Creating a computer assisted ICD coding system : performance metric choice and use of the ICD hierarchy

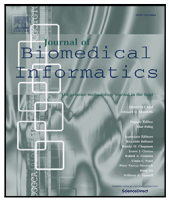
Q. Marcou, Laure Berti-Equille, N. Novelli

Journal of Biomedical Informatics 152 (2024) 104617

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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



Original Research

Creating a computer assisted ICD coding system: Performance metric choice and use of the ICD hierarchy

Quentin Marcou [a,b,*], Laure Berti-Equille [c], Noël Novelli [b]

[a] Aix-Marseille Université, Faculté des sciences médicales et paramédicales, Marseille, France
[b] Aix-Marseille Université, UMR7020 CNRS, Laboratoire d'Informatique et Systèmes (LIS), Marseille, France
[c] IRD, UMR 228 ESPACE-DEV, Montpellier, France

ARTICLE INFO

Keywords:
Recommender systems
International Classification of Diseases (ICD)
MIMIC-III
OMOP
Medication
Hierarchical Multilabel Classification (HMC)

ABSTRACT

Objective: Machine learning methods hold the promise of leveraging available data and generating higher-quality data while alleviating the data collection burden on healthcare professionals. International Classification of Diseases (ICD) diagnoses data, collected globally for billing and epidemiological purposes, represents a valuable source of structured information. However, ICD coding is a challenging task. While numerous previous studies reported promising results in automatic ICD classification, they often describe input data specific model architectures, that are heterogeneously evaluated with different performance metrics and ICD code subsets.

This study aims to explore the evaluation and construction of more effective Computer Assisted Coding (CAC) systems using generic approaches, focusing on the use of ICD hierarchy, medication data and a feed forward neural network architecture.

Methods: We conduct comprehensive experiments using the MIMIC-III clinical database, mapped to the OMOP data model. Our evaluations encompass various performance metrics, alongside investigations into multitask, hierarchical, and imbalanced learning for neural networks.

Results: We introduce a novel metric, R@R, tailored to the ICD coding task, which offers interpretable insights for healthcare informatics practitioners, aiding them in assessing the quality of assisted coding systems. Our findings highlight that selectively cherry-picking ICD codes diminish retrieval performance without performance improvement over the selected subset. We show that optimizing for metrics such as NDCG and AUPRC outperforms traditional F1-based metrics in ranking performance. We observe that Neural Network training on different ICD levels simultaneously offers minor benefits for ranking and significant runtime gains. However, our models do not derive benefits from hierarchical or class imbalance correction techniques for ICD code retrieval.

Conclusion: This study offers valuable insights for researchers and healthcare practitioners interested in developing and evaluating CAC systems. Using a straightforward sequential neural network model, we confirm that medical prescriptions are a rich data source for CAC systems, providing competitive retrieval capabilities for a fraction of the computational load compared to text-based models. Our study underscores the importance of metric selection and challenges existing practices related to ICD code sub-setting for model training and evaluation.

1. Introduction

The different versions of the International Classification of Diseases (ICD) have been used for annotating clinical data in tens of countries for several decades. While this annotation is primarily used for healthcare billing and planning purposes, it should also constitute a wealth of structured data for large-scale epidemiological studies and personalized predictions [1].

However, accurate ICD coding is both difficult and time consuming. This coding difficulty results in rather low data quality, e.g., with inter-coder agreement between fair and poor for principal diagnostic coding at the billing level, even for professional coders [2].

With the release of freely-available datasets such as MIMIC [3], there has been an endeavor of computer science and health informatics researchers to help the medical community with this clinical coding

* Corresponding author at: Aix-Marseille Université, Faculté des sciences médicales et paramédicales, Marseille, France.
E-mail address: qm.sci@protonmail.com (Q. Marcou).

https://doi.org/10.1016/j.jbi.2024.104617
Received 11 September 2023; Received in revised form 23 February 2024; Accepted 24 February 2024
Available online 1 March 2024

burden. In the last decade, many different approaches from rule-based algorithms [4] to supervised learning algorithms (such as Support Vector Machines (SVMs) [5,6] or Neural Networks (NNs)) have been explored. The NN approach is the most prevalent nowadays. NNs are particularly used for automated ICD coding based on unstructured text (using Transformers [7–9], CNNs [10–15] or RNNs [16–18]) but also more marginally for ICD automated coding based on structured [19,20] or multimodal [11,21] data (see Table S1 for a quick review of existing methods on the MIMIC-III database). The utilization of medication information in ICD coding has been relatively understudied with, to our knowledge, only one published study on this topic [20]. Nevertheless, medication information may present several assets: (1) It is closely linked to actual diagnosis for medical conditions; (2) Systematic recording by prescription softwares; (3) It is language-insensitive and uses normalized nomenclatures such as RxNorm; (4) Anonymization for multi-center training is straightforward; (5) Medical treatments follow strict recommendations, offering a priori good generalizability; (6) Being represented as structured data, decent performance could be expected even with a simple model and limited preprocessing.

Despite promising results, a limited number of studies document the use of such methods in a production environment [22,23]. The adoption of such systems by the medical informatics community may be hindered by several factors. For one, conscious of the limited quality of training data and the difficulty of the coding task, coders seem to expect CAC systems, i.e., semi-automatic systems with human in the loop systems, rather than fully automated coding systems [24,25]. In turn, commonly reported performance metrics such as F1 score and analysis on small code subsets, such as top 50 billing codes (see Table S1), may seem irrelevant for the clinical coding community.

Finally, while existing studies concentrate on optimizing input-specific model architectures, a notable gap lies in the underexplored territory of leveraging the inherent properties of the ICD hierarchy, a feature inherent to any ICD coding system and model architecture. Indeed, albeit few studies using NNs leverage the ICD hierarchy information using an attention mechanism in conjunction with a graph NN [14,15,26,27] or text labels [7,11,16] within their architecture, only SVMs based studies [5,28] have made use of generic, model architecture agnostic, Hierarchical Multilabel Classification (HMC) techniques. Similarly, class imbalance, though inherent to HMC, is also seldom addressed.

In this paper, we present several significant model architecture-agnostic contributions to ICD coding using a CAC system through a systematic study of techniques to exploit CAC systems and ICD properties. First, we define a set of performance metrics, `RE@R`, tailored to the ICD coding task, and ensuring interpretability by clinical coders who may not be experts in machine learning. Second, we investigate and compare various performance metrics, revealing that optimizing NDCG or precision–recall based metrics leads to improved ranking performance compared to traditional F1 based metrics. Third, we examine the impact of utilizing the entire set of ICD codes instead of cherry-picked subsets, finding that it maintains performance on selected codes and enables more efficient recovery of additional codes from seemingly limited medication input data. Fourth, we conduct a systematic study of generic approaches with low computational overhead to test whether exploiting the hierarchical properties of the ICD classification can improve CAC systems. This includes multitask learning scenarios, hierarchical multilabel classification techniques, and class imbalance correction techniques. Our findings suggest that NNs benefit from learning from the whole hierarchy; however, neither HMC nor class imbalance correction methods improved results upon the simple multitask learning NNs. Finally, our study serves as a valuable replication of predicting ICD-9 codes based on medication information [20] using the MIMIC-III dataset [3], achieving superior classification and ranking results, as well as promoting broader applicability through the use of the OMOP-CDM data-standard.

Statement of significance	
Problem or Issue	Accurate ICD coding is challenging and time consuming, leading to low data quality.
What is Already Known	Machine learning algorithms, could improve ICD coding but their use in production environments remains limited. Existing work focuses on input data specificities and automated classification on heterogeneous subsets of ICD codes.
What this Paper Adds	We introduce interpretable performance metrics tailored for computer assisted coding, and identify target metrics to improve ranking performance. We show that utilizing the full set of ICD codes is beneficial even for input data with seemingly low information. Furthermore, we explore multitask, hierarchical and class imbalance correction methods demonstrating their limited benefits.

2. Material and methods

2.1. Data

2.1.1. MIMIC-III OMOP

We performed all our experiments on the MIMIC-III v1.4 clinical database [3], a freely-available database of 58,976 ICU stays. The raw database was mapped to the OMOP-OHDSI common data model using scripts from [29], in order to exploit the mapping of the non-standard drug prescription representation in MIMIC-III to the RxNorm standard.

2.1.2. Medication prescription data

We then transformed the resulting RxNorm codes into their corresponding RxNorm ingredients, thereby discarding any clinical form, dose or brand information. This process resulted in a set of 1164 unique RxNorm ingredients. Subsequently, we further simplified the medication prescription data by transforming it into one binary hot-encoded vector per stay, indicating whether a particular RxNorm ingredient had been prescribed at least once during the patient's stay.

2.1.3. ICD9-CM

MIMIC-III diagnoses use the ICD9-CM nomenclature, which is hierarchical with a maximal depth of 5 levels denoted, in increasing depth, as Chapters, Subchapters, 3 digits, 4 digits, and 5 digits codes. Stays are generally annotated only by leaf nodes, as only leaf nodes are used for billing purposes. Notably, not all branches have the same depth, resulting in Billing codes of either 3, 4, or 5 digits.

In order to obtain complete annotations for each ICU stay, we performed a roll-up of the ICD9 hierarchy. The resulting numbers of unique codes and the numbers of codes per ICU stay are shown in Tables 1 and 2, respectively.

2.1.4. Cherry-picked ICD9 codes

As shown in Table S1 many studies investigating automatic ICD coding using MIMIC-III focus on a cherry-picked subset of ICD9 codes. In particular in Hansen et al. [20], the only existing study focusing on the use of medication to predict ICD diagnoses, codes were cherry-picked according to both frequency and researchers' prior belief regarding medication's information content about the different diagnoses. Aiming to reproduce Hansen et al. [20]'s filters and construct our cherry-picked code set, we discarded the following codes:

- codes with less than 100 occurrences in the complete dataset;

Table 1
Number of unique ICD codes per level. The cherry-picked subset corresponds to a subset of codes filtered on both frequency and a prior belief about the potential amount of information contained in the input data, as defined in Section 2.1.4.

	ICD9-CM	MIMIC III	Cherry-picked
Chapters	19	19	16
Subchapters	154	151	68
3 digits	1,234	1070	312
4 digits	7,473	4380	520
5 digits	8,846	3716	318
Billing	12,167	5871	613
All levels	17,726	9336	1234

Table 2
Number of codes per stay for different ICD levels on the complete MIMIC-III code set.

	Min	Max	Median	Mean	SD
Chapters	0	16	6	5.77	2.97
Subchapters	0	28	7	8.10	4.38
3 digits	0	39	9	10.18	5.70
4 digits	0	39	9	10.48	6.15
5 digits	0	28	4	5.02	3.64
Billing	0	39	8	9.13	5.83
All levels	0	137	34	39.54	21.95

- codes belonging to chapters *Injury And Poisoning* (800–999), *Supplementary Classification Of External Causes Of Injury And Poisoning* (E000-E999), and *Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services* (V01-V91);
- codes belonging to *Disorders relating to short gestation and low birth-weight* (765 3-digits code).

The resulting number of unique codes and number of codes per ICU stay for the cherry-picked code set are respectively shown in Table 1 and Table S2. Note that despite our efforts to replicate their code filtering we ended up retaining significantly more codes than reported in Hansen et al. [20].

2.2. Models

2.2.1. Architecture

As the emphasis of our study is not to optimize a model structure specific to drug prescriptions, but rather perform a systematic evaluation of techniques making use of the ICD hierarchy properties, we favored a computationally frugal approach and used simple sequential NNs with hyperparameters described in Table S3.

Our neural networks take as input a vector of length 1166 containing:

- a binary indicator for patient's sex;
- the age of the patient clipped and normalized by 89 years, the maximum age reported in MIMIC;
- 1164 binary indicators for hot encoded RxNorm ingredients prescribed at least once during the patient's stay.

We performed a three-fold Monte Carlo cross-validation using 48,976 visits for training, 5000 for development, and another 5000 for testing. The model training and selection process is detailed in Appendix B.

2.2.2. Dummy model

For the sake of comparison, we introduce a dummy ranking model that ranks codes solely based on their frequencies estimated from the training set. Consequently, the predictions of this dummy model remain constant for any input.

2.3. Performance metrics

Table S1 presents research conducted on ICD code prediction using the MIMIC-III dataset, where various performance metrics, primarily derived from the field of automatic classification, have been employed. However, these metrics might not be ideal for evaluating a recommendation system designed to assist clinical coders in the coding process. Their practical significance for clinical coders, who are the primary users, could be particularly challenging to grasp. In turn, this lack of clarity may hinder the adoption of CAC systems. In this section, we introduce a dedicated set of metrics of performance quantifying ranking errors. We defer the definition of other usual metrics that we will use as comparison to Appendix D.

Our objective is to identify a performance metric that possesses the following essential characteristics:

- it captures the idea that all relevant codes must be retrieved by the clinical coder;
- it quantifies each non-relevant code mistakenly ranked before a relevant one as one unit of lost time for clinical coders;
- it allows for meaningful comparisons between different datasets or hospital wards;
- it can be easily understood and interpreted by clinical coders and healthcare informatics staff who may not have expertise in machine learning.

While widely used for assessing the performance of recommender systems, Normalized Discounted Cumulative Gain (NDCG) presents challenges in interpretation due to its nonlinear discounting, which assigns different weights to errors based on label rank. Additionally, its value is influenced by the number of labels considered [30], making it difficult to compare models trained on different code subsets or databases. In contrast, `Precision@Recall` implicitly assigns different weights to errors based on the number of positive labels per sample. `Recall@K` is more straightforward to interpret, but the choice of the relevant K depends on the average number of codes expected to be found in a given patient stay. For example, a clinical coder may be willing to browse a longer list of recommended codes annotating an ICU stay than annotating a simple day clinic visit for chemotherapy. Lastly, `Coverage`, a metric that measures the number of labels that must be examined in a ranked list to achieve 100% recall, is easily interpretable but also influenced by the number of positive labels per stay.

We introduce a set of related metrics `RankErrors@Recall` or `RE@R`. These metrics represent the number of negative labels that are mistakenly ranked above the positive labels until a fraction R of the positive labels is recovered. This set of metrics is a generalization of the notion of `Coverage` made independent of the number of positive labels per example:

$$\mathrm{RE@R}(y, \hat{y}) = \sum_{\{i | r(i) \leq R\}} \mathbb{1}_{\tilde{y}_i = 0},$$

where $r(k)$ is the recall at rank k, or `Recall@K`, and $\tilde{y}_i$ the vector of ground truth labels sorted in decreasing order according to the predicted score $\hat{y}$. Note that given this definition:

$$\mathrm{RE@100}(y, \hat{y}) = \mathrm{Coverage}(y, \hat{y}) - \sum_i y_i,$$

meaning that the sample averaged `RE@R` is simply a more granular version of coverage that is independent of the average number of positive labels per sample in the dataset.

2.4. Multitask learning

Multitask Learning (MTL) is the process of training an algorithm to perform multiple related tasks simultaneously, with the aim to improve both performance and generalization compared to isolated task training [31]. In that sense, performing multilabel ranking or classification for a single level of the ICD hierarchy, already is an instance of multitask learning.

In the remaining of this paper, we will however refer to the notion of multitask with the hierarchical nature of the ICD in mind, where ranking or classification at each level of the hierarchy corresponds

to a different task. We evaluate whether learning simultaneously on different hierarchy levels is beneficial using three distinct learning strategies:

- Learning per level: a different model is trained for each level of interest in the hierarchy (Chapters, Subchapters, 3-digits and Billing levels), yielding four separate models;
- Naive multitask learning: a flat classifier is trained with an output layer containing an output neuron for each node of the hierarchy (9336 labels);
- Re-weighted multitask learning: by construction, there exist more labels at each level as we progress towards the leaves of the hierarchy, resulting in higher importance of finer grained levels of the hierarchy in the cost function. To give equal importance to each task, or level, we have re-weighted the cost of each example such that each task have the same weight in the cost function (Appendix F3).

2.5. Hierarchical Multilabel Classification (HMC)

While multitask learning leverages the ICD hierarchy in an implicit manner, dedicated methods exist to explicitly exploit the known hierarchical links in multilabel classification or ranking. In our study, we implemented several of them, employing different approaches to incorporate hierarchical knowledge. These include the use of simple logical rules (roll-down and roll-up) to enforce hierarchical consistency of predictions, modifications of the cost functions based on logical rules such as TreeMin loss [32] and MCLoss [33], as well as hierarchical regularization techniques [34] (Appendix G).

2.6. Class imbalance

Large scale hierarchical problems inherently exhibit imbalanced label distributions, often following a power-law like distributions for deep hierarchies [35] (Fig. S1). Severe imbalance can lead an algorithm to ignore positive examples of the minority class to limit the amount of false negative for the majority class, or lead to poor generalization on a dataset with different class frequencies. In classification settings, class imbalance can be mitigated using regularization, cost function modifications or more frequently using resampling methods. Despite being the most frequent, resampling approaches increase computational load and impose constraints on model types and learning objectives for HMC. Thus, we focused on applying a regularization approach, using L2 regularization on the last classification layer, and three cost sensitive approaches: two based on upweighting the cost of positive examples by the imbalance ratio, namely Imbalance Ratio Weighting (IRW) and Imbalance Ratio Normalized Weighting (IRNW), and one variant of the cross-entropy loss, initially used in computer vision, and referred to as focal loss (Appendix H).

3. Results

3.1. ICD code cherry-picking degrades retrieval performance

A distinctive aspect of medication data as input for building a CAC system is that medication could be expected to convey little to no information about some diagnoses, such as, for example, diagnosis codes belonging to the 3-digits node *V50-V59 Persons Encountering Health Services For Specific Procedures And Aftercare*. This has led some authors to design algorithms only on cherry-picked subsets of codes filtered both by code frequency and researchers' prior beliefs about the information content of input data regarding ICD codes [20] (see Section 2.1.4). Such filtering can be justified assuming that (1) the input data conveys no information about some diagnostics and/or (2) that data is too scarce to allow any learning and/or (3) that the inclusion of such codes in the model deteriorates predictions for the selected cherry-picked codes.

We challenged the validity of these first two assumptions by training models on the complete code set and quantifying the amount of information a model, using a naive multitask architecture and trained to maximize μF1, extracted for different code subsets. Fig. 1(a) shows the resulting cumulative distribution of per-code entropy reduction, i.e., the reduction in uncertainty regarding the presence or absence of a code once the model's output is known. We find that, despite a generally lower information content, the model was still able to extract some information for more than 75% of low frequency codes with $\leq$100 occurrences in the entire MIMIC dataset, with a reduction of entropy of at least 10% for 31.8% of such codes. In fact, the algorithm was able to extract some information even for some codes with less than 5 occurrences in the whole dataset (see Fig. S2). Strikingly, Fig. 1(a) also suggests that codes filtered out based on researcher's prior beliefs about information content are not harder to predict than cherry-picked codes. In fact, the entropy reduction distribution was comparable for both code sets (p = .155, Mann–Whitney U test).

Fig. 1(b) illustrates how this extracted information content translates into ranking capabilities by comparing ranks for positive examples versus ranks for negative examples for a given code. 89.2% of codes were better ranked for positive examples than negative examples, 59.3% of codes had their rank at least halved on positive examples and 13.5% had their rank on positive examples being less than a tenth of their rank on negative examples. Every code in the cherry-picked subsets or filtered out codes based on researcher's prior beliefs had better ranks on positive examples, and 83.65% of them had a rank more than halved on positive examples.

Then, we challenged the assumption that adding codes on which the algorithm would fail to extract information could worsen the algorithm's predictions on cherry-picked codes. To that end, we carried hyperparameter optimization and training both using the complete code set and the cherry-picked code subset. Fig. 1(c), displaying the `Recall@K` on the cherry-picked code subset, shows that both models had equivalent code retrieval performance on that subset. Similar conclusions could be drawn for automatic coding systems, relying on hard classification as shown by the `μF1` scores in Table S4.

Finally, one might still wonder whether cherry-picking codes actually matters from a clinical coder perspective. Fig. 1(d) shows that using a model trained on the complete code set does improve `Recall@K` on billing codes even for `K` as low as 9, the average number of billing codes per stay (see Table 2), with for instance 4.16% of added recall for `K` = 30.

3.2. NDCG and AUPRC variants are better target metrics than F1

ICD codes prediction can be viewed as two different tasks: fully automatic coding or building a coding assistant system. The latter can be seen as a learning to rank problem with the specificity that all codes that must be coded should be found in the recommendations. While building a perfect classifier for automatic coding entails building a perfect recommender system, classical classification metrics may not convey complete or intuitive information on how useful an imperfect soft classification algorithm is as a coding assistant.

Having this hybrid system between multi-label classification and recommender or learning to rank system in mind, we defined the `RE@R` set of metrics in Section 2.3. This metric denotes the average number of non relevant codes that have to be seen by a clinical coder browsing the ranked list of codes before achieving a recall of `R`.

While achieving the minimal `RE@100` of 0 would mean achieving a perfect recommender system, `RE@100` alone is insufficient to describe the performance of an imperfect recommender system, and `RE@R` for different values of `R` is of interest. Trying to minimize `RE@R` for different values of `R` at the same time would, however, be impractical. We thus searched for a more traditional scalar performance metric that by optimizing, we would also minimize `RE@R` for different values of `R`. With that aim, we performed hyperparameter search and training on the complete code set targeting various widely used ranking and

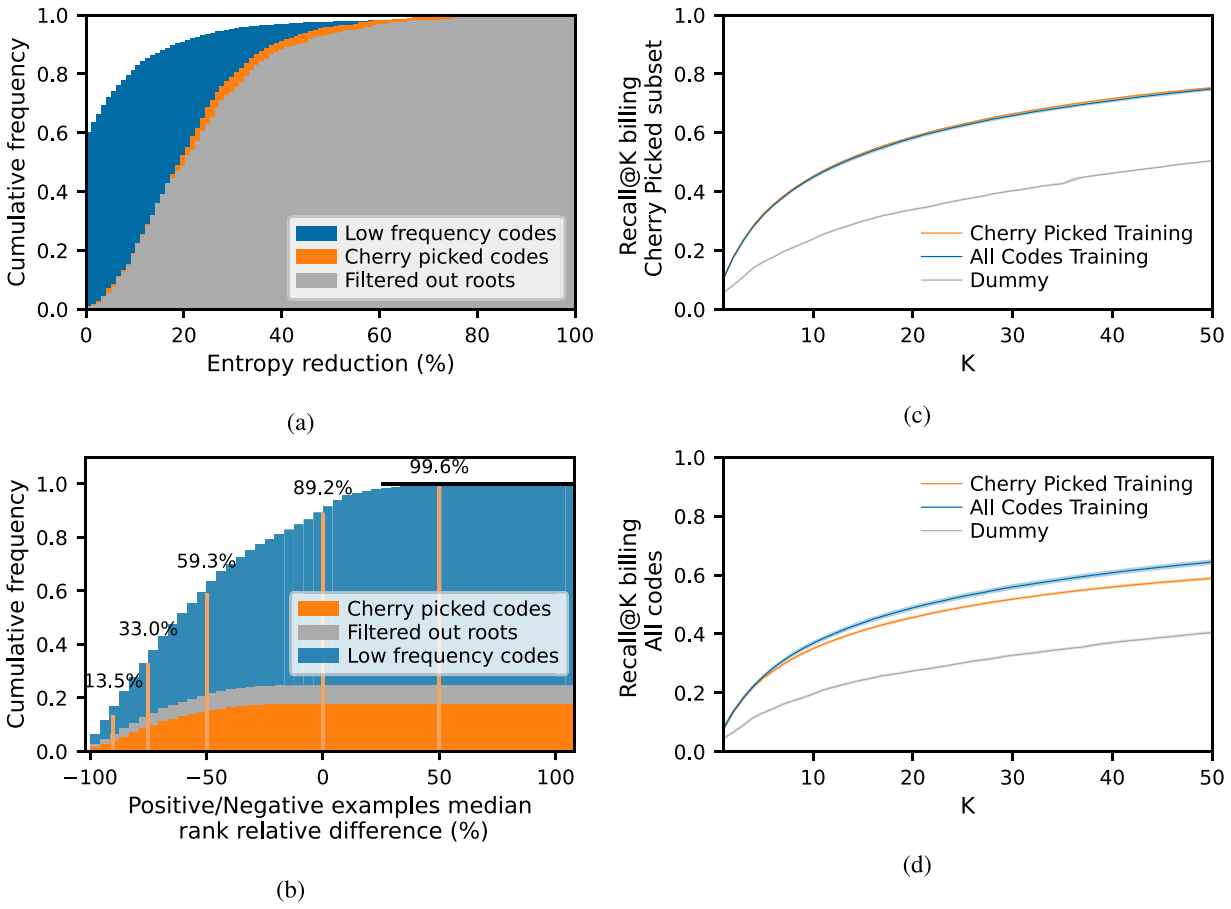


Fig. 1. a. Cumulative distribution of the entropy reduction, or information gain, per ICD code at any level of the hierarchy by performing soft classification using our naive multitask neural network model. Low frequency codes designate labels with ≤100 occurrences in the complete dataset. Filtered out roots designate codes excluded based on prior beliefs (Section 2.1.4) **b.** Cumulative distribution of the relative difference of median ranks between positive and negative examples for each ICD code. ICD codes that are, as expected, better ranked on positive than on negative examples exhibit negative values, while codes with worse ranking on positive examples have positive values. For legibility, we added a horizontal black line at y = 1. **c.** `Recall@K` on the cherry-picked code subset using a dummy model (grey), or neural networks models trained respectively on the cherry-picked subset only (orange) or the complete code subset (blue). Shades around the lines show the standard error on the estimated sample averaged `Recall@K` computed via three-fold Monte-Carlo cross-validation. **d.** `Recall@K` on the complete code set using the same models. ICD codes not seen at training time by the model trained on the cherry-picked subset only were assigned 0 probability at inference time. The four sub-figures were made using models built to optimize μF1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

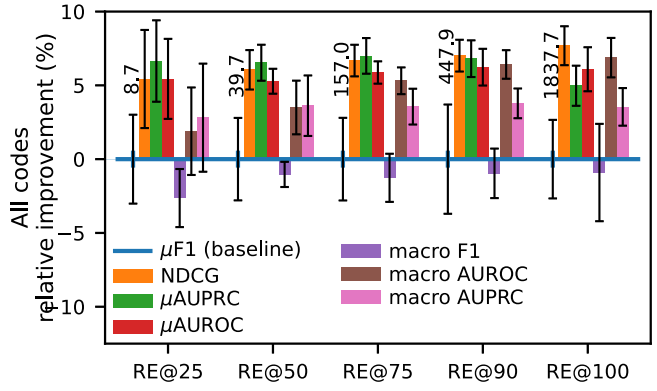


Fig. 2. Relative improvement in ranking performance, measured by `RE@R` for different values of R, achieved by optimizing various scalar performance metrics instead of the commonly reported μF1. Error bars represent the standard error of the mean computed over cross-validation runs. We indicate, as reference, the absolute `RE@R` values, averaged over cross validation runs, achieved by the baseline model optimizing the μF1.

classification metrics. The results are shown in Fig. 2, as relative `RE@R` improvement over a model selected and trained trying to optimize the μF1 score.

Despite being by far the most reported performance metric for automated ICD coding on MIMIC (see Table S1), Fig. 2 shows that optimizing F1 based metrics leads to lower ranking performance, and that optimizing either NDCG or a micro averaged AUC variant leads to ranking performance improved by ∼5% for any recall. In general, macro-averaged metrics seem to lead to lower ranking performance compared to their micro averaging counterparts. Similar conclusions can be drawn on a per hierarchy level basis from Fig. S3. To further strengthen these points, without extra model training computations, we show in Fig. S4 taking into account all models we have trained, that NDCG has highest correlation with `RE@R` for all values of `R`, closely followed by Precision-Recall based metrics.

Unsurprisingly, the NDCG, a dedicated ranking metric, seemed to be the most effective target metric to optimize ranks. However, because the computation of NDCG only relies on rank, it may not always prioritize models whose predictions can be readily interpreted as probabilities. Interestingly, our best NDCG models already demonstrated strong calibration. Nevertheless, our models chosen to maximize Area Under the Precision Recall Curve (AUPRC) exhibited notably enhanced

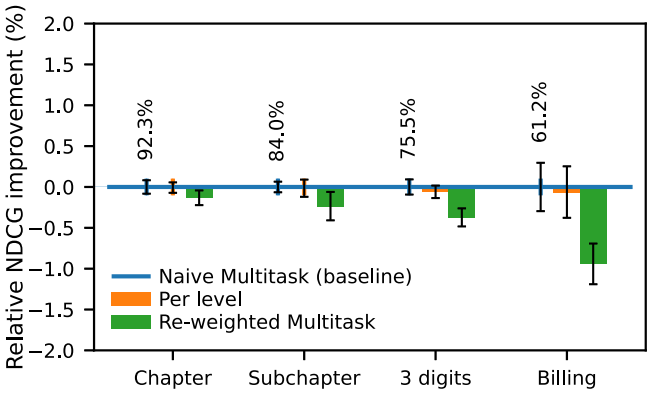


Fig. 3. Relative improvement in ranking performance, measured by NDCG, across different levels of the ICD hierarchy using various multitask learning strategies compared to the baseline naive multitask model. Absolute NDCG values, averaged over cross-validation runs, for the baseline naive multitask model are provided as reference.

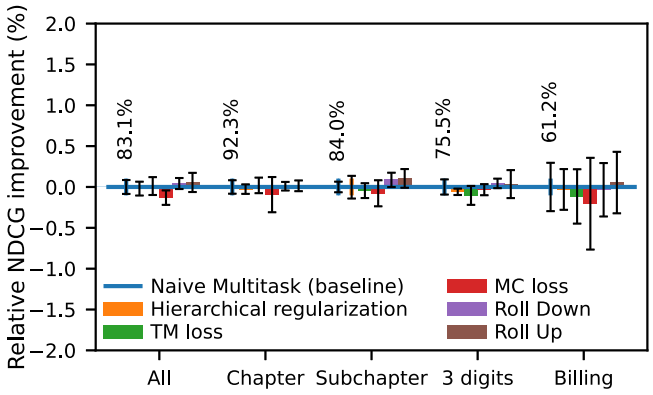


Fig. 4. Relative improvement in ranking performance, measured by NDCG, across different levels of the ICD hierarchy using various hierarchical multilabel classification strategies compared to the baseline naive multitask model. Absolute NDCG values, averaged over cross-validation runs, for the baseline naive multitask model are provided as reference.

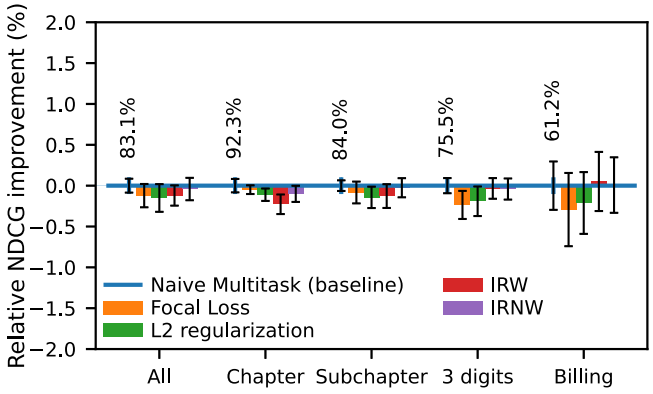


Fig. 5. Relative improvement in ranking performance, measured by NDCG, across different levels of the ICD hierarchy using various class imbalance handling strategies compared to the baseline naive multitask model. Absolute NDCG values, averaged over cross-validation runs, for the baseline naive multitask model are provided as reference.

calibration, reflected in significantly improved Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) values (Table S5).

3.3. Multitask learning improves ranking and runtime

Using multitask learning a NN may discover and exploit hierarchical relationships between tasks in hidden layers [31]. In this section, we investigate three distinct training strategies (see details in Section 2.4) to assess whether a naive use of the hierarchy information, by training a NN to rank (or classify) simultaneously codes from different levels of the hierarchy, can impact performance.

As depicted in Fig. 3, when using NDCG as a ranking performance indicator, the re-weighted multitask approach led to a significant degradation in performance across all levels of the ICD hierarchy. In contrast, while the average NDCG per level was generally higher for the naive multitask approach compared to the per-level strategy, the observed differences remained within error bounds (Fig. 3). Similarly, looking at RE@R (Fig. S5), the estimated performance was greater for the naive multitask model across various values of R and all hierarchy levels when compared to the per level models. These estimated differences, were however deemed statistically significant (paired t-test) only at the Billing level and for recalls higher than 75%. Notably, hyperparameters selected for the per-level models closely resembled those obtained for the naive multitask model (see Table S10). This observation indicate a runtime advantage associated with the naive multitask approach, which scales linearly with the number of hierarchy levels.

3.4. Hierarchical learning methods do not improve ICD code retrieval

In the previous section, we have shown that simply adding hierarchy levels as complementary tasks only marginally improves performance. In this section, we investigate whether explicit use of known hierarchical relationships, at learning or inference time, can further improve performance. We compared a flat classifier (naive multitask) with a diverse set of global hierarchical approaches using logical constraints (roll-up and roll-down), special cost functions (MCLoss, TreeMin), or regularization (HierL2) (see Section 2.5).

The NDCG achieved by these different models are illustrated in Fig. 4. We find that, analyzing performance based on NDCG, none of the hierarchical methods tested significantly change ranking performance. In fact judging by rank errors statistics, the two cost based methods, TreeMin and MCLoss, even seem to degrade ranking performance on billing codes below 50% recall (Fig. S5).

3.5. Class imbalance correction does not benefit ICD codes retrieval

As illustrated in Fig. S1, the breadth and depth of the ICD classification result in highly imbalanced labels. Taking into account class imbalance has proven beneficial for some classification tasks, but has been less studied for recommender systems.

We have implemented and applied several cost sensitive and regularization based methods used for imbalanced classification (see Section 2.6) to assess whether addressing class imbalance could improve our recommending system. Our results, shown in Fig. 5, suggest that these different methods either did not improve or even worsened the global ranking performance at every level of the ICD hierarchy. In fact, our baseline naive multitask model seem to exploit jointly code frequency (Fig. S6) and input data information (Fig. 1) for rank predictions, while still being able to output high ranks (e.g., top-10) for some very low frequency codes (Fig. S6).

Still, relying on code frequency for ranking purposes raises questions about the models generalization and equity in performance, especially across different wards within the same hospital. To assess the latter we evaluated our model's performance on exclusive newborn, surgical and medical patients subsets (Appendix J). Overall we found that, despite being a minority class, newborn stays exhibited significantly better ranking performance (Table S9). Moreover, despite the categories having similar frequencies, which may not translate directly into ICD code frequencies, significantly better ranking performance were obtained on the medical stays compared to the surgical ones.

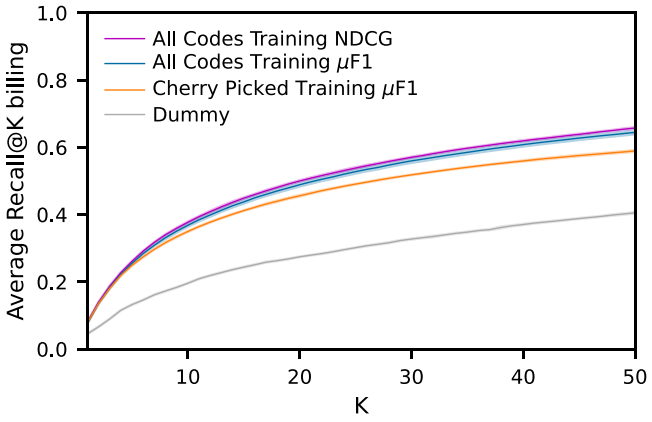


Fig. 6. Comparison of ICD code retrieval performance, focused on the billing level, between our top-performing model (shown in purple) and a model employing a strategy similar to the state-of-the-art for ICD code prediction using medication data as described in Hansen et al. [20] (represented in orange). Shades around the lines show the standard error on the estimated sample averaged `Recall@K` computed via three-fold Monte-Carlo cross-validation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.6. Overall performance summary

3.6.1. Predictions

Taken altogether we show in Fig. 6 that our proposed model, with naive multitask learning selected to maximize NDCG, improves the sample averaged `Recall@K` at all values of K when compared to the strategy similar to the one presented in Hansen et al. [20], the only existing study predicting ICD codes based on medications only. On billing codes average sample recall was increased by 5.22% for K = 30 to reach a `Recall@30` of 57.05%. On 3-digits codes, the improvement was even greater with 14.77% extra codes retrieved with a `Recall@30` of 70.66% (Fig. S7). Beyond ranking, we found that by fine-tuning a model initially optimized for μAUPRC and adjusting the probability threshold at each hierarchy level to maximize μF1 (see Appendix K) we achieved superior μF1 than by direct maximization of μF1 as a target metric (Table S4). Detailed tables presenting ranking and soft classification performance of our models can be found as Table S7 and Table S8.

3.6.2. Runtime

In our extensive experiments and across various cross-validation runs, we consistently observed the emergence of models with similar hyperparameters. Notably, the top-performing models tended to be relatively shallow yet wide networks, featuring 2 to 3 hidden layers, each containing approximately 2–3k units and a dropout rate of approximately 60% (details available in Table S10).

Our model, achieving the best NDCG performance, demonstrated remarkable efficiency in processing the comprehensive MIMIC-III dataset, encompassing 58,976 hospital stays, in 1.92 ± 0.19 s employing an inference batch size of 1024 samples on a T4 NVIDIA GPU.

4. Discussion

The first, model-agnostic step towards developing improved CAC systems and encouraging their adoption is the thoughtful selection or definition of a suitable and interpretable performance evaluation metric. To this end, we introduced the `RE@R` set of evaluation metrics based on existing literature [24,25] and discussions with health informatics professionals. Subsequently, we demonstrated that filtering ICD codes was detrimental to code retrieval without yielding any positive effect on the model's performance within the cherry-picked subset (Section 3.1). Additionally, we showed that using ranking-based metrics, such as NDCG or Precision-Recall curve-derived metrics as proxies to optimize `RE@R`, proved more effective in selecting superior models compared to μF1 (Section 3.2). These findings mark a second step towards enhancing ICD CAC systems in a model-agnostic manner. Regarding generic NNs training strategies, our results suggest that minor ranking improvements and significant computational load reduction can be achieved using a naive multitask design (Section 3.3). In evaluating the explicit use of the ICD hierarchy (Section 3.4), our results indicate that HMC methods did not yield improvements. Concerning class imbalance (Section 3.5), our results suggest that our neural network models exploit both input information and code frequency to build predictions without negative impact on minority classes. Furthermore, we observed no added performance gains, from a ranking perspective, when attempting to correct for label imbalance through various approaches with low computational overhead.

A previous real-world study [23] showed that a text-based automatic classification model optimized for μF1 improved coding quality but failed to reduce coding time for medical coders, highlighting the importance of performance metric selection and CAC system design. While our proposed evaluation metric fulfills all identified usability and interpretability criteria, further research is needed to validate its usage by studying its correlation with both coding quality and coding time. Regarding target metric choice, although the ranking-based metric NDCG demonstrated the highest correlation with `RE@R`, our results indicate that AUPRC-based metrics selects better-calibrated models, offering enhanced interpretability for clinical coders. While the observed difference in calibration remained marginal with shallow NNs, this distinction might become critical with deep networks, such as text models, which have shown poor calibration without additional correction [36]. Beyond its adverse impact on code retrieval, code cherry-picking posed significant challenges for reproducibility, as illustrated by our struggles to replicate filtered code sets across different studies [19,20,28].

At first glance, our results regarding multitask learning may seem to contradict previous results reporting significant improvement on ICD classification with a multimodal SVM adapted for hierarchical multitask learning [28]. However, the proposed SVM only shares information between explicitly hierarchy-related tasks, while our neural network approach enables multitask learning even using a single hierarchy level through common hidden features [31]. This difference already makes it more akin to explicit HMC methods. Second, the small training set size (3750 MIMIC-III patients, compared to our 48,976 stays) used in Malakouti and Hauskrecht [28] may have placed their algorithm in a more challenging learning situation where hierarchical multitask learning would be more beneficial. Overall, using our naive multitask neural network approach we obtained much higher classification performances, with 11,64% macro AUROC and 11.72% macro AUPRC differences on comparable code subsets, using only medication information compared to the best reported multimodal SVM in Malakouti and Hauskrecht [28] (see Table S6).

Our results also contrast with reports of substantial gains using HMC methods with neural networks [32,33], in particular for ICD classification using medication data [20]. For the latter, we found that our model's classification performance was generally superior to results reported by Hansen et al. [20], as discussed later in this section. Moreover, their hierarchical strategy, where parent node predictions are set to be the average of children nodes' predicted probabilities, inherently creates hierarchical violations. This design forces parent node probabilities to be inferior to the most probable child node. We hypothesize that their hierarchical loss formulation promoted a compensatory inflation of leaf node probabilities, resulting in higher recall at the expense of precision and possibly calibration, ultimately contributing to a generally improved μF1. Regarding other proper HMC methods, Giunchglia and Lukasiewicz [33] reported positive results

over several dataset with structured input. However the training set size in their experiments was generally much lower than our setting, on the order of ∼1.5 K examples compared to our ∼50k examples. We hypothesize that while HMC methods could show benefits on small datasets, this advantage might diminish with increasing dataset size. Indeed, through multitask learning NN are able to learn implicit hierarchical representations [31]. This implicit hierarchical learning could lead to an implicit hierarchy rewiring and ultimately compensate for hierarchical ontological inconsistencies, that would be enforced by an explicit HMC method [35]. The quality of the ICD9 ontology could thus constitute an alternative hypothesis for this lack of improvement using HMC methods.

While class imbalance correction techniques are frequently applied for classification purposes, their exploration within the recommender system domain has been limited. In general, these correction techniques might only alter the position of the decision plane, leaving the embedding space unaffected and resulting in identical ranking performance, as evidenced by our results. The absence of improvement in ranking performance may also stem from the intrinsic multitask nature of recommender systems when using neural networks. This could be particularly beneficial for rare labels, as suggested by the observed ranking benefits of multitask learning on codes with least favorable ranks, and the correlation between code rank and code frequency (Fig. S5 and Fig. S6).

Regarding computational efficiency, our model can process medication data with a very high throughput (Section 3.6.2). Comparatively, Gao et al. [10] reported runtimes for Convolutionnal Neural Network (CNN) and Transformer text models. These models processed 1000 samples in 8.40 and 75.86 s, respectively, using a comparable V100 NVIDIA GPU, and focusing on a single level of the hierarchy. In the same experimental conditions, our model exhibited a considerable speed advantage, completing predictions for 1000 samples in 0.077 ± 0.001 s. In addition to being faster by several orders of magnitude, our model output predictions for the entire hierarchy of ICD codes.

Regarding classification performance using medication data alone, our naive multitask model surpassed the classification results on the cherry-picked code subset reported in Hansen et al. [20] at every level of the hierarchy, except for the 3-digit level (Table S4). Our results also have to be put in perspective with other approaches using different input data. For instance, Rodrigues-Jr et al. [19] reported a 58% `Recall@20` using ICD codes history. However this result was obtained on a subset of 855 ICD codes which we were unable to replicate. Additionally, their method is limited to patients who have been previously hospitalized in the database. Our model seemed to be generally surpassed by deep Natural Language Processing (NLP) models that leveraged medical text data [10,12,13,27] on F1 and AUROC based metrics. It is worth noting that these NLP models were typically evaluated on a broader range of ICD data, encompassing both diagnoses and procedures, making direct comparisons with our work challenging. However, in a noteworthy exception, Gao et al. [10] reported significantly higher μF1 scores using a CNN text model for 3 and 5-digit codes compared to our adaptive threshold model (Table S1 and Table S4). From a computational intensity perspective, though, our feed forward NN aligns more closely with a text-based SVM model employing a bag of words as input [5]. In that study, the SVM achieved a reported μF1 of 29.3% on billing codes. In contrast, our adaptive threshold model exhibits a considerably superior hard classification performance with a 38.04% μF1 on billing codes.

We conducted our systematic analysis of plug-in methods to enhance ICD coding using MIMIC-III, a widely utilized real-world ICU clinical dataset. This choice was motivated by its availability, extensive benchmarks against other algorithms, and the facilitation of reproducibility in our work. Although the reliance on a single dataset might seem to restrict the generalization of our results, particularly given its use of ICD-9 labels instead of the more contemporary ICD-10, prior studies employing text-based models [17,23], diagnosis history [19], and even medication data [20] have demonstrated generalization with similar performance across ICD-10 [17,20,23] variants and other clinical ontologies [19], both on national databases [20] and hospital wide settings [17,19,23], when compared to identical architectures trained and evaluated on MIMIC-III. Still, it is important to note that MIMIC-III and other retrospective real-world databases likely contain numerous ICD labeling errors, which may in turn impact our model's performance evaluation. Furthermore, MIMIC-III contains only ICU stays, which might not fully represent the diversity of medical practices, and potential ward balance issues that could arise on a hospital level despite our reassuring initial experiments.

In summary, we have demonstrated the efficacy of various approaches to enhance CAC systems in a model-architecture agnostic manner, while refuting the efficacy of others through a systematic evaluation, using medication data and feed-forward NNs. While our model appear to be outperformed by deep NLP models leveraging medical text data, the significant discrepancy in observed runtimes underscores the importance of employing a feed-forward NN architecture for conducting our assessment of target metric choice, multitask, hierarchical, and class imbalance correction techniques. Employing more computationally intensive architectures would have rendered this systematic evaluation computationally intractable. Furthermore, our approach demonstrated significantly better performance than existing ones employing similarly complex models, highlighting the valuable information contained in medication data for ICD coding. In addition to providing precious insights for the construction of CAC systems, our findings may be applicable to any semi-automatic system designed to assist humans in data labeling from large-scale ontologies. Such systems are indispensable in healthcare for alleviating the data collection burden on practitioners while enhancing data quality for further reuse. Although MIMIC-III only contains ICU stay data, the volume of data it offers should be attainable even by small hospitals, thereby making our conclusions about HMC and multitask learning portable to real clinical settings. Building on our results, future research is needed to validate our proposed evaluation metric `RE@R` in a production system. Additionally, there is a need to develop more sophisticated architectures that can fully leverage medication information, including dosage, route of administration, and temporal patterns in medication intake, to further enhance prediction quality.

5. Conclusions

We have explored the construction of more effective CAC systems using generic approaches, with medication data and a simple neural network architecture as examples. We demonstrated that the practice of ICD code cherry-picking reduce overall retrieval performance without gain in the cherry-picked subset. This practice not only compromises performance but also poses challenges for reproducibility in research. Our investigation, introducing a novel metric `RE@R` tailored for the ICD coding task, revealed that optimizing for metrics like NDCG and AUPRC leads to superior ranking performance compared to the commonly used F1-based metrics. We found that multitask learning, by training NNs simultaneously on different ICD hierarchy levels, provides minor benefits for ranking as well as runtime gains. Surprisingly, our sequential NN models did not draw benefits from HMC methods or class imbalance correction techniques.

Our generic experiments offer valuable insights for researchers and healthcare practitioners interested in developing or evaluating CAC systems. Despite employing a straightforward sequential NN model, we confirmed that medical prescriptions represent a rich data source for CAC systems, providing competitive ICD codes retrieval capabilities for a fraction of the computational load compared to text-based models. Future work should validate our findings in a production setting and explore more sophisticated architectures to fully leverage medication information.

Acronyms

AUPRC	Area Under the Precision Recall Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BCE	Binary Cross-Entropy
CAC	Computer Assisted Coding
CNN	Convolutional Neural Network
ECE	Expected Calibration Error
HMC	Hierarchical Multilabel Classification
ICD	International Classification of Diseases
IRNW	Imbalance Ratio Normalized Weighting
IRW	Imbalance Ratio Weighting
MCE	Maximum Calibration Error
MTL	Multitask Learning
NDCG	Normalized Discounted Cumulative Gain
NLP	Natural Language Processing
NN	Neural Network
RNN	Recurrent Neural Network
SVM	Support Vector Machine

CRediT authorship contribution statement

Quentin Marcou: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Laure Berti-Equille:** Supervision, Writing – review & editing. **Noël Novelli:** Supervision, Writing – review & editing.

Declaration of competing interest

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

Acknowledgments

V. Pradel, F. Antonini and C. Fraboulet for initial discussions. M. Bertrand for his support and maintenance of the LIS computing cluster.

Code availability

Computer code will be released under a FOSS license on QM's Github profile: https://github.com/qmarcou.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2024.104617.

References

[1] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, npj Digit. Med. (ISSN: 2398-6352) 4 (1) (2021) 1–13.

[2] J. Stausberg, N. Lehmann, D. Kaczmarek, M. Stein, Reliability of diagnoses coding with ICD-10, Int. J. Med. Inform. (ISSN: 1386-5056) 77 (1) (2008) 50–57.

[3] A.E.W. Johnson, T.J. Pollard, L. Shen, L.-w.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data (ISSN: 2052-4463) 3 (1) (2016) 160035.

[4] R. Farkas, G. Szarvas, Automatic construction of rule-based ICD-9-CM coding systems, BMC Bioinformatics (ISSN: 1471-2105) 9 (3) (2008) S10.

[5] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, N. Elhadad, Diagnosis code assignment: Models and evaluation metrics, J. Amer. Med. Inform. Assoc. 21 (2) (2014) 231–237, ISSN 1067-5027, 1527-974X.

[6] S. Malakouti, M. Hauskrecht, Not all samples are equal: Class dependent hierarchical multi-task learning for patient diagnosis classification, in: Proceedings of the Thirty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2020), 2020, https://aaai.org/papers/323-flairs-2020-18456/.

[7] I. Chalkidis, M. Fergadiotis, S. Kotitsas, P. Malakasiotis, N. Aletras, I. Androutsopoulos, An empirical study on large-scale multi-label text classification including few and zero-shot labels, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7503–7515, https://aclanthology.org/2020.emnlp-main.607.

[8] V. Yogarajan, J. Montiel, T. Smith, B. Pfahringer, Transformers for multi-label classification of medical text: An empirical comparison, in: A. Tucker, P. Henriques Abreu, J. Cardoso, P. Pereira Rodrigues, D. Riaño (Eds.), Artificial Intelligence in Medicine, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, ISBN: 978-3-030-77211-6, 2021, pp. 114–123.

[9] P. Blinov, M. Avetisian, V. Kokh, D. Umerenkov, A. Tuzhilin, Predicting clinical diagnosis from patients electronic health records using BERT-based neural networks, in: M. Michalowski, R. Moskovitch (Eds.), Artificial Intelligence in Medicine, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, ISBN: 978-3-030-59137-3, 2020, pp. 111–121.

[10] S. Gao, M. Alawad, M.T. Young, J. Gounley, N. Schaefferkoetter, H.J. Yoon, X.-C. Wu, E.B. Durbin, J. Doherty, A. Stroup, L. Coyle, G. Tourassi, Limitations of transformers on clinical text classification, IEEE J. Biomed. Health Inf. (ISSN: 2168-2208) 25 (9) (2021) 3596–3607.

[11] K. Xu, M. Lam, J. Pang, X. Gao, C. Band, P. Mathur, F. Papay, A.K. Khanna, J.B. Cywinski, K. Maheshwari, P. Xie, E.P. Xing, Multimodal machine learning for automated ICD coding, in: Proceedings of the 4th Machine Learning for Healthcare Conference, PMLR, (ISSN: 2640-3498) 2019, pp. 197–215.

[12] 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, 34, (05) 2020, pp. 8180–8187.

[13] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein, Explainable prediction of medical codes from clinical text, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1101–1111.

[14] A. Rios, R. Kavuluru, Few-shot and zero-shot multi-label learning for structured label spaces, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3132–3142.

[15] F. Teng, W. Yang, L. Chen, L. Huang, Q. Xu, Explainable prediction of medical codes with knowledge graphs, Front. Bioeng. Biotechnol. (ISSN: 2296-4185) 8 (2020) 867.

[16] F. Catling, G.P. Spithourakis, S. Riedel, Towards automated clinical coding, Int. J. Med. Inform. (ISSN: 13865056) 120 (2018) 50–61.

[17] S.-M. Wang, Y.-H. Chang, L.-C. Kuo, F. Lai, Y.-N. Chen, F.-Y. Yu, C.-W. Chen, Z.-W. Li, Y. Chung, Using deep learning for automatic icd-10 classification from free-text data, Eur. J. Biomed. Inform. (2020).

[18] W. Sun, S. Ji, E. Cambria, P. Marttinen, Multitask recalibrated aggregation network for medical code prediction, in: Y. Dong, N. Kourtellis, B. Hammer, J.A. Lozano (Eds.), Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, ISBN: 978-3-030-86514-6, 2021, pp. 367–383.

[19] J. Rodrigues-Jr, M. Gutierrez, G. Spadon, B. Brandoli, S. Amer-Yahia, LIG-doctor: Efficient patient trajectory prediction using bidirectional minimal gated-recurrent networks, Inf. Sci. Inf. Sci. (2021).

[20] E.R. Hansen, T. Sagi, K. Hose, G.Y.H. Lip, T.B. Larsen, F. Skjøth, Assigning diagnosis codes using medication history, Artif. Intell. Med. (ISSN: 0933-3657) 128 (2022) 102307.

[21] E. Choi, M.T. Bahadori, A. Schuetz, W.F. Stewart, J. Sun, Doctor AI: Predicting Clinical Events via recurrent neural networks, in: Proceedings of the 1st Machine Learning for Healthcare Conference, PMLR, 2016, pp. 301–318.

[22] L. Zhou, C. Cheng, D. Ou, H. Huang, Construction of a semi-automatic ICD-10 coding system, BMC Med. Inform. Decis. Mak. (ISSN: 1472-6947) 20 (1) (2020) 67.

[23] P.-F. Chen, S.-M. Wang, W.-C. Liao, L.-C. Kuo, K.-C. Chen, Y.-C. Lin, C.-Y. Yang, C.-H. Chiu, S.-C. Chang, F. Lai, Automatic ICD-10 coding and training system: Deep neural network based on supervised learning, JMIR Med. Inform. 9 (8) (2021) e23230.

[24] S. Campbell, K. Giadresco, Computer-assisted clinical coding: A narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals, Health Inf. Manage. J. (ISSN: 1833-3583) 49 (1) (2020) 5–18.

[25] K.E. Henry, R. Kornfield, A. Sridharan, R.C. Linton, C. Groh, T. Wang, A. Wu, B. Mutlu, S. Saria, Human–machine teaming is key to AI adoption: Clinicians' experiences with a deployed machine learning system, npj Digit. Med. (ISSN: 2398-6352) 5 (1) (2022) 1–6.

[26] E. Choi, M.T. Bahadori, L. Song, W.F. Stewart, J. Sun, GRAM: Graph-based attention model for healthcare representation learning, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Halifax NS Canada, ISBN: 978-1-4503-4887-4, 2017, pp. 787–795.

[27] P. Cao, Y. Chen, K. Liu, J. Zhao, S. Liu, W. Chong, HyperCore: Hyperbolic and co-graph representation for automatic ICD coding, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3105–3114.

[28] S. Malakouti, M. Hauskrecht, Hierarchical adaptive multi-task learning framework for patient diagnoses and diagnostic category classification, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, San Diego, CA, USA, ISBN: 978-1-72811-867-3, 2019, pp. 701–706.

[29] N. Paris, A. Lamer, A. Parrot, Transformation and evaluation of the MIMIC database in the OMOP common data model: development and usability study, JMIR Med Inform (ISSN: 2291-9694) 9 (12) (2021) e30970, https://medinform.jmir.org/2021/12/e30970.

[30] Y. Wang, L. Wang, Y. Li, D. He, T.-Y. Liu, A theoretical analysis of NDCG type ranking measures, in: S. Shalev-Shwartz, I. Steinwart (Eds.), Proceedings of the 26th Annual Conference on Learning Theory, in: Proceedings of Machine Learning Research, 30, PMLR, Princeton, NJ, USA, 2013, pp. 25–54, https://proceedings.mlr.press/v30/Wang13.html.

[31] R. Caruana, Multitask learning, Machine Learning (ISSN: 1573-0565) 28 (1) (1997) 41–75, https://doi.org/10.1023/A:1007379606734.

[32] L. Li, T. Zhou, W. Wang, J. Li, Y. Yang, Deep hierarchical semantic segmentation, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1236–1247.

[33] E. Giunchiglia, T. Lukasiewicz, Coherent hierarchical multi-label classification networks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in neural information processing systems, 33, Curran Associates, Inc., 2020, pp. 9662–9673, https://proceedings.neurips.cc/paper_files/paper/2020/file/6dd4e10e3296fa63738371ec0d5df818-Paper.pdf.

[34] S. Gopal, Y. Yang, Recursive regularization for large-scale classification with hierarchical and graphical dependencies, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Chicago Illinois USA, ISBN: 978-1-4503-2174-7, 2013, pp. 257–265.

[35] A. Naik, H. Rangwala, Large Scale Hierarchical Classification: State of the Art, SpringerBriefs in Computer Science, Springer International Publishing, Cham, 2018, ISBN 978-3-030-01619-7 978-3-030-01620-3.

[36] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning, PMLR, 2017, pp. 1321–1330.