Evaluating Clinical Coding Alignment Across ICD-9, ICD-10, and ICD-11 with a Semantic Retrieval Approach

Irene Siragusa[1,*], Roberto Pirrone[1]

[1]Department of Engineering, University of Palermo, Palermo, 90128, Sicily, Italy

Abstract

In this article, we address the Clinical Coding Alignment problem in the three most recent revisions of the International Classification of Diseases (ICD), namely ICD-9, ICD-10, and ICD-11. Using the clinical records in MedPix 2.0 data set, we automatically mapped the manually annotated diseases, reported in a free-textual form, to standard ICD formats. Our approach combines an exact-match strategy with a semantic retrieval pipeline that couples cosine similarity with a BERT-F1 Score as re-ranking mechanism. We evaluated three retrieval queries ranging from isolated disease names to general disease definitions and the full clinical reports. Furthermore, we conducted an open-evaluation analysis to observe how these distinct coding systems handle free-form clinical text. Our findings indicate that the updated taxonomy of ICD-11 aligns more naturally with real-world clinical narratives than its predecessors and that the definition-based query achieves the best overall performance.

Keywords

LLMs, Clinical Coding, ICD Classifications, Semantic Retrieval

1. Introduction

In everyday clinical practice, physicians are asked to generate medical reports, discharge summaries, and any other clinical notes in the Electronic Health Record (EHR). These documents are often written in a free-form style, in which terms related to diseases, body parts, or diagnostic tests follow a classification system derived from common practice in the specific medical institution rather than standard terminology.

To ensure both standardisation in medicine at the global level and to provide interoperability in the recording, reporting, and monitoring of diseases, the World Health Organisation (WHO) has implemented and currently maintains a standardisation procedure in the International Classification of Diseases (ICD). ICD serves to record and report health and health-related conditions globally, ensuring interoperability of digital health data [2]. Moreover, ICD has been continuously revised through the years, reaching the ninth (ICD-9) [3], the tenth (ICD-10) [4] and the eleventh (ICD-11) [2] versions, and translated in more than 40 languages. Although the 11th revision was adopted in 2019, and formally came into effect in 2022, when member states are expected to use the most current version of the ICD for reporting death and disease statistics, ICD-11 is not widely used and some institutions still use ICD-10 or ICD-9 for clinical coding.

In the context of this paper, we analyse how a simple alignment strategy behaves across the various ICD versions. In particular, we used an exact-match strategy and a retrieval pipeline across the three target taxonomies [5]. In the proposed approach, plausible candidates are first retrieved following a similarity-based criterion, and then re-ranked according to their BERT-F1 score metric [6] with the annotated value. Three different retrieval queries have been tested, each of them providing a different amount of information, from the disease name to its general definition and the full clinical report. As an evaluation data set, we use MedPix 2.0 [7], which collects clinical reports with annotations on the diagnosed diseases.

This work is structured as follows. Section 2 describes the ICD taxonomies used in this paper. We

NL4AI 2026: Ninth Workshop on Natural Language for Artificial Intelligence, October 6-9th, 2026, Perugia, Italy [1]

*Corresponding author.

irene.siragusa02@unipa.it (I. Siragusa); roberto.pirrone@unipa.it (R. Pirrone)

0009-0005-8434-8729 (I. Siragusa); 0000-0001-9453-510X (R. Pirrone)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-05

CEUR-WS.org/Vol-4315/paper1.pdf

describe the data set used and the alignment strategy in Section 3. We discuss our findings in Section 4 and present concluding remarks in Section 5.

2. ICD Taxonomy

The International Classification of Diseases (ICD) is designed to promote international comparability in the collection, processing, classification, and presentation of mortality statistics. Moreover, ICD classification aims to standardise in a unique taxonomy the medical terminology, thus allowing worldwide comparison and information sharing such as health trends and statistics. ICD codes have been continuously revised through the years by WHO which released the 9th, 10th and 11th versions. All taxonomies follow the same structure in which diseases are organised in chapters, sections, and subsections, each disease is associated with an unique alphanumeric code, and the entire coding system is organised hierarchically [2].

ICD-9 was published in 1977 after it started developing in 1975 [3]. It was further extended in its Clinical Modification version (ICD-9-CM) which consists of three volumes that collect diagnosis codes and procedural codes. It collects more than 14,500 diseases, which are organised in 19 chapters, two of which cover supplementary codes for Classification Of Factors Influencing Health Status And Contact With Health Services and for Classification Of External Causes Of Injury And Poisoning. Each disease is coded with 3-digits and additional specialisations are coded with increased granularity with 4-digit and 5-digit codes.

Developing of **ICD-10** started in late 80s and its ICD-10-CM version was then introduced for its annual process of review in 1992 [4]. ICD-10 is much more granular than ICD-9, with an expansion from 17,000 codes to approximately 155,000, arranged in 22 chapters. The internal coding system follows the same structure as ICD-9: the first 3 digits in both represent common traits, with each subsequent character providing greater specificity. The ICD-10-CM is alphanumeric, with a possible 7 digits of specificity as opposed to the 5 digits of ICD-9 [4].

ICD-11 is the latest revision that was released in 2018 and officially came into effect in 2022 [2]. ICD-11 is more extensive and has greater implications for what can be done with the ICD, and how, than any revision since the 6th in 1948. It is organized into 28 chapters providing substantial improvement to the previous revisions by adding references to Traditional Medicine (Chapter 26) and supplementary sections for functioning assessment and additional extension codes, in Chapters 27 and 28. The ICD-11 Mortality and Morbidity Statistics (MMS) is the reference ICD-11 coding and contains nearly 85,000 entities, in which also residual categories, referred to as "Other specified" and "Unspecified" diseases, are kept. Coding system remains alphanumerical and follows a hierarchical structure.

In medical practice, the Clinical Coding task can be manually performed by specialised clinicians, but recent developments in Artificial Intelligence (AI) led to the development of Automatic Clinical Coding techniques. The Clinical Coding task can be formalised as a multi-class classification problem on a given clinical note, which can be addressed with both traditional classifiers based on Convolutional Neural Networks [8, 9, 10] or Graph Neural Network [11, 12] and using pre-trained language models [13]. All these strategies require a training step that is avoided in the approach proposed in this work, as Large Language Models (LLMs) are inferred with a semantic retrieval approach [5].

3. Methodology

The objective of Clinical Coding is to report the diagnosed disease, annotated in a free textual form to its standard form in ICD. For example, the clinical note, which reports *Vertebral Artery Dissection* as its manually annotated disease, is a non-standard term in ICD terminology as an exact match cannot

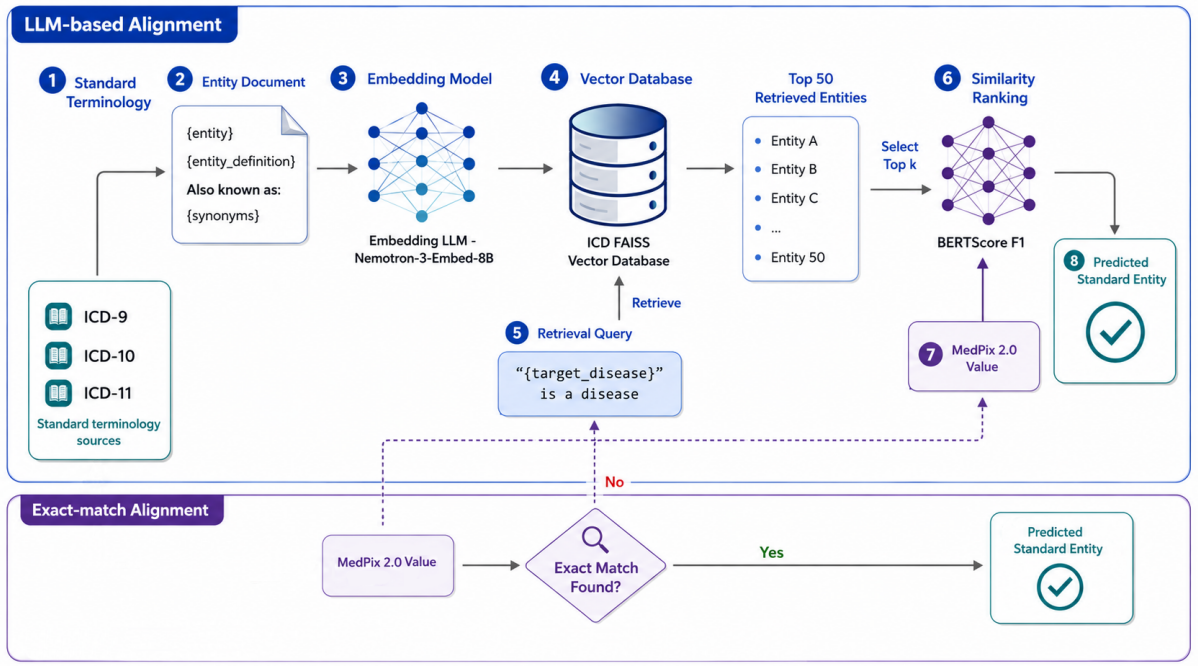


Figure 1: Schematic overview of the coding alignment strategy.

be found. In contrast, *Dissection of vertebral artery*, *Aneurysm and dissection of vertebral artery* and *Aneurysm or dissection of vertebral artery* can be found as its corresponding and valuable ICD standard terms in 9th, 10th and 11th versions respectively.

The methodology proposed in this work relies on the application of the alignment procedure as in [5], which is schematically reported in Figure 1 and follows a Retrieval Augmented Generation (RAG) approach [14].

Initially, exact-string matching is applied to identify any direct correspondence between the free-text entity and an entry in the standard terminology. In doing this, the alignment is performed directly, and both the standard term and its code can be associated with the target report. If direct correspondence fails, a query is automatically generated to perform a cosine similarity search, extracting the top-k candidates from the pre-build vector database, collecting documents representing the disease in ICD. From this pool, the entity that maximises the BERT-F1 score [6] is selected as the representative standardised term, following a re-ranking approach. This score is calculated only between the target term and the retrieved one, and, if present, any of its synonyms. On doing this, we ensure that the re-ranking, and the subsequent selection of the standard term is not influenced by any other vocabularies. In the pure retrieval phase, on the contrary, synonyms or definitions can more effectively guide the retrieval phase, which keeps the computational effort low, compared to the usage of BERT-score metrics.

A vector database has been created for each ICD disease corpora and for each entity an *ad hoc* document was created including the entity definition and its synonyms if present. Only disease terms have been selected and included in the vector database. The raw ICD tables for the 10th and the 11th revision were retrieved via ICD-API[1]. In particular, ICD-10-CM in its 2019 version was selected for ICD-10 and ICD-11-MMS of 2025 for ICD-11. For ICD-9, the latest 2014 ICD-9-CM version has been downloaded[2].

In the context of this work, the system has been modified from the original one as Nemotron-3-Embed-8B-BF16 [15] was used as Embedding model for the vector databases. This change was made because this model reached, in July 2026, the best performance in the Retrieval Embedding Benchmark (RTEB) [16], as well as in RTEB Healthcare split among the open-source models. Other parts of the

[1] https://icd.who.int/docs/icd-api/APIDoc-Version2/

[2] https://www.cms.gov/medicare/coding-billing/icd-10-codes/icd-9-cm-diagnosis-procedure-codes-abbreviated-and-full-code-titles

pipeline have been left unchanged, as the FAISS vector database [17] and BioSyn [18] have been used for the calculations of BERT Score. BioSyn [18], is a fine-tuned version of BioBERT [19] trained on biomedical data and disease synonym corpora [20], making it the most valuable model for the scope of this pipeline. We considered evaluation driven by BERT Score metric since it offers an overall good comparison in the embedding space among the two candidate terms, compare with string-matching or n-grams based strategies.

We considered the MedPix 2.0 data set [7], which collects nearly 700 clinical reports derived from MedPix portal[3] as benchmark data set. These documents are arranged in a JSON format in which both demographic and verbose health data about the patient are reported, along with the final diagnosed disease and the macro category to which it belongs.

We designed three retrieval query strategies, namely *Simple*, *Definition*, and *Report*, where different amounts of information are provided from the clinical case, as in 1.

Table 1
Structure of the three retrieval queries.

Strategy	Retrieval query
Simple	`<target_disease>` is a disease.
Definition	`<target_disease>` is a disease. `<disease discussion>`.
Report	`<target_disease>` is a disease. `<patient_history>`, `<exam_notes>`, `<diagnosis>`, `<disease_discussion>`, `<differential_diagnosis>`, `<treatment>`.

The *Simple* strategy relies exclusively on the target disease entity, whereas the *Definition* query supplements this entity with general background information about the pathology. Neither strategy incorporates patient-specific data, thus isolating the alignment between disease nomenclature and standard ICD terminology. In contrast, the *Report* strategy incorporates comprehensive clinical details, including patient history, diagnostic findings, differential diagnoses, and treatment plans. Therefore, the system effectively emulates an expert annotator operating with complete case knowledge. In section 4, we will focus more on the simple strategy, as it better emulates the agnostic alignment behaviour from a free-textual annotation without additional details, to standard terminology.

We extend the standardisation procedure from the target disease to its belonging macro category, thus their corresponding term according to ICD taxonomy can be both identified. To assess a higher level of accuracy, only terms in all ICDs taxonomies associated with a 3-digit code and not marked as higher level terms, such as block or chapter names, have been compared with the disease field. As for the macro disease field, two experiments have been conducted: on one hand, the same retriever base as the simple disease have been used (macro diseases) and, on the other hand, its restriction to the proper macro chapter and blocks in ICD (macro category diseases). This double analysis has the scope to evaluate how the clinical narratives align to the standard coding.

The experiments were executed on single NVIDIA A100 64GB GPUs on the Leonardo supercomputer [21].

4. Results

In Table 2 are reported the average retrieval scores expressed in terms of shifted cosine similarity, thus ranging in [0,2], with the simple retrieval query. Starting from the top 50 retrieved items, we perform an ablation study on the optimal number of retrieved terms, evaluating the impact of the entire pipeline at different retrieval thresholds, namely selecting the first 5%, 10%, 20%, 50% and 100% retrieved elements. As aforementioned, we considered the evaluation settings for the macro disease fields: columns marked

[3] https://medpix.nlm.nih.gov/home

as *macro diseases* consider the same retrieval space as for the disease, while in *macro category diseases* the retrieval space is restricted to the items which are marked as chapters or blocks in ICD taxonomy. In Table 3 the average BERT-F1 scores obtained between the target term in the MedPix data set and the top retrieved ICD term after the BERT-F1 re-ranking are reported.

Table 2
Retrieval scores across ICD-9, ICD-10 and ICD-11 with simple retrieval query. Higher values are marked in bold, while starred ones indicates the overall best results for each field case.

Retrieve score (cosine similarity)									
	icd9			**icd10**			**icd11**		
Retrieve %	**disease**	**macro diseases**	**macro category diseases**	**disease**	**macro diseases**	**macro category diseases**	**disease**	**macro diseases**	**macro category diseases**
5	**0.8744**	**0.7660**	**0.8757**	**0.8493**	**0.8455**	**0.8899***	**0.9820***	**0.8485***	**0.8676**
10	0.8504	0.7200	0.8473	0.8290	0.8002	0.8602	0.9586	0.8236	0.8352
20	0.8333	0.6754	0.8185	0.8112	0.7560	0.8329	0.9437	0.7947	0.8088
50	0.8126	0.6280	0.7757	0.7936	0.6908	0.7890	0.9261	0.7499	0.7721
100	0.8009	0.5741	0.7407	0.7776	0.6433	0.7587	0.9160	0.7049	0.7371

Table 3
BERT-F1 Scores across ICD-9, ICD-10 and ICD-11 with the target values in MedPix and the selected standard term. These results are obtained with simple retrieval query. Higher values are marked in bold.

BERT-F1 Score									
	icd9			**icd10**			**icd11**		
Retrieve %	**disease**	**macro diseases**	**macro category diseases**	**disease**	**macro diseases**	**macro category diseases**	**disease**	**macro diseases**	**macro category diseases**
5	0.8154	0.6648	0.6860	0.8217	0.6692	0.6808	0.8550	0.6780	0.6802
10	0.8331	0.6811	0.7003	0.8357	0.6807	0.6933	0.8680	0.6859	0.6992
20	0.8403	0.6916	0.7110	0.8433	0.6921	0.7051	0.8754	0.6985	0.7122
50	0.8484	0.7028	0.7268	0.8498	0.7076	0.7227	0.8812	0.7090	0.7266
100	**0.8520**	**0.7122**	**0.7371**	**0.8533**	**0.7163**	**0.7330**	**0.8833***	**0.7197***	**0.7420***

The strongest retrieval performance is achieved at the lowest retrieval thresholds, which retain only the top 5%, i.e. the first two candidates from the 50 retrieved items. Conversely, the highest BERT-F1 scores occur at the largest retrieval thresholds.

This trade-off reflects expected behaviour: expanding the retrieval pool size, by increasing k, introduces lower-similarity, potentially noisy candidates, which degrades overall retrieval precision (Table 2). However, a larger candidate pool provides the subsequent BERT-F1 re-ranking stage with a broader candidate space, thereby enabling it to identify closer semantic matches and driving the performance gains observed in Table 3.

Furthermore, the consistently strong baseline scores suggest an optimal trade-off point exists, particularly given the modest relative gains observed across different k thresholds. Specifically, moving from the top 5% to 100% of retrieved candidates yields an average performance difference of approximately 18% for the initial retrieval stage and only 6% for the downstream BERT-F1 score.

However, a notable discrepancy emerges across target fields: performance on broader disease categories (both macro diseases and macro category diseases) lags behind the standard disease field across both metrics. We attribute this degradation to the inherent heterogeneity of high-level clinical categories, which lack strict ontological and taxonomical structuring and are therefore more difficult to align semantically. This is further compounded by clinical documentation practices, where categorical descriptions exhibit significantly less standardization than specific disease entities.

Furthermore, zero exact matches occur between the standard terminology and macro categories across all evaluated ICD versions. Within specific disease fields, exact matches appear predominantly in ICD-11 (8%), compared to only 0.44% in ICD-10 and none in ICD-9. These findings, corroborated by the retrieval and BERT-F1 metrics in Tables 2 and 3, demonstrate that the structural revisions introduced

in ICD-11, despite increasing overall system complexity, align substantially closer with the natural, free-text phrasing used in clinical documentation.

Similar patterns emerge when using both definition-based and report-based retrieval queries (see Appendix A). For both query types, scaling k yields smaller relative improvements in retrieval and BERT-F1 performance compared to the baseline simple query, with the definition-based formulation achieving the highest overall scores. This suggests that injecting superfluous contextual information does not necessarily yield optimal retrieval performance. Notably, while downstream BERT-F1 scores remain relatively stable across query formulations, definition-based queries provide a marked improvement in the initial retrieval stage.

5. Conclusions and Future Works

This study presents a robust methodology for the Clinical Coding Alignment problem across the ICD-9, ICD-10, and ICD-11 taxonomies using clinical reports from the MedPix 2.0 data set. The workflow [5] effectively combines an initial exact-match strategy with a fallback semantic retrieval pipeline. By leveraging the Nemotron-3-Embed-8B model to query a vector database via cosine similarity, and subsequently re-ranking candidates using the BERT-F1 metric, we successfully mapped free-form medical annotations to standardised codes.

Our findings demonstrate that alignment success depends heavily on target field standardization: while specific disease fields yielded strong retrieval and re-ranking metrics, macro categories proved challenging due to their unstructured nature in clinical practice. Furthermore, comparing query formulations revealed that definition-based retrieval queries achieved the best overall retrieval performance, showing that targeted semantic definitions outperform both simple queries and detailed queries which incorporates the clinical report. Most notably, exact-match analysis revealed that ICD-11 achieves an exact-match rate of nearly 8% with free-text narratives, significantly outperforming ICD-10 (0.44%) and ICD-9 (0%). Ultimately, these results underscore that despite its increased structural complexity, ICD-11's modernized vocabulary aligns far more naturally with real-world clinical documentation. Future work will extend this framework to additional clinical corpora and refine the retrieval strategy to build a fully automated clinical narrative alignment system.

Acknowledgements

We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy).

Declaration on Generative AI

During the preparation of this work, the authors used Gemini in order to: paraphrase and reword, improve writing style. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Bondielli, G. Bonetta, E. Leonardelli, I. Siragusa, Preface to the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026) co-located with 24th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2026), 2026.

[2] ICD-11, ICD-11, 2025. Https://icd.who.int/en/.

[3] J. H. Stephens, G. R. Ledlow, T. V. Fockler, Converting icd-9 to icd-10, Hospital topics 94 (2016) 1–7.

[4] J. Hirsch, G. Nicola, G. McGinty, R. Liu, R. Barr, M. Chittle, L. Manchikanti, Icd-10: history and context, American Journal of Neuroradiology 37 (2016) 596–599.

[5] L. C. Cante, M. Graziano, I. Siragusa, B. Di Martino, R. Pirrone, An automatic alignment strategy for medical ontologies: the medpix 2.0 case study, in: Artificial Intelligence for Healthcare, and Hybrid Models for Coupling Deductive and Inductive Reasoning, Springer Nature Switzerland, Cham, 2026, pp. 257–269.

[6] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, Preprint arXiv:1904.09675, 2020.

[7] I. Siragusa, S. Contino, M. L. Ciura, R. Alicata, R. Pirrone, Medpix 2.0: A comprehensive multimodal biomedical data set for advanced ai applications with retrieval augmented generation and knowledge graphs, Data Science and Engineering (2025) 1–17.

[8] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein, Explainable prediction of medical codes from clinical text, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1101–1111. URL: https://aclanthology.org/N18-1100/. doi:10.18653/v1/N18-1100.

[9] F. Li, H. Yu, Icd coding from clinical text using multi-filter residual convolutional neural network, in: proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 8180–8187.

[10] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, N. Elhadad, Multi-label classification of patient notes: Case study on icd code assignment., in: AAAI Workshops, 2018, pp. 409–416.

[11] X. Deng, Graph inference towards icd coding, 2026. URL: https://arxiv.org/abs/2601.07496. arXiv:2601.07496.

[12] S. Xi, J. Shi, J. Yan, M. Lin, X. Zhou, Y. Cheng, H. Ding, C. C. Kang, Breaking barriers in icd classification with a robust graph neural network for hierarchical coding, Scientific reports 15 (2025) 25676.

[13] C.-W. Huang, S.-C. Tsai, Y.-N. Chen, PLM-ICD: Automatic ICD coding with pretrained language models, in: T. Naumann, S. Bethard, K. Roberts, A. Rumshisky (Eds.), Proceedings of the 4th Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Seattle, WA, 2022, pp. 10–20. URL: https://aclanthology.org/2022.clinicalnlp-1.2/. doi:10.18653/v1/2022.clinicalnlp-1.2.

[14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Advances in Neural Information Processing Systems (2020).

[15] Nemotron-3-Embed-8B-BF16, Nemotron-3-Embed-8B-BF16, 2026. Https://huggingface.co/nvidia/Nemotron-3-Embed-8B-BF16.

[16] F. Liu, K. Enevoldsen, R. Solomatin, I. Chung, T. Aarsen, Z. Fődi, Introducing rteb: A new standard for retrieval evaluation, 2025.

[17] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The Faiss library, Preprint arXiv:2401.08281, 2024.

[18] M. Sung, H. Jeon, J. Lee, J. Kang, Biomedical entity representations with synonym marginalization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3641–3650.

[19] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics (2019) 1234–1240.

[20] R. I. Doğan, R. Leaman, Z. Lu, Ncbi disease corpus: A resource for disease name recognition and concept normalization, Journal of Biomedical Informatics (2014) 1–10.

[21] M. Turisini, G. Amati, M. Cestari, Leonardo: A pan-european pre-exascale supercomputer for hpc and ai applications, arXiv preprint arXiv:2307.16885 (2023).

A. Appendix

Table 4
Retrieval scores across ICD-9, ICD-10 and ICD-11 with definition retrieval query. Higher values are marked in bold, while starred ones indicates the overall best results for each field case.

Retrieve score (cosine similarity)									
	icd9			**icd10**			**icd11**		
Retrieve %	**disease**	**macro diseases**	**macro category diseases**	**disease**	**macro diseases**	**macro category diseases**	**disease**	**macro diseases**	**macro category diseases**
5	**0.8747**	**0.6596**	**0.8697**	**0.9048**	**0.7168**	**0.9081**	**1.1264***	**0.8138***	**1.063***
10	0.8539	0.6360	0.8493	0.8856	0.6967	0.8843	1.0980	0.7821	1.0303
20	0.8401	0.6196	0.8325	0.8688	0.6771	0.8652	1.0818	0.7575	1.0079
50	0.8238	0.5899	0.8113	0.8504	0.6520	0.8441	1.0628	0.7162	0.9850
100	0.8119	0.5702	0.7957	0.8376	0.6362	0.8269	1.0513	0.6944	0.9665

Table 5
BERT-F1 Scores across ICD-9, ICD-10 and ICD-11 with the target values in MedPix and the selected standard term. These results are obtained with definition retrieval query. Higher values are marked in bold.

BERT-F1 Score									
	icd9			**icd10**			**icd11**		
Retrieve %	**disease**	**macro diseases**	**macro category diseases**	**disease**	**macro diseases**	**macro category diseases**	**disease**	**macro diseases**	**macro category diseases**
5	0.8210	0.6984	0.7965	0.8239	0.7003	0.7973	0.8556	0.7089	0.8170
10	0.8354	0.7084	0.8088	0.8395	0.7112	0.8116	0.8697	0.7236	0.8339
20	0.8430	0.7135	0.8180	0.8466	0.7185	0.8217	0.8759	0.7320	0.8435
50	0.8501	0.7189	0.8274	0.8528	0.7256	0.8308	0.8810	0.7417	0.8512
100	**0.8539**	**0.7219**	**0.8317**	**0.8555**	**0.7295**	**0.8361**	**0.8835***	**0.7466***	**0.8555***

Table 6
Retrieval scores across ICD-9, ICD-10 and ICD-11 with report retrieval query. Higher values are marked in bold, while starred ones indicates the overall best results for each field case.

Retrieve score (cosine similarity)									
	icd9			**icd10**			**icd11**		
Retrieve %	**disease**	**macro diseases**	**macro category diseases**	**disease**	**macro diseases**	**macro category diseases**	**disease**	**macro diseases**	**macro category diseases**
5	**0.8407**	**0.6356**	**0.8213**	**0.8681**	**0.6924**	**0.8516**	**1.0657***	**0.7589***	**0.9621***
10	0.8226	0.6120	0.8020	0.8489	0.6732	0.8284	1.0428	0.7278	0.9365
20	0.8080	0.5945	0.7844	0.8337	0.6534	0.8080	1.0266	0.7023	0.9162
50	0.7897	0.5676	0.7629	0.8156	0.6265	0.7855	1.0106	0.6678	0.8921
100	0.7773	0.5477	0.7477	0.8030	0.6080	0.7659	0.9989	0.6410	0.8770

Table 7
BERT-F1 Scores across ICD-9, ICD-10 and ICD-11 with the target values in MedPix and the selected standard term. These results are obtained with report retrieval query. Higher values are marked in bold.

BERT-F1 Score									
	icd9			**icd10**			**icd11**		
Retrieve %	**disease**	**macro diseases**	**macro category diseases**	**disease**	**macro diseases**	**macro category diseases**	**disease**	**macro diseases**	**macro category diseases**
5	0.8170	0.6965	0.7755	0.8195	0.6950	0.7755	0.8522	0.7016	0.8006
10	0.8325	0.7080	0.7915	0.8353	0.7073	0.7956	0.8678	0.7197	0.8206
20	0.8401	0.7137	0.8027	0.8427	0.7151	0.8069	0.8745	0.7286	0.8308
50	0.8479	0.7187	0.8146	0.8493	0.7231	0.8172	0.8794	0.7385	0.8432
100	**0.8519**	**0.7220**	**0.8229**	**0.8530**	**0.7274**	**0.8264**	**0.8820***	**0.7448***	**0.8494***