period in the past.

6.4.3 Advisory perspective

This perspective is aware of the larger picture and is the closest to mimicking the coder task. In a consultative role, an NLP application must identify the most likely diagnoses and symptoms for a given patient, considering all documents regarding the patient's stay or course. This annotation perspective works as a coder: rather than generating text annotations, a consultative approach produces patient-centred diagnostic candidate. The approaches discussed in Section 3.2 mostly use machine learning techniques due to the complexity of this task.

The consultative approach is not constrained by pieces of text and it can possibly consider the entire patient record. The possible ICD may result from a combination of different pieces of text from different sections or even different documents. NLP consultative applications are designed to identify the most likely ICD codes that correspond to patient diagnoses, and then to be accurately checked by an experienced coder. The output cardinality

is patient-centered, with possibly multiple diagnoses per stay, in which each diagnosis is not necessarily linked to a specific part of a document. In this case, the two examples discussed previously, E1 and E2 could be potentially identified as strong candidates, depending on all other possible diagnoses and whether those symptoms are relevant to the cause of admission.

7. Conclusions

We explored the effectiveness of less complex baselines for solving the task of identifying ICD codes using patient discharge notes. Despite the widespread use of deep learning models, we demonstrate that simpler methods can achieve comparable results while requiring fewer computational resources. Our results highlight the importance of thoroughly evaluating the performance of different models and methods for different tasks to avoid over-reliance on a particular approach. By considering the importance of a vocabulary versus concept approach, researchers can improve the efficiency and practicality of their solutions while still achieving competitive results. By surpassing current baselines, the developed approach exhibits enhanced capabilities in capturing and extracting relevant information using a less computationally expensive method. This suggests that the proposed method possesses distinctive characteristics that make it well-suited for handling the intricacies and challenges inherent to healthcare data.

While deep learning methods have gained popularity and demonstrated success in various domains, their application to healthcare datasets often faces specific complexities, such as limited labeled data and the need for interpretations from medical professionals. In this regard, the results obtained highlight the potential of baseline approaches in accurately extracting valuable insights to ensure fairness when evaluating more sophisticated deep learning alternatives. These findings hold significant implications for advancing the state of the art in healthcare data analysis and decision making. By surpassing current baselines and outperforming deep learning approaches, the developed methodology contributes to the refinement of techniques that can effectively harness the potential of healthcare data sets for improved patient care, clinical decision support, and medical research.

Less frequent ICD codes tend to be more specific, and baseline models designed on such imbalanced data may have difficulty capturing the nuances and specifics of rare diseases. Consequently, models with limited training samples may struggle to generalize well to rare classes [84], and imbalances between classes may lead to biased model performance, as models may prioritize accuracy on the majority class at the expense of the rare classes [85]. With less frequent ICD codes, the data may become sparser, as there are fewer examples available for the rare classes. We plan to extend our work by designing baseline models for rare diseases with less frequent ICD codes (less than 3 percent of the data), with some important considerations and potential impacts to be aware of.

Most of the previously published coding work is based on ICD-9 evaluated on the MIMIC-III dataset. In contrast, results typically report model performance on real-world datasets that are proprietary and not publicly available. With the recent release of de-identified free text clinical notes as part of the MIMIC-IV dataset, we plan to incorporate this new dataset into your next endeavor by designing a more comprehensive hybrid approach to test on a larger dataset (MIMIC-IV), hoping to resolve coding even when ICD-9 and ICD-10 codes coexist and are concomitantly found in the same dataset.

Authors' contribution

Jessica Jha was primarily involved in examining baseline machine learning models with a specific focus on identifying ICD-9 codes within discharge notes from the MIMIC database. She also conducted a comprehensive evaluation of various models and approaches to ensure a thorough understanding of their effectiveness. Mario Almagro's contribution focused on the conception, design, and manuscript preparation of the review section about Clinical Terminologies and the comparison of the differences between terming, coding and grouping. Hegler Tissot was mainly involved in designing all the experiments, providing support to the other authors, and final manuscript preparation and review. Finally, all authors equally contributed to the preparation of the Discussion section.

Funding information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of interest

The authors declare no competing financial interests.

Consent for publication

The authors (we) fully consent and agree with this research work to be published in the Journal of Digital Health. We also agree that the text and any figures published in this article will be used only in educational publications, and as published on an open -access basis, we understand that it will be freely available on the internet and can be viewed by the general public.

References

[1] Paz KB, Halverstam C, Rzepecki AK, McLellan BN. A National Survey of Medical Coding and Billing Training in United States Dermatology Residency Programs. *Journal of drugs in dermatology*. 2018; 17(6):678-682. Available from: http://europepmc.org/abstract/MED/29879256.

[2] Dong H, Falis M, Whiteley W, Alex B, Matterson J, Ji S, Chen J, Wu H. Automated clinical coding: what, why, and where we are? *npj Digital Medicine*. 2022 5(1):159. doi: 10.1038/s41746-022-00705-7.

[3] W. H. O. *WHO, ICD-10 : international statistical classification of diseases and related health problems*. World Health Organization, 10th ed. World Health Organization, Geneva, 2004.

[4] Adams DL, Norman H, Burroughs VJ. Addressing medical coding and billing part II: a strategy for achieving compliance. A risk management approach for reducing coding and billing errors. *Journal of the National Medical Association*. 2002; 94(6):430-47. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2594405/.

[5] Raghavendra Chalapathy, Ehsan Zare Borzeshi, Massimo Piccardi. Bidirectional LSTM-CRF for clinical concept extraction. In: *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*. Osaka, Japan: The COLING 2016 Organizing Committee; 2016. p.7-12. Available from: https://aclanthology.org/ W16-4202.

[6] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad. Intelligible models for Health Care. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM; 2015. p.1721-1730. doi: 10.1145/2783258.2788613.

[7] Rasheed K, Qayyum A, Ghaly M, Al-Fuqaha A, Razi A, Qadir J. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Computers in Biology and Medicine*. 2022;149:106043. doi:10.1016/j.compbiomed.2022.106043.

[8] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*. 2017; 19(6): 1236-1246. doi:10.1093/bib/bbx044.

[9] Zhao L, Bao J, Qiao X, Jin P, Ji Y, Li Z, Zhang J, Su Y, Ji L, Shen J, Zhang Y, Niu L, Xie W, Hu C, Shen H, Wang X, Liu J, Tian J. Predicting clinically significant prostate cancer with a deep learning approach: a multicentre retrospective study. *European Journal of Nuclear Medicine and Molecular Imaging*. 2022; 50(3):727-741. doi:10.1007/s00259-022-06036-9.

[10] Weiss J, Raghu VK, Bontempi D, Christiani DC, Mak RH, Lu MT, Aerts HJWL. Deep learning to estimate lung disease mortality from chest radiographs. *Nature Communications*. 2023; 14(1): 2797. doi:10.1038/s41467-023-37758-5.

[11] Teo K, Yong CW, Chuah JH, Hum YC, Tee YK, Xia K, Lai KW. Current Trends in Readmission Prediction: An Overview of Approaches. *Arabian Journal for Science and Engineering*. 2021; 16:1-18. doi:10.1007/s13369-021-06040-5.

[12] Kessler S, Schroeder D, Korlakov S, Hettlich V, Kalkhoff S, Moazemi S, Lichtenberg A, Schmid F, Aubin H. Predicting readmission to the cardiovascular intensive care unit using recurrent neural networks. *Digital Health*. 2023; 9;9:20552076221149529. doi:10.1177/20552076221149529.

[13] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, Sundberg P, Yee H, Zhang K, Zhang Y, Flores G, Duggan GE, Irvine J, Le Q, Litsch K, Mossin A, Tansuwan J, Wang D, Wexler J, Wilson J, Ludwig D, Volchenboum SL, Chou K, Pearson M, Madabushi S, Shah NH, Butte AJ, Howell MD, Cui C, Corrado GS, Dean J. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*. 2018; 1:18. doi: 10.1038/s41746-018-0029-1.

[14] Luo J, Wu M, Gopukumar D, Zhao Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomedical informatics insights*. 2016; 8:1-10. doi: 10.4137/BII.S31559.

[15] Cowie JM, Wanger KM, Cartwright A, Bailey H, Millar JA, Price S, Henry M. A review of Clinical Terms Version 3 (Read Codes) for speech and language record keeping. *International Journal of Language & Communication Disorders*. 2001; 36(1): 117-126. doi:10.1080/13682820150217608.

[16] Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*. 2008; 77(5): 291-304. doi: 10.1016/j.ijmedinf.2007.09.001.

[17] Stuart-Buttle CD, Read JD, Sanderson HF, Sutton YM. A language of health in action: Read Codes, classifications and groupings. *Proceedings : a conference of the American Medical Informatics Association*. AMIA Fall Symposium; 1996. p. 75-79.

[18] Vardy DA, Gill RP, Israeli A. Coding medical information: classification versus nomenclature and implications to the Israeli medical system. *Journal of Medical Systems*. 1988; 22(4): 203-210. doi:10.1023/A:1022643216122.

[19] Read JD, Sanderson HF, Drennan YM. Terming, encoding, and grouping. *MEDINFO*. 1995; 8(1):56-64.

[20] Mannion R, Marini G, Street A. Implementing payment by results in the English NHS: changing incentives and the role of information. *Journal of Health Organization and Management*. 2008; 22(1): 79-88. doi:10.1108/14777260810862425

[21] De Silva TS, MacDonald D, Paterson G, Sikdar KC, Cochrane B. Systematized nomenclature of medicine clinical terms (SNOMED CT) to represent computed

tomography procedures. *Computer Methods and Programs in Biomedicine.* 2011; 101(3):324-329. doi:10.1016/j.cmpb.2011.01.002.

[22] Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute CG, Warren J. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. *Journal of the American Medical Informatics Association*. 1997; 4(3): 238-251. doi: 10.1136/jamia.1997.0040238.

[23] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*. 2004; 32(1): D267-D270. doi: 10.1093/nar/gkh061.

[24] H. S. C. I. C. HSCIC, *OPCS Classification of Interventions and Procedures Version 4.7 combined Volumes I & II / Health and Social Care Information Centre*, 4th ed. TSO (The Stationery Office), 2014.

[25] Robert B. Fetter. Diagnosis related groups: Understanding hospital performance. *Interfaces*.1991; 21(1): 6-26. doi: http://dx.doi.org/10.1287/inte.21.1.6.

[26] Street A, Dawson D. Costing hospital activity: the experience with healthcare resource groups in England. *The European Journal of Health Economics.* 2002; 3(1): 3-9. doi: 10.1007/s10198-001-0086-1.

[27] Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J*ournal of biomedical informatics.* 2009; 42(5):760-772. doi: 10.1016/j.jbi.2009.08.007.

[28] Chenwei Yan, Xiangling Fu, Xien Liu, Yuanqiu Zhang, Yue Gao, Ji Wu, Qiang Li. A survey of automated International Classification of Diseases coding: development, challenges, and applications. *Intelligent Medicine*. 2022; 2(3):161-173. doi: 10.1016/j.imed.2022.03.003.

[29] Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*. 2004; 11(5): 392-402. doi: 10.1197/jamia.M1552

[30] M. T. Chiaravalloti, R. Guarasci, V. Lagani, E. Pasceri, R. Trunfio. A Coding Support System for the ICD-9-CM Standard. *2014 IEEE International Conference on Healthcare Informatics*. Verona, Italy. 2014; p. 71-78, doi: 10.1109/ICHI.2014.17.

[31] S. G. Rizzo, D. Montesi, A. Fabbri, G. Marchesini. Icd code retrieval: Novel approach for assisted disease classification. In: I*nternational Conference on Data Integration in the Life Sciences*. Springer. 2015; p. 147-161.

[32] D. Zhang, D. He, S. Zhao, L. Li. Enhancing automatic icd-9-cm code assignment for medical texts with pubmed. BioNLP, *Association for Computational Linguistics*. 2017; p. 263-271. doi:10.18653/v1/W17-2333.

[33] Chen Y, Lu H, Li L. Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity. *PloS One*. 2017; 12(3): e0173410. doi: 10.1371/journal.pone.0173410.

[34] Ning W, Yu M, Zhang R. A hierarchical method to automatically encode Chinese diagnoses through semantic similarity estimation. *BMC Medical Informatics and Decision Making*. 2016; 16(1): 30. doi: 10.1186/s12911-016-0269-4.

[35] Damla Arifoğlu, Onur Deniz, Kemal Aleçakır, Meltem Yöndem. Codemagic: semi-automatic assignment of icd-10-am codes to patient records. In: *Information Sciences and Systems 2014*. Springer, 2014; p. 259-268.

[36] Sheng-Wei Chen, Po-Ting Lai, Yi-Lin Tsai, Jay Kuan-Chieh Chung, Sherry Shih-Huan Hsiao, Richard Tzong-Han Tsai. NCU IISR System for NTCIR-11 MedNLP-2 Task. In: Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies. *National Institute of Informatics*. Tokyo, Japan. 2014; 9-12.

[37] S. Boytcheva. Automatic matching of icd-10 codes to diagnoses in discharge letters. In: *Proceedings of the Second Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics. 2011; p. 11-18.

[38] P. Zweigenbaum and T. Lavergne. Hybrid methods for icd-10 coding of death certificates. In: *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics. 2016;p. 96-105. doi: 10.18653/v1/W16-6113.

[39] P. Jatunarapit, K. Piromsopa, and C. Charoeanlap. Development of thai text-mining model for classifying icd-10 tm. In: *2016 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*. Ploiesti, Romania. IEEE. 2016; p. 1-6. doi: 10.1109/ECAI.2016.7861163.

[40] J. Seva, M. Kittner, R. Roller, and U. Leser. Multilingual icd-10 coding using a hybrid rule-based and supervised classification approach at clef ehealth 2017. In: *Conference and Labs of the Evaluation Forum (Working Notes)*. 2017.

[41] E. M. Van Mulligen, Z. Afzal, S. Akhondi, D. Dang, and J. Kors. Erasmus mc at clef ehealth 2016: Concept recognition and coding in french texts. In: *Conference and Labs of the Evaluation Forum (Working Notes)*. 2016.

[42] Schmidt D, Budde K, Sonntag D, Profitlich HJ, Ihle M, Staeck O. A novel tool for the identification of correlations in medical data by faceted search. *Computers in biology and medicine*. 2017; 85: 98-105. doi: 10.1016/j.compbiomed.2017.04.011.

[43] L.-M. Ho-Dac, C. Fabre, A. Birski, I. Boudraa, A. Bourriot, M. Cassier, L. Delvenne, C. Garcia-Gonzalez, E.-B. Kang, E. Piccinini et al. Litl at clef ehealth 2017: automatic classication of death reports.

In: *CLEF eHealth 2017*, 2017.

[44] M. Subotin and A. Davis. A system for predicting icd-10-pcs codes from electronic health records. In: *Proceedings of BioNLP. 2014*: 59-67. doi: 10.3115/v1/W14-3409.

[45] Z. Miftahutdinov and E. Tutubalina. Kfu at clef ehealth 2017 task 1: Icd-10 coding of english death certificates with recurrent neural networks. In: *Conference and Labs of the Evaluation Forum (Working Notes)*. 2017.

[46] Byung-Hak Kim, Varun Ganapathi. Read, attend, and code: Pushing the limits of medical codes prediction from clinical notes by machines. In: *Proceedings of the 6th Machine Learning for Healthcare Conference*. 2021; 149: 196-208. Available from: https://proceedings.mlr.press/v149/kim21a.html.

[47] Jinmiao Huang, Cesar Osorio, Luke Wicent Sy. An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes. *Computer Methods and Programs in Biomedicine*. 2019; 177:141-153. doi: 10.1016/j.cmpb.2019.05.024.

[48] J. Edin, A. Junge, J. D. Havtorn, L. Borgholt, M. Maistro, T. Ruotsalo, and L. Maaløe. Automated medical coding on MIMIC-III and MIMIC-IV: A critical review and replicability study. In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2023. doi: 10.1145/3539618.3591918.

[49] R. Kavuluru, A. Rios, and Y. Lu. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial Intelligence in Medicine*. 2015; 65(2):155-166. doi: 10.1016/j.artmed.2015.04.007.

[50] L. Liu, O. Perez-Concha, A. Nguyen, V. Bennett, and L. Jorm. Hierarchical label-wise attention transformer model for explainable ICD coding. *Journal of Biomedical Informatics*. 2022; 133: 104161. doi: 10.1016/j.jbi.2022.104161

[51] A. H. Peden. An overview of coding and its relationship to standardized clinical terminology. *Topics in Health Information Management*. 2000; 21(2):1-9.

[52] J. R. Campbell, H. Brear, R. Scichilone, S. White, K. Giannangelo, B. Carlsen, H. R. Solbrig, and K. W. Fung. Semantic interoperation and electronic health records: context sensitive mapping from snomed ct to icd-10. *Studies in Health Technology and Informatics*. 2013; 192: 603-607.

[53] J. A. Feinstein, P. J. Gill, and B. R. Anderson. Preparing for the international classification of diseases, 11th revision (ICD-11) in the US health care system. *JAMA Health Forum*. 2023; 4(7): e232253. doi: 10.1001/jamahealthforum.2023.2253.

[54] W. R. Hersh, M. G. Weiner, P. J. Embi, J. R. Logan, P. R. Payne, E. V. Bernstam, H. P. Lehmann, G. Hripcsak, T. H. Hartzog, J. J. Cimino, and J. H. Saltz. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical Care*. 2013; 51: S30–S37. doi: 10.1097/mlr.0b013e31829b1dbd.

[55] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*. 2016; 6(1). doi:10.1038/srep26094.

[56] A. Johnson, T. Pollard, and R. Mark. Mimic-iii clinical database. PhysioNet. 2016. doi: 10.13026/cd7z-wg25.

[57] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*. Wiley, 2013. doi: 10.1002/9781118548387.

[58] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016: 785-794. doi: 10.1145/2939672.2939785.

[59] Z. Tran, A. Verma, T. Wurdeman, S. Burruss, K. Mukherjee, and P. Benharash. ICD-10 based machine learning models outperform the trauma and injury severity score (TRISS) in survival prediction. *PLOS ONE*. 2022; 17(10): e0276624. doi: 10.1371/journal.pone. 0276624.

[60] Z. Tran, W. Zhang, A. Verma, A. Cook, D. Kim, S. Burruss, R. Ramezani, and P. Benharash. The derivation of an international classification of diseases, tenth revision-based trauma-related mortality model using machine learning. *The journal of trauma and acute care surgery*. 2022; 92(3): 561-566. doi: 10.1097/TA.0000000000003416.

[61] A. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. Proceedings. *AMIA Symposium*. 2001:17-21.

[62] X. Schmitt, S. Kubler, J. Robert, M. Papadakis, and Y. LeTraon. A replicable comparison study of ner software: Stanfordnlp, nltk, opennlp, spacy, gate. In: *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. Granada, Spain. 2019:338-343. doi: 10.1109/SNAMS.2019.8931850.

[63] V. Balakrishnan and L.-Y. Ethel. Stemming and lemmatization: A comparison of retrieval performances. *Lecture Notes on Software Engineering*. 2014; 2(3): 262-267. doi:10.7763/lnse.2014.v2.134.

[64] S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. In: *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg. 2004:22-30. doi: https://doi.org/10.1007/978-3-540-24775-3 5

[65] H.-F. Yu, P. Jain, P. Kar, and I. Dhillon. Large-scale multi-label learning with missing labels. In: *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine

Learning Research. E. P. Xing and T. Jebara, Eds. Bejing, China. 2014; 32(1): 593-601. Available from: https://proceedings.mlr.press/v32/yu14.html.

[66] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 2002;16: 321-357. doi:10.1613/jair.953.

[67] K. Duan, S. Keerthi, and A. N. Poo. Evaluation of simple performance measures for tuning SVM hyperparameters. *Neurocomputing*. 2003;51:41-59. doi: 10.1016/s0925-2312(02)00601-x.

[68] D. Pascual, S. Luck, and R. Wattenhofer. Towards BERT-based automatic ICD coding: Limitations and opportunities. In: *Proceedings of the 20th Workshop on Biomedical Language Processing*. Association for Computational Linguistics. 2021:54-63. Available from: https://aclanthology.org/2021.bionlp-1.6.pdf.

[69] F. Li and H. Yu. ICD coding from clinical text using multi-filter residual convolutional neural network. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020;34(05):8180-8187. doi: 10.1609/aaai.v34i05.6331.

[70] Z. Yang, S. Wang, B. P. S. Rawat, A. Mitra, and H. Yu. Knowledge injected prompt based fine-tuning for multi-label few-shot ICD coding. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.2022: 1767-1781. Available from: https://aclanthology.org/2022.findings-emnlp.127.

[71] K. Xu, M. Lam, J. Pang, X. Gao, C. Band, P. Mathur, F. Papay, A. K. Khanna, J. B. Cywinski, K. Maheshwari, P. Xie, and E. P. Xing. Multimodal machine learning for automated icd coding. In: *Proceedings of the 4th Machine Learning for Healthcare Conference, ser. Proceedings of Machine Learning Research*. F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, Eds. 2019;106: 197-215. Available from: https://proceedings.mlr.press/v106/xu19a.html.

[72] C. Song, S. Zhang, N. Sadoughi, P. Xie, and E. Xing. Generalized zero-shot text classification for ICD coding. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization*. 2020: 4018-4024. doi: 10.24963/ijcai.2020/556.

[73] H. Schafer and C. M. Friedrich. UMLS mapping and word embeddings for ICD code assignment using the MIMIC-III intensive care database. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Berlin, Germany. IEEE. 2019: 6089-6092. doi: 10.1109/embc.2019.8856442.

[74] J. Huang and C. Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*. 2005; 17(3): 299-310. doi: 10.1109/TKDE.2005.50.

[75] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*. 2001; 42(3) : 203-231. doi: 10.1023/A:1007601015854.

[76] J. Davis and M. Goadrich. The relationship between precision-recall and ROC curves. In: *Proceedings of the 23rd international conference onMachine learning-ICML'06*. ACM Press. 2006. doi:10.1145/1143844.1143874.

[77] T. Saito and M. Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*. 2015;10(3):e0118432. doi: 10.1371/journal.pone.0118432.

[78] P. A. Flach and M. Kull. Precision-recall-gain curves: Pr analysis done right. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*. 2015;1(NIPS'15): 838-846. Cambridge, MA, USA: MIT Press.

[79] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*. 2004; 6(1): 20-29. doi: 10.1145/1007730.1007735.

[80] L. A. Jeni, J. F. Cohn, and F. D. L. Torre. Facing imbalanced data–recommendations for the use of performance metrics. In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 2013: 245-251. doi: 10.1109/acii.2013.47.

[81] T. Saito and M. Rehmsmeier. Precrec: fast and accurate precision–recall and ROC curve calculations in r. *Bioinformatics*. 2016; 33(1): 145-147. doi: 10.1093/bioinformatics/btw570.

[82] G. Gorrell, X. Song, and A. Roberts. Bio-yodie: A named entity linking system for biomedical text. *arXiv preprint*. 2018. doi: 10.48550/arXiv.1811.04860.

[83] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*. 2010; 17(5): 507-513. doi: 10.1136/jamia.2009.001560.

[84] M. Längkvist, L. Karlsson, and A. Loutfi. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*. 2014; 42: 11-24. doi:10.1016/j.patrec.2014.01.008.

[85] H. He and Y. Ma, Eds. Imbalanced learning: Foundations, Algorithms, and Applications. Hoboken, NJ: Wiley-Blackwell. 2013. doi:10.1002/9781118646106.