

Assigning Diagnosis Codes Using Medication History

Hansen, Emil Riis; Sagi, Tomer; Hose, Katja; Lip, Gregory Y. H.; Larsen, Torben Bjerregaard; Skjøth, Flemming

Published in:
Artificial Intelligence in Medicine

DOI (link to publication from Publisher):
[10.1016/j.artmed.2022.102307](https://doi.org/10.1016/j.artmed.2022.102307)

Creative Commons License
CC BY 4.0

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

[Link to publication from Aalborg University](#)

Citation for published version (APA):
Hansen, E. R., Sagi, T., Hose, K., Lip, G. Y. H., Larsen, T. B., & Skjøth, F. (2022). Assigning Diagnosis Codes Using Medication History. *Artificial Intelligence in Medicine*, 128(1), Article 102307. <https://doi.org/10.1016/j.artmed.2022.102307>

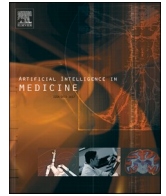
General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy

If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.



Assigning diagnosis codes using medication history[☆]

Emil Riis Hansen^{a,*}, Tomer Sagi^{a,b,**}, Katja Hose^{a,*}, Gregory Y.H. Lip^{c,f,***},
Torben Bjerregaard Larsen^{c,e,***}, Flemming Skjøth^{c,d,***}

^a Department of Computer Science, Aalborg University, Aalborg, Denmark

^b Department of Information Systems, University of Haifa, Haifa, Israel

^c Aalborg Thrombosis Research Unit, Department of Clinical Medicine, Aalborg University, Aalborg, Denmark

^d Unit of Clinical Biostatistics, Department of Research and Innovation, Aalborg University Hospital, Aalborg, Denmark

^e Thrombosis and Drug Research Unit, Department of Cardiology, Aalborg University Hospital, Aalborg, Denmark

^f Liverpool Centre for Cardiovascular Sciences, University of Liverpool and Liverpool Heart & Chest Hospital, Liverpool, United Kingdom

ARTICLE INFO

Keywords:

Diagnosis assignment
Patient profiling
Medication
Machine learning

ABSTRACT

Diagnosis assignment is the process of assigning disease codes to patients. Automatic diagnosis assignment has the potential to validate code assignments, correct erroneous codes, and register completion. Previous methods build on text-based techniques utilizing medical notes but are inapplicable in the absence of these notes. We propose using patients' medication data to assign diagnosis codes. We present a proof-of-concept study using medical data from an American dataset (MIMIC-III) and Danish nationwide registers to train a machine-learning-based model that predicts an extensive collection of diagnosis codes for multiple levels of aggregation over a disease hierarchy. We further suggest a specialized loss function designed to utilize the innate hierarchical nature of the disease hierarchy. We evaluate the proposed method on a subset of 567 disease codes. Moreover, we investigate the technique's generalizability and transferability by (1) training and testing models on the same subsets of disease codes over the two medical datasets and (2) training models on the American dataset while evaluating them on the Danish dataset, respectively. Results demonstrate the proposed method can correctly assign diagnosis codes on multiple levels of aggregation from the disease hierarchy over the American dataset with recall 70.0% and precision 69.48% for top-10 assigned codes; thereby being comparable to text-based techniques. Furthermore, the specialized loss function performs consistently better than the non-hierarchical state-of-the-art version. Moreover, results suggest the proposed method is language and dataset-agnostic, with initial indications of transferability over subsets of disease codes.

1. Introduction

The practice of coding diagnoses of medical conditions using standardized vocabularies of disease codes such as ICD-10 [1] has steadily grown. However, while coding systems are in widespread use, coding quality is uneven. Coding a medical diagnosis is notoriously complex. There exist multiple hierarchies and choosing the appropriate code requires a deep understanding of their structure and the relationships. For example, in a review of 1800 injury discharges from a New Zealand hospital, Davie et al. [2] found 2% to be uncoded, and 14% of principal

injury diagnosis codes and 26% of external cause codes to be inaccurately coded. Wockenfuss et al. [3] determined that ICD-10 three and four level codes are too detailed to be reliable for general practitioners by measuring the Kappa inter-rater agreement scores. Some work exists on predicting diagnoses from laboratory results (e.g., [4]), however, it is limited to cases where such results are available and relevant. A large body of work exists on extracting diagnoses from clinical notes and reports (see review [5]). However, the performance of these systems relies on techniques that tend to work much better in English and must be retrained for every new language [6].

[☆] This article belongs to Special issue: AIME 2020

^{*} Corresponding authors.

^{**} Correspondence to: T. Sagi, Department of Computer Science, Aalborg University, Aalborg, Denmark.

^{***} Corresponding authors at: Aalborg Thrombosis Research Unit, Department of Clinical Medicine, Aalborg University, Aalborg, Denmark.

E-mail addresses: emilrh@cs.aau.dk (E.R. Hansen), tsagi@cs.aau.dk (T. Sagi), khose@cs.aau.dk (K. Hose), Gregory.Lip@liverpool.ac.uk (G.Y.H. Lip), tobl@rn.dk (T.B. Larsen), fls@rn.dk (F. Skjøth).

<https://doi.org/10.1016/j.artmed.2022.102307>

Received 12 April 2021; Received in revised form 31 March 2022; Accepted 16 April 2022

Available online 20 April 2022

0933-3657/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

A patient's current medication can shed valuable light on their existing medical conditions. For example, observing that a patient has a long-term prescription for Metoprolol usually indicates that he/she is suffering from hypertension or ischaemic heart disease. Generalizing upon this observation, in this work, we develop a machine-learning-based model able to predict the list of diagnoses assigned to a patient based on his/her medications. Thus, such a model could provide emergency responders and critical care facilities with a rapid assessment of a patient's existing conditions in addition to the model's utility in diagnosis quality control. For example, an unconscious patient with a history of diabetes will be first assessed for hyper/hypoglycemia. In contrast, one without a history of diabetes but with a history of heart disease will be first assessed for acute heart conditions, such as a heart attack. We assess the viability of our approach using the publicly available American dataset (MIMIC-III) [7] and a Danish dataset combining prescription and diagnosis register data [8,9] denoted DNPR in the following. While MIMIC-III contains rigorously anonymized and detailed medical records for over 50 K intensive care unit (ICU) patients, DNPR contains data from an unselected population on disease codes from Danish hospital admissions and medication prescription history from Danish pharmacies.

This work extends our previous paper [10] in three ways. We investigate the generalizability and transferability of our approach by extensive experimentation on the Danish DNPR dataset. We investigate different aspects of heterogeneity between MIMIC-III and DNPR and provide results for comparable non-medication-based methods.

The rest of the paper is structured as follows. In Section 2 we review related work. In Section 3 we describe MIMIC-III and DNPR, comparing and contrasting the two datasets. In Section 4 we detail our proposed method. In Section 5 we describe our experimental setup. In Section 6 we report experimental findings and provide results for text-based methods. We discuss the implications of the results in Section 7 while concluding and providing opportunities for future work in Section 8.

2. Related work

Several studies justify the need to perform quality control of diagnosis code assignment. Cooke et al. [11] have shown that an ICD-9 code as a predictor of true chronic obstructive pulmonary disease had a sensitivity of 76% and specificity of 67% using spirometry as their gold standard. A comprehensive review of Danish validation studies on the Danish national patient registry [12] showed that the positive predictive values of disease and treatments varies from 15% to 100%. Recent work attempted to predict ICD-9 assignment in MIMIC-III from discharge notes [13]. Their solution to the multi-label multi-level problem was to limit the number of labels or aggregate predicted codes into categories, thereby solving two different problems, namely to predict the top-10/50 codes or the top 10/50 categories. In this work, we aim to predict a large set of codes at different aggregation levels to examine which codes and code groups are predictable from medication data.

There have been a few attempts to use prescription data to predict a single or at most two conditions. Schmidt et al. developed and validated an algorithm with 87% accuracy able to identify herpes zoster [14]. In another study, prescription data was used to classify whether or not patients had preexisting conditions of diabetes or hypertension [15]. In a recent review [16] of algorithms designed to extract cases for medical research from electronic medical records data, some of the studies use medication data. However, all studies extract cases for a single condition, often aggregating several diagnosis codes. In our scenario, we identify the probable diagnosis codes of multiple conditions at once and thus identify cases where improbable diagnosis codes have been used.

3. Data and heterogeneity

In this section we introduce the MIMIC-III and DNPR datasets and specify our steps of data preprocessing. Furthermore, to understand the

heterogeneity between the datasets, we investigate and highlight their main differences.

3.1. MIMIC-III

We use MIMIC-III [7] from PhysioNet [17], electronic health record (EHR) data for 50 K patients who stayed in critical care units (ICU) of the Beth Israel Deaconess Medical Center for 11 years. MIMIC-III contains an extensive variety of data, including lab results, vital signs, medical notes, and most importantly for our needs, drugs administered, and diagnoses ascertained. MIMIC-III is structured as a relational database consisting of multiple tables. For instance, MIMIC-III contains a table for drug data, a table for diagnosis data, and a table for general patient information enclosing patient age, gender etc. The drug data table (model input) contains four million rows of drugs administered during 58,976 admissions. There are 4,525 different drug names in the DRUG field, which are often the same drug, with different spelling or with an added comment, e.g., *Basiliximab* and **NF* Basiliximab*. To disambiguate and standardize the codes we use a mapping of MIMIC-III terms to the Observational Medical Outputs Partnerships (OMOP) Common Data Model (CDM) concepts [18] and group them by *Clinical Drug Form* to receive 1,602 RxNorm drug codes.

The diagnosis table (expected output) contains 651,047 diagnoses for 58,976 admissions using 6,984 different ICD-9 codes. ICD-9 is a hierarchical grouping of disease codes that consists of 5 levels starting from 0 (most general), to 4 (most specific). ICD-9 is built on the basis of grouping similar diseases. Upon review, we omit 6,110 codes for which less than 100 cases exist as it is typically not possible to generalize from such a low number. We further omit several codes focusing on diagnoses for persistent conditions not treatable by medication. A complete and detailed description of omissions can be found in Appendix A.

We use the patient table to add the *age* in years upon admission and *gender* to the model input normalized as described in Section 3.3.

3.2. DNPR

To evaluate the generalizability of the proposed method, including its language-agnostic nature, we combine the two Danish datasets “The Danish National Patient Register” [8] and “The Danish National Prescription Registry” [9]. The Danish National Patient Register is the Danish national register of diagnosis data, which contains diagnosis codes assigned during patient hospitalizations. The register contains patient records since 1977. The Danish National Prescription Registry contains prescription data for all prescriptions sold in Denmark through pharmacies since 1994. The registers can be combined patient-wise though the Danish unique personal identification number which is used throughout in Danish registers. Demographic and vital status information is obtained from the Central Person Register [19]. Throughout this work we refer to the combination of these three registers as DNPR. Due to the continuous approval of new drugs and expanding hierarchy of disease, we limit data from DNPR to the same range of years as MIMIC-III (2002–2012).

The combined register DNPR is structured as a relational database. It contains tables for patient diagnoses, prescribed drugs and general patient information among others. A main difference between MIMIC-III and DNPR is their different utilization of drug and disease vocabularies. Whereas MIMIC-III uses ICD-9 to code disease, DNPR uses a Danish extension of ICD-10 called The Danish Health Authority Classification System (SKS). Furthermore, DNPR utilizes the World Health Organization's (WHO's) Anatomical Therapeutic Classification (ATC) [20] for coding prescription drugs. DNPR contains 6,273,158 prescriptions (model input) and 2,351,769 diagnoses (expected output) for 2,093,987 admissions. Each admission (both inpatient and outpatient) consists of one or multiple diseases (both primary and secondary codes) diagnosed during hospitalization and all prescriptions administered to the patient within 30 days before and after diagnosis as illustrated in

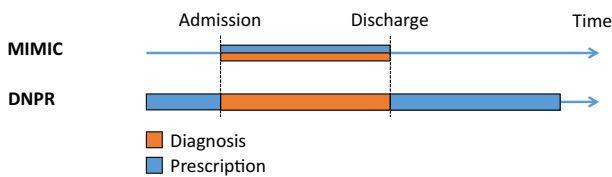


Fig. 1. Differences between MIMIC-III and DNPR in terms of prescription data gathering. An orange box represents the time of diagnosis assignment and a blue box represents the time span for which prescription medicine consumption data is gathered. Whereas MIMIC-III contains information on prescription data from time of admission until release, DNPR only contains prescription data taken before and after the patient is released from the hospital. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 1. In addition, we add patient *age* and *gender* to the model input normalized as described in Section 3.3.

3.3. Homogenization

Due to the differences between MIMIC and DNPR, a homogenization of the datasets is required. As MIMIC-III and DNPR are coded using different disease and medication vocabularies, we created mappings to convert the DNPR disease and prescription codes to the code systems used in MIMIC-III. As detailed in Sections 3.4.1 and 3.4.2, based on many-to-many general equivalence mappings (GAMs) [21] and the OMOP CDM concept mappings, we managed to create a mapping for converting disease codes between ICD-9 and ICD-10-SKS as well as one for converting prescriptions from RxNorm to ATC. We map 567 unique ICD-9 codes to 320 unique ICD-10-SKS codes and 1602 unique RxNorm drug concepts to 834 unique ATC codes. Furthermore, MIMIC-III hides elderly patients (over 89 years) due to anonymization concerns and reports the age of 92.4 for each of these. We normalize the age of all patients from MIMIC-III by dividing it by 92.4; a practice that is beneficial in machine learning techniques. We normalize the age of patients from DNPR using the same approach by first calculating the average age of elderly patients (over 89 years), then reporting elderly patients with the average age, and finally dividing all patients by the average age of the elderly. When joining the prescription, diagnosis, and patient tables for MIMIC-III, we end up with 48K admissions for 38K different patients using 567 unique codes, referred to as labels in the following.

3.4. Data heterogeneity

MIMIC-III and DNPR both consist of drug prescriptions and diagnosed diseases albeit they are collected for different purposes. Whereas MIMIC-III is collected in an insurance financed setting, DNPR is collected for administrative purposes but in a tax financed setting naturally leading to data heterogeneity.

The main difference between MIMIC-III and DNPR is the way prescription data is gathered. While diagnosis codes from MIMIC-III and DNPR are both assigned while the patient is hospitalized, prescription data from DNPR differs from MIMIC-III by not consisting of the medicine administered during hospitalization but rather the medicine taken before and after release as illustrated in Fig. 1. Furthermore, since MIMIC-III consists of ICU patients often hospitalized with acute disease, the purpose of drug administration will initially be patient stabilization. On the other hand, the purpose of DNPR prescription data is directed at treating the disease diagnosed at release, as well as chronic conditions present before and after hospitalization (e.g., diabetes).

3.4.1. Disease vocabularies

Although MIMIC-III and DNPR both utilize the ICD disease code hierarchy for standardized patient diagnosis, the hierarchy is used in different ways based on the purpose of the databases. Since subtle

changes in disease codes can cause major changes to the final patient bill, MIMIC-III disease codes have to be as specific as possible. Comparatively, Danish physicians are not too concerned with the precision of specifying diagnosis codes as long as other clinicians can understand the patient's symptomatology. As an example, a patient from MIMIC-III might get diagnosed with the billable diagnosis code 280.1 - "Iron deficiency anemias - secondary to inadequate dietary iron intake", while a patient from DNPR will be diagnosed with the less specific diagnosis code 280 - "Iron deficiency anemias", which is a non-billable ICD code.

For many years, ICD has been used globally and has thus gone through several iterations to accommodate new disease and better disease hierarchy structures. Whereas Denmark has been using the 10th version of ICD (ICD-10) since 1994, MIMIC-III patients have been diagnosed using the ICD-9 disease hierarchy. Furthermore, DNPR is coded using a Danish extension of ICD-10 called The Danish Health Authority Classification System (SKS) which extends the ICD-10 by introducing new branches of diseases and removing some codes that were originally in ICD-10. A bijective mapping between ICD-9 and ICD-10 is not possible due to the big changes between ICD versions [21]; however, a many-to-many mapping exists¹ as illustrated in Fig. 2(b) by the subset s' mapping to the subset p' . Additionally, we create subsets $s'' \subset s'$ and $p'' \subset p'$ of ICD-9 and ICD-10-SKS codes respectively for which there exists a one-to-one mapping between the sets; this is illustrated in Fig. 2 as the sets s'' and p'' . From the initial 567 ICD-9 codes with more than 100 MIMIC-III patient cases, we managed to map 320 unique ICD-9 codes to 532 ICD-10-SKS codes using the following procedure.

We utilize a many-to-many general equivalence mapping (GAM) between leaf nodes of the ICD-9 and ICD-10 disease hierarchies and consequently map 558 ICD-9 codes to 2525 ICD-10 codes. However, 1967 ICD-10 codes do not automatically correspond to ICD-10-SKS codes which results in 558 mappings from ICD-9 to ICD-10-SKS with 320 Unique ICD-9 codes mapping to 532 unique ICD-10-SKS codes forming a many-to-many relational mapping. Furthermore, we found a subset of 148 relations forming a one-to-one mapping between the two vocabularies.

3.4.2. Prescription vocabularies

Adding to the heterogeneous nature of prescription data, MIMIC-III and DNPR use different medicine vocabularies. Whereas MIMIC-III can be mapped to the RxNorm drug vocabulary using the OMOP CDM maps, DNPR is coded using the anatomical therapeutic classification (ATC). To compare the datasets we create a mapping from ATC codes to the RxNorm drug vocabulary using the OMOP CDM concept hierarchy. This results in a many-to-many mapping as seen in Fig. 2(a). Furthermore, the mapping is only partial since the OMOP CDM concept hierarchy has missing links between the two vocabularies. From the initial 1602 RxNorm drug codes, we were able to map 1257 unique RxNorm drug codes, illustrated as the set $x' \subset x$ in Fig. 2, to 834 unique ATC drug codes, illustrated as the set $y' \subset y$ in Fig. 2, with 1351 relations between x' and y' .

3.4.3. Statistical heterogeneity

Of the resulting 834 mappable ATC codes, 771 are used at least once for patients from DNPR. Furthermore, counting only the 1,257 mappable drugs, the total number of drugs given to patients from MIMIC-III is 1,129,677 with an average of 23.72 drugs per patient case. In contrast, 6,273,158 drugs are prescribed to patients from DNPR averaging at 3.00 drugs per patient case. Likewise, using the many-to-many disease code mapping, we found that of the 320 unique ICD-9 disease codes, 307 have been assigned to patients from the DNPR dataset. MIMIC-III has 47,634 patient cases with a total of 282,150 assigned disease codes which gives

¹ <https://www.cms.gov/Medicare/Coding/ICD10/2018-ICD-10-CM-and-GEMs>.

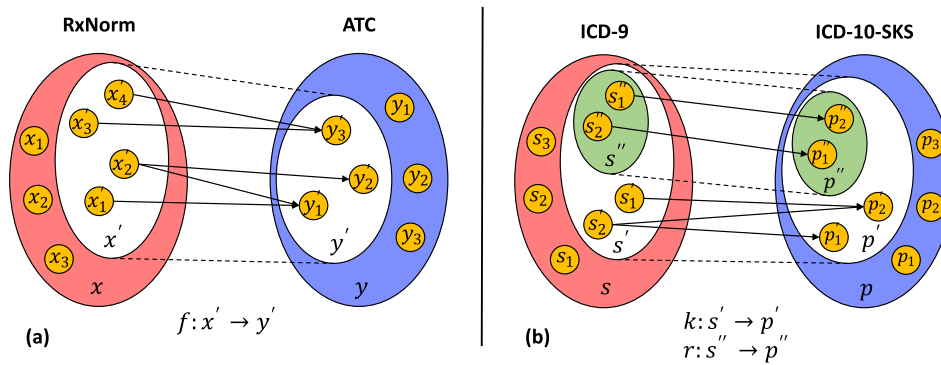


Fig. 2. (a) RxNorm to ATC mapping. The mapping between RxNorm concepts and ATC codes forms a many-to-many relationship between the two vocabularies. This can be seen by the two RxNorm concepts x_3' and x_4' both mapping to the ATC concept y_3' and the RxNorm concept x_2' mapping to the two ATC concepts y_1' and y_2' . As an example, the RxNorm concepts “Digoxin Injection” and “Digoxin Oral Tablet” both map to the ATC concept “digoxin” and the ATC concepts “triamterene” and “hydrochlorothiazide” both map to the RxNorm concept “Hydrochlorothiazide/Triamterene Oral Tablet”. (b) ICD-9 to ICD-10-SKS mapping. The mapping forms a many-to-many relationship between the two vocabularies as seen by subsets s' and p' . Furthermore, some codes only have one corresponding code from the other vocabulary, thus we create the sets s''

$\subset s'$ and $p'' \subset p'$ which have a one-to-one relationship. All mappings have been made available through an online data repository [22].

an average of 5.94 diseases per patient. DNPR has 2,093,987 patient cases with a total of 2,351,769 diagnosed disease, averaging at 1.12 disease per patient. The distribution of patients diagnosed with each of the 320 mappable ICD-9 codes is illustrated in Fig. 3.

4. Hierarchical multi-label classification (HMC)

Binary classification problems (e.g., has this person received treatment related to sepsis) aim to correctly classify each task as either positive or negative. Single-label multi-class problems (e.g., is the following brain magnetic resonance imaging (MRI) normal or does it contain a glioblastoma, a sarcoma, or a metastatic bronchogenic carcinoma?) extend the classification to allow more than one class for each task. These two types of Machine Learning (ML) tasks are, by far, the most commonly studied in the medical domain. Less common are multi-label classification problems, which attempt to assign a set of labels to each example (e.g., which of the ICD-9 codes should be assigned following this medical report [23]), each of the labels is drawn from a possible set of classes. Since each person may have multiple comorbidities, the task of assigning the correct set of diagnosis codes can be characterized as a multi-label classification problem [24]. The

hierarchical nature of diagnoses both complicates the task and offers an opportunity to improve the applicability of an ML model. If an algorithm predicts a patient suffering from non-specified chiroisis (ICD-9 code 571.5) to be suffering from alcoholic chiroisis (ICD-9 code 571.2) it should be more appreciated than if no chiroisis related diagnoses are returned since both codes share a common ancestor. Further hierarchical constraints may dictate that a person cannot have more than one label from the same sub-tree of codes. Since ICD-9 is indeed hierarchical and imposes such constraints on some of its sub-trees, we can classify our task as a hierarchical multi-label classification (HMC) problem.

4.1. Machine learning and loss functions

Many approaches to HMC include splitting the problem into multiple simple (single label) classification tasks, each of which is trained separately. Within these approaches, local and global approaches [25] differ by the number of classifiers trained. In the local case, multiple classifiers are trained over a binary label pertaining to a single node in the hierarchy and the predictions of each level are subsequently propagated [26]. In the global case, the labels are selected from a set of all possible labels. In this work, we follow the observation of Cerri et al. [27] that by

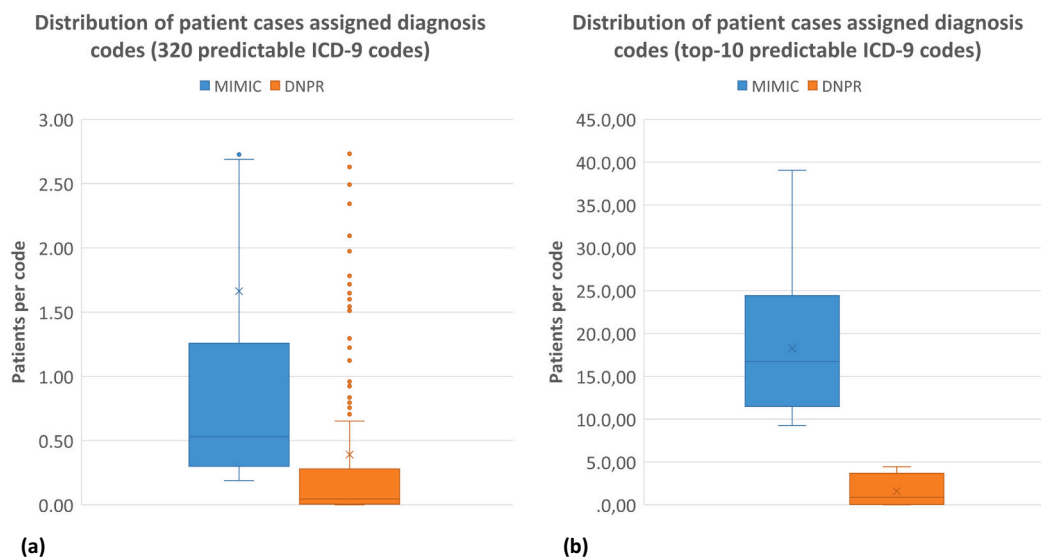


Fig. 3. Distribution of patient cases with assigned ICD-9 codes. (a) The distribution of diagnosed patients for each of the 320 predictable ICD-9 codes. As illustrated, Q3 of DNPR codes is below the interquartile range of MIMIC-III codes. For the sake of readability, no percentage above 3 is shown. However, MIMIC-III has 29 outliers not shown on the figure and DNPR has 3. (b) The distribution of the top-10 used ICD-9 codes in MIMIC. As illustrated, all disease codes are used more frequently in MIMIC-III as compared to DNPR.

training a single global classifier based on a multi-level neural network representation, one can effectively reuse the high-level features learned to discriminate between high levels in the hierarchy and then refine these to more accurate code assignments using the subsequent levels of the neural network. Furthermore, deep neural networks (DNN) have repeatedly shown superiority over other techniques in the medical domain (e.g., [28,29]). We therefore employ a multi-layer perceptron, or fully connected neural network. The input layer for this network consists of one node for each RxNorm code in the data (one for normalized age and one for biological sex) and the output layer of one node for each ICD-9 code at the chosen roll-up level.

Machine learning, in particular deep learning, uses a loss function during the training phase to quantify the error of the current iteration of the model with respect to the expected output. Choosing an appropriate loss function is crucial and in general must reflect the structure of the expected output. Thus, specific loss functions have been suggested for the multi-label case [30] as well as hierarchical multi-label functions [31]. However, these are tied directly to the structure of the global classifier, and none have been applied in the medical data setting using the inherent hierarchy of a medical taxonomy.

We therefore experiment with two types of loss functions, *ml* and *hml* as described below. One suitable for the multi-label case, where each missed label is treated the same regardless of the extent of the mistake (*ml*, Eq. (1)), and one designed for the HMC case. For the general multi-label case, we chose the multi-label soft margin loss function [32], defined as follows with C being the number of classes, y being the class indicator, and x the current value of the corresponding output node (i iterates over all classes).

$$\text{loss}(x, y) = -\frac{1}{C} \sum_i y[i] \cdot \log((1 + \exp(-x[i]))^{-1}) + (1 - y[i]) \cdot \log\left(\frac{\exp(-x[i])}{(1 + \exp(-x[i]))}\right) \quad (1)$$

We model our HMC loss function (*hml*, Eq. (2)) after the one developed for HMCN-F [31] while adjusting it to account for the differences between a text-classification problem and our own task and minimize a function comprised of two components.

$$\mathcal{L}_{hml} = \mathcal{L}_L + \mathcal{L}_G \quad (2)$$

\mathcal{L}_L is the local loss – calculation of Eq. (1) at the leaf level. \mathcal{L}_G is calculated by rolling up the results one layer at a time until the ICD-9 chapter level (0). At each phase of the roll-up, the predictions for each inner node are set to the average of the predictions over its children. The loss of each level is calculated and summed to the other levels. Since our neural network does not directly predict the global scores, we do not suffer from hierarchical violations and do not require the third

component that penalizes them in HMCN-F. We employ the Roll Up method to aggregate diagnoses given the ICD-9 hierarchy (see example in Fig. 4). Leaf node of the ICD-9 hierarchy can be assigned to patients. However, not all leaves are on the same level. As an example, 322.2 is a level 3 code, which represents *Chronic meningitis*, whereas code 003.22 is a level 4 code for *Salmonella pneumonia*. Each patient starts with one or more codes from the ICD-9 hierarchy.

5. Experimental setup

In this section, we introduce the experimental setups for evaluating different aspects of our proposed method. We evaluate the proposed method's overall performance by investigating the model's performance on the MIMIC-III dataset. Furthermore, we relate the model's performance to baseline results from several textual-based diagnosis assignment methods. To evaluate the method's generalizability, we investigate the model's performance on the Danish DNPR dataset comparing it to the performance of MIMIC-III when trained and evaluated on the same sets of ICD-9 disease codes. Finally, we investigate the model's transferability properties by training a model on the MIMIC-III dataset while testing the model on the DNPR dataset.

5.1. Diagnosis assignment using medication data (proposed method)

To evaluate the proposed method of using medication data to assign diagnosis codes, we train, evaluate and test *hml* and *ml* models on the MIMIC-III dataset with an 80/10/10 train/evaluate/test data split.

Utilizing the *roll up* method for initial data transformation, we

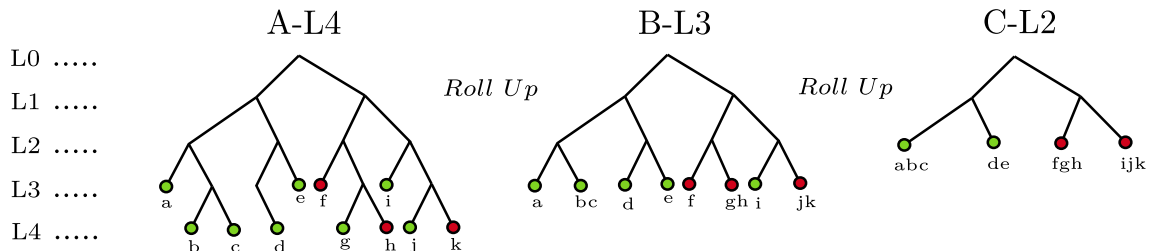


Fig. 4. Example of the roll up algorithm. An example level 4 code assignment is shown as tree A-L4. Disease codes {b, c, d, g, h, j, k} are level 4 codes, whereas codes {a, e, f, i} are codes on level 3. Red circles are the registered comorbidities of the patient. Green circles are diseases not recorded in the patient. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1

Experimental settings for evaluating the generalizability of the proposed method. M and D stand for MIMIC-III and DNPR respectively. Experiment is the name of the experiment where letters on the left and right side of the dash stand for the dataset used for training and testing respectively. Codes are the number of different disease predicted in the experiment. Train and Evaluation are the number of admissions for training and testing the model. Due to server limitations, hyperparameter optimization through standard grid search was not possible. Instead, model parameters were held constant for all experiments with the following settings - Batch Size: 256, Activation Function: Rectified Unit, Layer Dropout: 0.01, Layer Sizes: [512, 256, 128, 256].

Experiment	Codes	Disease mapping	Train	Test
M - M (multi)	320	$k^{-1} : p' \rightarrow s'$	38,107	4,763
M - M (bijection)	148	$r^{-1} : p'' \rightarrow s''$	32,228	4,028
M - M (Top 50)	50	$k^{-1} : p' \rightarrow s'$	30,367	3,795
M - M (Top 10)	10	$k^{-1} : p' \rightarrow s'$	23,286	2,910
D - D (multi)	306	$k^{-1} : p' \rightarrow s'$	1,675,189	209,398
D - D (bijection)	142	$r^{-1} : p'' \rightarrow s''$	561,789	70,223
D - D (Top 50)	50	$k^{-1} : p' \rightarrow s'$	315,007	39,375
D - D (Top 10)	10	$k^{-1} : p' \rightarrow s'$	171,372	21,421

5.2. Generalizability

Generalizability should be understood as the model's ability to perform well on new datasets and in new settings. To investigate the proposed method's generalizability, we evaluate the model on the Danish DNPR dataset. DNPR is an ideal target for evaluating the method's generalizability due to the heterogeneity between MIMIC and DNPR as detailed in Section 3.4. Since DNPR is coded using a different disease vocabulary and prescription vocabulary than that of MIMIC-III, the generalizability experiment reveals the dataset-agnostic and language-agnostic nature of the proposed method.

The experimental setup for the generalizability experiments are summarized in Table 1. Experiments are performed using ICD-9 (level 4) codes on *hml* and *ml* models. Furthermore, all experiments are trained, evaluated, and tested on an 80/10/10 data split. The DNPR data is hosted on a government server with severely restricted access and computational power thus limiting our ability to perform parameter grid search to tune the models. Hence, all experiments use the same parameter settings which can be found in the legend of Table 1.

5.3. Transferability

Transferability is the model's ability to work in a different setting from the setting in which it has been originally trained. We evaluate the proposed method's transferability by training a model on the MIMIC-III dataset while testing the model on the DNPR dataset. The transferability experiments are listed in Table 2. To ensure data compatibility, we preprocess the DNPR dataset by translating the model input and output according to the taxonomy mappings developed in Section 3.4.2 and Section 3.4.1 as illustrated in Fig. 2. All experiments are done for the most detailed level of ICD-9 (level 4) using both an *hml* and *ml* model. Due to server limitations, model parameter optimization is not possible. Model parameter settings are held constant, as listed in Table 2. All transferability experiments are trained and evaluated on an 80/20 MIMIC data split while tested on all DNPR data.

Table 2

Experimental settings for evaluating the transferability of the proposed method. M and D stand for MIMIC-III and DNPR, respectively. The description of the legend and experiments follows the same format as that of Table 1.

Experiment	Codes	Disease mapping	Train	Test
M - D (multi)	320	$k^{-1} : p' \rightarrow s'$	47,634	2,093,987
M - D (bijection)	148	$r^{-1} : p'' \rightarrow s''$	40,286	693,950
M - D (Top 50)	50	$k^{-1} : p' \rightarrow s'$	37,959	389,344
M - D (Top 10)	10	$k^{-1} : p' \rightarrow s'$	29,108	211,579

5.4. Experimentation settings

For the generalizability and transferability experiments as described in Sections 5.2 and 5.3, we experiment with multiple settings of disease codes and hierarchies. Each setting has a different rationale and clinical application in hospital settings.

The multi experiments utilize the k^{-1} disease mapping as described in Section 3.4.1. k^{-1} establishes a many-to-many link between diseases of the ICD-9 vocabulary and that of the ICD-10 vocabulary. In total, we were able to map 320 ICD-9 codes to ICD-10 codes using this mapping. The mapping is a naive conversion method since the mapping from ICD-10 to ICD-9 merges several ICD-10 codes into a single ICD-9 code. However, since most groups of merged ICD-10 codes are very similar, it should be uncommon for patients to lose important disease information when using the mapping. Furthermore, this mapping keeps many of the original disease codes from the 567 ICD-9 code set. A model based on such a mapping can be used in a clinical setting for various purposes such as automatic disease code assignment, as a validation tool for manual disease code assignment, for finding registry errors, or as a clinical tool for assessing the disease history of a patient based on the patients prescription history.

To evaluate the performance of one-to-one corresponding codes from ICD-9 and ICD-10, we created a bijective mapping function R^{-1} to map ICD-10 disease codes to ICD-9 disease codes. The experiments using this mapping are mainly used to investigate the performance of a model when mitigating the problems introduced by many-to-many mappings.

Top 10 (level 4) and top 50 (level 4) experiments use the top 10 and top 50 diagnosed codes. Previous diagnosis assignment approaches [33,34] have used top 10 and top 50 codes for experimentation. To be comparable with other approaches for diagnosis assignment on the MIMIC-III dataset, we chose to incorporate these experimental settings as well.

5.5. Comparison to text-based methods

We evaluated several textual-based approaches similar to those proposed by [33] for diagnosis assignment on different sets of the 567 MIMIC-III codes described in Section 3.1. We evaluated a Convolutional Neural Network (CNN) [35], a Recurrent Neural Network followed by a Gated Recurrent Unit (GRU), and a Convolutional Neural Network with Attention (CNN-att) [33].

The evaluated text-based methods treat ICD-9 code prediction as a multi-label classification problem. The input for text-based methods are the textual discharge summaries for patient stays, and the output is the ICD-9 codes assigned to the patient. To compare against our approach, we evaluate each of the three text-based models in a Top-10 (level 4) setting and a Raw (Level 4) setting, with Top-10 occurring MIMIC-III (level 4) codes and the set of all 567 MIMIC-III (level 4) codes, respectively.

The convolutional neural network we evaluate against, as described in [33], works as follows. As an initial data transformation step, the discharge summary notes are transformed into a feature matrix by substituting each word using pre-trained d_e -dimensional word embeddings to create an embedding matrix $X = [x_1, x_2, \dots, x_N]$, where N is the length of the document. A convolution layer then applies a convolutional filter $W_c \in \mathbb{R}^{k \times d_e \times d_c}$, where d_c is the size of the filter output, to X , to produce a convolution matrix H . A global average pooling layer is then applied to H to generate a feature for each corresponding disease to classify. The only difference between the CNN and the GRU network architecture is that a gated recurrent unit layer replaces the convolution layer from the CNN-based architecture. The CNN-att model utilizes a per-label attention mechanism since different parts of the convolution H may be relevant for different labels. The attention mechanism learns a vector parameter $u_l \in \mathbb{R}^{d_c}$ for each disease label. By doing matrix multiplication between u_l and H and using a softmax function to normalize over all words from the input file, an attention vector a_l is

learned for each label. The intuition behind a_l is that it learns which words in a document are important for classifying a specific label l .

5.6. Baseline

We introduced a statistics based disease code assignment approach as a baseline method for the task of disease code assignment. The approach is based on the statistical prior that patients are more likely to be diagnosed with common diseases than rare diseases. For each disease, we first calculate the dataset-specific probability of a patient having a disease. The assignment of patient diseases then follows a schema of generating a random floating point number between 0 and 100 for each patient for each disease. If the randomly generated number is lower than or equal to the probability of having the disease, we assign the diagnosis code to the patient. A good model should outperform this baseline by learning from the input features to choose against the statistical prior.

6. Experimental results

This section presents the results obtained from the experimental settings defined in Section 5. The obtained results are presented in separate sections according to their experimental setting. To allow easy comparison between our approach and techniques utilizing medical notes, we evaluate experimental results using the standard micro-averaged precision and recall and their harmonic mean F1. The choice of experimental settings is described in Section 5.4.

6.1. Diagnosis assignment using medication data (proposed method)

To evaluate the proposed method, we trained several models on the MIMIC-III dataset for each ICD-9 level according to the experimental setup described in Section 5.1. Table 3 presents the best results (by F1) obtained over MIMIC-III using an 80/10/10 split by an *hml* mode following a standard hyper-parameter grid search. In each task, the code assignments were rolled up before both the training and the test phase and not only for evaluation, such that the neural network encountered a different task for each level. For each ICD-9 level, we provide the number of codes in that level, the average branching factor, and the average number of eventual leaves of a node in this level's sub-tree. In addition to precision, recall, and F1, we show the number of diagnosis

Table 3

MIMIC-III diagnosis prediction results for our approach and for the baseline. Br. is the branching factor and Prec. is precision. F1 = 0 is the number of codes for which F1 was equal to zero.

Prediction task	Codes	Br.	Avg. leaves	Prec.	Recall	F1	F1 = 0
Baseline							
Top-10 (level 0)	10	NA	NA	42.04	40.10	41.05	1
Top-10 (level 4)	10	NA	NA	23.25	21.46	22.32	0
Raw (level 4)	567	0	0	8.79	8.24	8.51	402
Our approach							
Top-10 (level 0)	10	NA	NA	69.48	70.23	70.01	0
Top-10 (level 4)	10	NA	NA	52.38	70.00	59.92	0
Rolled Up (level 0)	16	5.7	565.1	68.46	69.27	68.86	0
Rolled Up (level 1)	65	8.4	108.3	58.05	57.21	57.63	10
Rolled Up (level 2)	236	6.6	14.0	48.45	47.19	47.81	83
Rolled Up (level 3)	461	1.6	1.6	37.36	41.61	39.37	195
Raw (level 4)	567	0	0	36.98	36.26	36.62	311

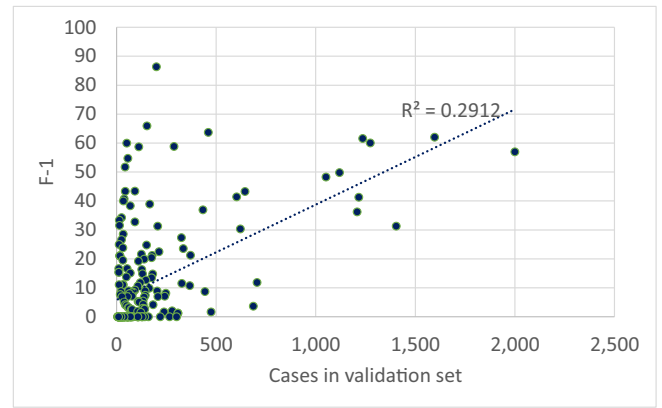


Fig. 5. F-1 by number of cases over level 2 codes.

codes for which F1 was equal to zero. Table 3 further presents the results of the baseline approach for easily comparing our proposed method against the baseline. We evaluate the Raw (level 4), Top-10 (level 0) and Top-10 (level 4) tasks for the baseline.

Since MIMIC-III is a relatively small dataset, the number of cases for many diagnoses is too low to expect good performance. When examining the effect of the number of cases on the model's performance (Fig. 5) we find that at least some of the variance can be explained by the small number of cases (R^2 of 0.29 for a linear model). Top-5/top-10 results by code are available as an online appendix containing the full results [22].

To assess the effect of using a hierarchical multi-label loss function (*hml*) versus a standard multi-label loss function (*ml*) we examine all experimental results from the *proposed method* experiment as described in Section 5.1 where the F1 was at least 5.0. Models trained using *hml* consistently out-performed those trained using *ml* with an average F1 result between 3 – 8% better. This result holds when comparing the max values obtained in each level with a 2 – 7% improvement for levels 2 – 4, although no significant improvement was seen for level 1. This last result is expected since the roll-up process for this level only rolls up to level 0.

6.2. Generalizability results

To investigate the proposed method's generalizability, we compare the performance of models trained on the MIMIC-III dataset to models trained on the same set of ICD-9 codes on the Danish DNPR dataset. The experimental setting is described in Section 5.2. Results in terms of F1 scores for *hml* and *ml* models grouped by experimental setting for all generalizability experiments are illustrated in Fig. 6.

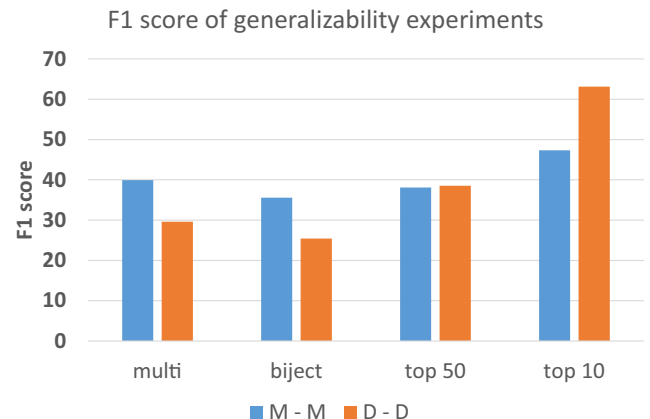


Fig. 6. F1 scores for *hml* models grouped by the type of experiment.

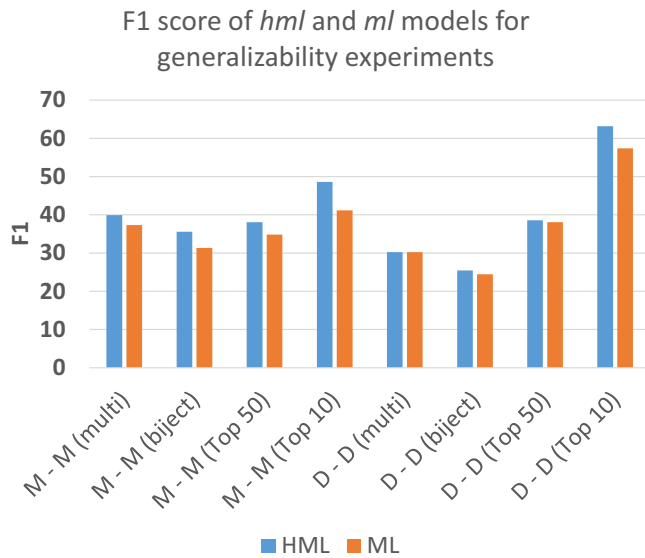


Fig. 7. F1 scores for generalizability experiments as listed in Table 1.

Table 4

F1, precision and recall of generalizability experiments for *hml* and *ml* models. M and D stand for MIMIC-III and DNPR respectively.

Experiment	HML				ML		
	Codes	F1	Prec.	Recall	F1	Prec.	Recall
M - M (multi)	320	39.93	40.16	39.71	37.30	36.53	38.12
M - M (bijection)	148	35.56	34.64	36.53	31.34	29.46	33.48
M - M (Top 50)	50	38.09	36.51	39.81	34.83	32.72	37.23
M - M (Top 10)	10	48.62	40.93	59.86	41.16	40.83	41.49
D - D (multi)	306	30.25	31.70	28.92	30.24	30.64	29.86
D - D (bijection)	142	25.42	20.43	33.61	24.42	24.36	26.57
D - D (Top 50)	50	38.55	36.25	41.16	38.09	35.14	41.58
D - D (Top 10)	10	63.16	59.54	67.25	57.38	52.42	63.38

Even though the two datasets are heterogeneous in nature, as described in Section 3.4, results indicate that the proposed method provides comparable predictive power for models trained on the MIMIC-III dataset and models trained on the DNPR dataset, for the same subsets of ICD-9 codes. Interestingly, models trained on MIMIC-III outperform models trained on DNPR when the number of predictable diseases is high. In contrast, the opposite is true when the number of predictable diseases is low.

Furthermore, results obtained from the generalizability experiments further validate the superiority of using an *hml* model as illustrated by Fig. 7, as *hml* models persistently out-performed *ml* models on F1 scores by up to 7.5% with an average performance increase of 3.1% (Table 4).

6.3. Transferability results

To assess the proposed method's transferability, we performed experiments described in Section 5.3. We trained and evaluated a model on the MIMIC-III dataset for each transferability experiment with an 80/20 data split while testing the model on the whole DNPR dataset. Results in terms of F1 score, precision and recall for all transferability experiments for *hml* and *ml* models are presented in Table 2. Results in terms of F1 score range from 6.28 when trained and tested on 320 disease codes to 28.25 when trained and tested on the top-10 most prevalent MIMIC-III ICD-9 codes as summarized in Table 5. Although transferability results indicate weak performance for models trained on the 320 ICD-9 codes, the performance improves as the prediction task gets easier.

Table 5

F1, precision and recall of transferability experiments for *hml* and *ml* models. M and D stands for MIMIC-III and DNPR respectively.

Experiment	HML				ML		
	Codes	F1	Prec.	Recall	F1	Prec.	Recall
M - D (multi)	320	6.28	7.38	5.46	5.46	4.72	6.49
M - D (bijection)	148	6.68	4.77	11.12	5.92	3.75	13.90
M - D (Top 50)	50	10.86	7.65	18.70	9.70	6.27	21.29
M - D (Top 10)	10	28.25	19.26	50.45	21.22	20.34	22.17

6.4. Results for text-based methods

To compare our work to text-based methods of diagnosis assignment, we experimented with implementations of several such methods. We evaluated state of the art Convolutional Neural Network (CNN), a Recurrent Neural Network followed by a Gated Recurrent Unit (GRU), and a Convolutional Neural Network with Attention (CNN-att) [33]. Results in terms of precision, recall, and F1 for all textual techniques are listed in Table 6. The best result in terms of F1 for the Top 10 (level 4) experiment was achieved with a CNN-att model with a score of 82.74. In comparison, the best result for our proposed method on the same set of codes is 59.92 as listed in Table 3. Similarly, whereas our proposed method achieved an F1 score of 36.62 when predicting the complete set of 567 ICD-9 codes (level 4), the best result for the text-based methods was achieved on the CNN-att model with a score of 55.76.

We further present the results for the Top-10 (level 4) assigned diagnosis codes for the text-based CNN model and our medication based HML model in Table 7. CNN predicts "Atrial fibrillation" with an F1 score of 89.66 whereas HML predicts the same disease with an F1 score of 73.09. The best performing class in terms of F1 score for the HML model is "Coronary atherosclerosis of native coronary artery" with an F1 score of 68.84. CNN predicts the same code with an F1 score of 77.23. Further results with the Top-10 assigned codes for the GRU and CNN-att text-based models are available as an online appendix [22].

7. Discussion

This section discusses and reflects upon the experimental results for the proposed method and its generalizability and transferability.

7.1. Proposed method

In the top-10 setting, an *hml* model was trained to assign one or more diseases to a patient among 10 unique ICD-9 disease codes. The model correctly assigned codes in 69.48% of all cases and was able to find 70.23% of all disease codes as summarized in Table 3.

The results of the performance of the top-10 assigned codes setting shows that text-based methods perform well in the diagnosis of all top-10 diseases as summarized in Table 7. The results indicate that diagnosis

Table 6

Results of text-based methods of diagnosis assignment. Codes are the sets of ICD-9 disease codes used in the experiment. Top 10 (level 4) is the 10 most frequently used ICD-9 codes in MIMIC-III from the initial set of 567 codes. Raw (level 4) is the complete set of 567 ICD-9 codes. For comparison, the table contains the best results for medication-based diagnosis code assignment for the same tasks.

Model	Codes	Precision	Recall	F1
CNN	Top 10 (level 4)	76.13	77.65	76.89
CNN	Raw (level 4)	44.51	46.33	42.82
GRU	Top 10 (level 4)	77.82	82.65	80.16
GRU	Raw (level 4)	62.02	49.96	55.34
CNN-att	Top 10 (level 4)	79.34	82.26	80.77
CNN-att	Raw (level 4)	57.51	59.38	55.76
<i>hml</i>	Top 10 (level 4)	54.24	67.92	60.31
<i>hml</i>	Raw (level 4)	36.98	36.26	36.62

Table 7

F1, precision and recall of top-10 assigned ICD-9 codes for our medication-based *hml* model and the text-based CNN model.

Disease	HML			CNN		
	Prec.	Recall	F1	Prec.	Recall	F1
Atrial fibrillation	69.03	77.65	73.09	87.62	91.78	89.66
Coronary atherosclerosis of native coronary artery	63.05	75.79	68.84	92.81	66.13	77.23
Unspecified essential hypertension	55.54	85.65	67.38	70.72	90.84	79.53
Congestive heart failure; unspecified	60.20	72.92	65.95	86.10	79.20	82.50
Acute respiratory failure	51.77	73.02	60.58	66.27	66.93	66.60
Acute kidney failure; unspecified	45.23	62.50	52.48	77.48	45.43	57.27
Diabetes mellitus without mention of complication	49.09	56.28	52.44	71.28	82.75	76.59
Urinary tract infection; site not specified	42.09	61.58	50.00	71.68	70.13	70.90
Other and unspecified hyperlipidemia	44.72	49.48	46.98	77.96	76.26	77.10
Esophageal reflux	36.77	21.32	26.99	82.10	67.19	73.90

observations are diligently written down in clinical discharge notes, with a precision such that text-based methods of diagnosis classification works well. Not surprisingly, it is more difficult to differentiate between diagnosis codes based on medication since some medications can be used in various contexts for treating multiple diseases. Furthermore, some diseases are not treated directly, but by adjusting some other treatments if the disease is a side effect, such as is often the case with Esophageal reflux. Hence, the F1 score of 26.99 for the medication based prediction of Esophageal reflux. However, for 8 out of 9 top-10 assigned codes, the F1 score for our medication based HML model was above 50.

The results are encouraging compared to the CNN, GRU and CNN-att textual methods of diagnosis assignment as illustrated in Fig. 6. In these days of computerized electronic health records, this approach offers a potential application to assign disease codes based on drugs prescribed automatically. The approach may also provide opportunities to create quality control mechanisms for diagnosis code assignments. The proposed method works in cases where registers do not contain medical notes but contain patient medication history, as in the Danish patient register DNPR.

As summarized in Table 3, F1 scores improve as the task is simplified with the worse performance obtained when the model tries to assign the correct code from a set of 567 possible codes at level 4. The best performance is on level 0 when the model only has 16 possible labels. Consistently, in all experimental conditions, precision and recall are approximately the same. Precision and recall are relatively low when predicting all 567 (level 4) codes. This result is partially explained by codes and groups that their medication cannot differentiate, and for which the model was unable to find any of the cases ($F1 = 0$). For example, at level 4, the model could not predict any assignment of codes from chapter 780–799 (Symptoms, Signs, And Ill-Defined Conditions). This chapter may not be differentiable by medication, as it comprises symptoms for many underlying conditions. Further analysis shows that prediction of neoplasms mostly fails, as cancer treatment can be surgical or radiation-based. Furthermore, since MIMIC contains only ICU records, the patient may not be currently undergoing any medication-based cancer treatment.

In addition, many diseases of the circulatory system were not differentiable by medication. Some diseases are asymptomatic and will thus rarely be treated by medication since the patient does not produce or show any symptoms regardless of the presence of the disease. The branch of diseases under code 426 (Conduction Disorders) are mostly asymptomatic, such as 426.0 (Atrioventricular Block, Complete), 426.4 (Right Bundle Branch Block), and 426.7 (Anomalous Atrioventricular

Excitation). Other diseases are either too general, as in 427.89 (Other Specified Cardiac Dysrhythmias), which makes it medically undiscernible, or does not have a specific medication treatment regime such as 437.0 (Cerebral Atherosclerosis). The treatment of cerebral atherosclerosis often involves administering statin, used for lowering cholesterol levels in the blood. However, statin is also used for various other atherosclerosis diseases such as aortic atherosclerosis and atherosclerosis of renal artery. Since no other discernable medication is used to treat cerebral atherosclerosis, this disease cannot be differentiated by medication.

Nonetheless, in some cases, diseases will have specific regimes of medication treatment, such as atrial fibrillation and hypertension. Patients with hypertension will often be treated by beta-blockers, ACE inhibitors or angiotensin II inhibitors. If two of these have been prescribed to a patient, there is a high probability of suffering from hypertension.

Another issue that is difficult to capture is that doses information of some drugs may vary depending on the disease indication. For example, rivaroxaban 2.5 mg BID is licensed for high-risk patients with acute coronary syndrome, while rivaroxaban 20 mg OD is for stroke prevention in atrial fibrillation. In this paper, we focused our analysis on static patient information, which means that we do not model changes in drugs over time. For example, the medication warfarin will often be prescribed to patients with venous thromboembolism and patients with atrial flutter. Whereas patients with atrial flutter will be prescribed warfarin for their entire life, venous thromboembolism patients will often stop taking warfarin after a certain period. Designing a model that can capture temporal drug information is an interesting aspect that we plan to address in our future work. Also, some patients may swap their drug into another agent from the same class of drugs, causing a further dilution of the number of cases a model can learn from. Some drugs are also in combination therapies, for example, combining ACE inhibitors and a diuretic in a single *combo* pill for the treatment of hypertension.

7.2. Generalizability

We evaluated the generalizability of the proposed method by experimenting with the Danish DNPR dataset. We compared results obtained from *hml* and *ml* models created over sets of ICD-9 codes from the MIMIC-III dataset to results obtained over the same sets of ICD-9 codes from the DNPR dataset. Experiments are summarized in Table 1. Despite their different aspects of heterogeneity, experimental results indicate comparable predictive model power for both datasets as illustrated in Fig. 6. This finding demonstrates the proposed method's dataset-agnostic properties. Furthermore, as the Danish and American datasets use distinct prescription and diagnosis vocabularies with different naming conventions for medications and diseases, we created mappings to convert between the vocabularies as described in Sections 3.4.2 and 3.4.1. Even though the mappings are incomplete and include many-to-many relations, results indicate that such conversion does not hurt the predictable properties of the proposed model when used on the Danish dataset. This result demonstrates the proposed method's language-agnostic properties.

7.3. Transferability

As indicated by the results gained from investigating the model's transferability, patient data's heterogeneous nature negatively affects the proposed methods predictive power. We evaluated the transferability of the proposed method by training models on subsets of ICD-9 disease codes of the MIMIC-III dataset while evaluating the models on the same sets of ICD-9 codes for the DNPR dataset as listed in Table 2. Results are summarized in Table 5. We achieve the F1 score of 6.28% from training an *hml* model on 320 ICD-9 level 4 codes while testing on the same subset of ATC-converted ICD-9 codes from the DNPR dataset. Furthermore, for 229 out of 320 disease codes, the model could not

provide any accurate predictions ($F1 = 0$). The results suggest that the heterogeneity between patient data across countries is too considerable to create a model with good transferability. As investigated in Section 3.4, the variability in purpose, collection method, and utilization of diverse vocabulary standards for prescription and disease code hierarchies arguably add to the variance between MIMIC and DNPR. Notwithstanding, when limiting to subsets of ICD-9 codes, model transferability significantly improves. An *hml* model trained on the top 10 occurring ICD-9 codes from the *s* code subset achieves an F1 score of 28.25 when tested on the same 10 ATC converted codes from the DNPR dataset. Noticeably, 4 out of 10 ICD-9 codes achieve an F1 score below 5.00, which indicates that the proposed method could potentially have a high transferability on specific sets of disease codes.

7.4. Domain knowledge

As with the majority of AI models today, domain knowledge is required to train models with satisfactory performance in real world applications. Although the proposed method incorporates external knowledge such as the ICD-9 disease code hierarchy and the RxNorm medication vocabulary, the proposed method is in fact agnostic towards these. Given a medical dataset coded using arbitrary disease and medication vocabularies, one could train a model using the proposed method either with or without a hierarchical taxonomy over the vocabularies. While our model performs adequately without any added domain knowledge, we show that incorporating domain knowledge in the form of hierarchical taxonomies directly into the loss function for multi-label diagnosis prediction consistently improves model results.

7.5. Practical implications

Automatic diagnosis code assignment using medication history has multiple practical implications such as registry error correction, a supportive validation tool for manual code assignment, or indicative tools usable in cases where prescription information is present but diagnosis information is not. Disease registers with manually assigned disease codes have been shown to be error-prone [2]. Using a neural model to find general patterns of medication to disease indications could automatically find outliers in register data. Currently we achieve an F1 score of 36.98% on a model for the prediction of 567 codes as summarized in Table 3. Furthermore, 311 of these codes are not discernible by medication. Hence, a neural model using only patients' prescription history can not find registry errors for all disease codes. However, as our experiments show, medication history can for some diagnosis codes be used for highly accurate diagnosis prediction and thereby be used in a system for finding register errors.

Manual diagnosis assignment is a cumbersome and error-prone task. Using a supportive tool to validate medical inputs of clinicians could help catch errors before they enter the system. As summarized in Table 3 the model performs better on higher levels of prediction. Although the model might not catch wrongly assigned diagnosis codes at the most specific level (level 4), it could help catch cases on higher aggregation levels where the error's severity is large.

In countries such as Germany, disease and medication registers are not combined. This means that emergency health care providers in ambulant settings sometimes only know the patient's prescription history and not the disease history. This can have severe implications for

the treatment of the patients, such as in the case of a patient having diabetes where several treatment protocols drastically change. In this case, a medication based diagnosis prescription model could help identify serious diseases present in the patient to guide emergency health care providers in providing the correct treatment protocol in ambulant settings.

8. Conclusion and future work

We presented a proof-of-concept study of the feasibility of using a machine learning model to assign multiple diagnosis codes on multiple aggregation levels using a person's current medication. The proposed method correctly assigned diagnosis codes on multiple levels of the ICD-9 hierarchy over the MIMIC-III dataset. The detailed results allow identifying which codes and code-groups are predictable by medication data. The use of a hierarchical loss function improved the proposed method's performance by an average F1 of 3–8% on multiple levels of aggregation of the MIMIC-III dataset while also increasing generalizability results by up to 7.5% in terms of F1 score. The promising results support continued research into utilizing larger medication datasets to create quality control mechanisms for diagnosis code assignment and provide diagnostic information to caregivers in emergencies.

Future work will further explore applications to clinical care using medication based diagnosis. Generalizability experiments demonstrate the feasibility and efficiency of the technique when applied to new dataset. Generalizability results from experimentation on the Danish DNPR dataset indicate that the technique is language-agnostic and can be directly used over new datasets. The technique is also helpful in situations where prescription data is present, but clinical discharge notes are not, as is the case with DNPR. Although model transferability underperformed when tested on the Danish DNPR dataset, results indicate that specific subsets of codes could be trained to perform well, even in model transferability. Furthermore, integrating more and diverse patient information into a unified model for diagnosis prediction should be further investigated. Patient clinical notes, medical imaging, coding systems such as laboratory codes, symptom codes and others are but a few examples of the diverse information contained in patient EHR that combined could increase the predictive performance of medical AI systems.

Declaration of competing interest

Emil Riis Hansen None identified.

Tomer Sagi None identified.

Katja Hose None identified.

Gregory Y. H. Lip Consultant and speaker for BMS/Pfizer, Boehringer Ingelheim and Daiichi-Sankyo. No fees are received personally.

Torben Bjerregaard Larsen investigator for Janssen Scientific Affairs and Boehringer-Ingelheim; speaker for AstraZeneca, Bayer, Boehringer-Ingelheim, Bristol-Myers Squibb/Pfizer, Roche Diagnostics, Siemens Diagnostics, and Takeda.

Flemming Skjøth None identified.

Acknowledgements

This research was partially funded by the Poul Due Jensen Foundation and by the Obel Family Foundation.

Appendix A. Omitted codes

Table A.8 details the omitted codes from the diagnosis table and the reasons for omission. We omit all codes with a low number of cases. We further omit 61 codes used to describe symptoms, as these are shared by multiple causes and will, most-probably, supplant a diagnosis code following medical investigation. Injuries and foreign bodies (30 codes) are omitted as well as their treatment is usually orthopedic or surgical, rather than medicinal. We omit the codes used in ICD-9 to classify birth-age and pre-term phase for infants (14 codes) as these are more descriptive than

diagnostic. Finally, we omit the E and V series of codes that are used to provide additional details for statistical reasons and which do not cause differences in medicinal treatment. We remain with 567 codes and 54,419 cases (92.4%) that contain at least one of the remaining codes. Filtering out only admissions contained in both the diagnosis and prescription tables we remain with 48,516 admissions.

Table A.8

List of omitted ICD-9 codes and code groups.

Code(s)	Description	Reason
5994 different codes	A large collection of various codes	Low base rate (less than 100 cases)
.X XX and 9XX	Descriptive of gestation week or preterm weight Injury	Will be accompanied by the specific results of pre-term birth if such exist Treatment would be Surgical or Orthopedic and impossible to accurately specify from medication
0.31,93.41 .X	Foreign body Complications of medical care	Undiscernable medicinally Undiscernable medicinally
different codes	Collection of different symptoms such as pain, nausea, and nuances of mental state/faculties	Should be accompanied by the symptom's cause which is the main diagnosis

References

- Brämer G. International statistical classification of diseases and related health problems. Tenth revision. In: World health statistics quarterly. Rapport trimestriel de statistiques sanitaires mondiales. 41 (1); 1988. p. 32–6.
- Davie G, Langley J, Samaranyaka A, Wetherspoon ME. Accuracy of injury coding under ICD-10-AM for New Zealand public hospital discharges. *Inj Prev* 2008;14(5): 319–23.
- Wockenfuss R, Frese T, Herrmann K, Claussnitzer M, Sandholzer H. Three- and four-digit ICD-10 is not a reliable classification system in primary care. *Scand J Prim Health Care* 2009;27(3):131–6.
- Razavian N, Marcus J, Sontag DA. Multi-task prediction of disease onsets from longitudinal laboratory tests. In: Doshi-Velez F, Fackler J, Kale DC, Wallace BC, Wiens J, editors. Proceedings of the 1st Machine Learning in Health Care, MLHC 2016, Los Angeles, CA, USA, August 19–20, 2016, Vol. 56 of JMLR Workshop and Conference Proceedings. JMLR.org; 2016. p. 73–100.
- Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S, Liu H. Clinical information extraction applications: a literature review. *J Biomed Inform* 2018;77:34–49.
- Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than English: opportunities and challenges. *J Biomed Semant* 2018;9(1):1–13.
- Johnson AE, Pollard TJ, Shen L, Li-Wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. Mimic-iii, a freely accessible critical care database. *SciData* 2016;3(1):1–9.
- Lynge E, Sandegaard JL, Rebolj M. The Danish national patient register. *Scand J Public Health* 2011;39(7 suppl):30–3.
- Wallach Kildemoes H, Toft Sørensen H, Hallas J. The Danish national prescription registry. *Scand J Public Health* 2011;39(7 suppl):38–41.
- Sagi T, Hansen ER, Hose K, Lip GY, Larsen TB, Skjøth F. Towards assigning diagnosis codes using medication history. In: International Conference on Artificial Intelligence in Medicine. Springer; 2020. p. 203–13.
- Cooke CR, Joo MJ, Anderson SM, Lee TA, Udris EM, Johnson E, Au DH. The validity of using icd-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease. *BMC Health Serv Res* 2011;11(1):1–10.
- Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish national patient registry: a review of content, data quality, and research potential. *Clin Epidemiol* 2015;7:449–90.
- Huang J, Osorio C, Sy LW. An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes. *Comput Methods Programs Biomed* 2019;177:141–53.
- Schmidt SA, Vestergaard M, Baggesen LM, Pedersen L, Schønheyder HC, Sørensen HT. Prevalence epidemiology of herpes zoster in Denmark: quantification of occurrence and risk factors. *Vaccine* 2017;35(42):5589–96.
- Schmidt M, Sørensen HT, Pedersen L. Diclofenac use and cardiovascular risks: series of nationwide cohort studies. *BMJ* 2018;362.
- Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc* 2016;23(5):1007–15.
- Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. Physiobank, physiobank, and physionet: components of a new research resource for complex physiologic signals. *Circulation* 2000;101(23):e215–20.
- Hripacsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong ICK, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li Y-C, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015;216:574–8.
- Ebbesen AV. The creation of the central person registry in Denmark. In: IFIP Conference on History of Nordic Computing. Springer; 2014. p. 49–57.
- Ronning M. A historical overview of the atc/ddd methodology. *WHO Drug Inform* 2002;16(3):233.
- Cartwright DJ. Icd-9-cm to icd-10-cm codes: what? why? how?. In: Advances in wound care. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA: Mary Ann Liebert, Inc.; 2013. p. 588–92.
- Hansen ER, Sagi T, Hose K, Lip GYH, Larsen TB, Skjøth F. MIMIC prescriptions result files. 2020. <https://doi.org/10.7910/DVN/SVTBME>.
- Baumel T, Nassour-Kassis J, Cohen R, Elhadad M, Elhadad N. Multi-label classification of patient notes: case study on ICD code assignment. In: The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2–7, 2018, Vol. WS-18 of AAAI Workshops. AAAI Press; 2018. p. 409–16.
- Xu D, Shi Y, Tsang IW, Ong Y-S, Gong C, Shen X. Survey on multi-output learning. *IEEE Trans Neural Netw Learn Syst* 2019;31(7):2409–29.
- Fabris F, Freitas AA, Tullet JM. An extensive empirical comparison of probabilistic hierarchical classifiers in datasets of ageing-related genes. *IEEE/ACM Trans Comput Biol Bioinform* 2016;13(6):1045–58.
- Perotte A, Pivovarov R, Natarajan K, Weiskopf N, Wood F, Elhadad N. Diagnosis code assignment: models and evaluation metrics. *J Am Med Inform Assoc* 2014;21(2):231–7.
- Cerri R, Barros RC, De Carvalho AC. Hierarchical multi-label classification using local neural networks. *J Comput Syst Sci* 2014;80(1):39–56 (1).
- Hung CY, Chen WC, Lai PT, Lin CH, Lee CC. Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. Institute of Electrical and Electronics Engineers Inc.; 2017. p. 3110–3.
- Cheng X, Zhang L, Zheng Y. Deep similarity learning for multimodal medical images. *Comput Methods Biomech Biomed Eng Imaging Vis* 2018;6(3):248–52.
- Martins AFT, Astudillo RF. From softmax to sparsemax: a sparse model of attention and multi-label classification. In: Balcan M, Weinberger KQ, editors. Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19–24, 2016, Vol. 48 of JMLR Workshop and Conference Proceedings. JMLR.org; 2016. p. 1614–23.
- Wehrmann J, Cerri R, Barros R. Hierarchical multi-label classification networks. In: Dy J, Krause A, editors. Proceedings of the 35th International Conference on Machine Learning, Vol. 80 of Proceedings of Machine Learning Research, PMLR, Stockholm, Sweden; 2018. p. 5075–84.
- Zhang ML, Zhou ZH. A review on multi-label learning algorithms. *IEEE Trans Knowl Data Eng* 2014;26(8):1819–37.
- Reys AD, Silva D, Severo D, Pedro S, Sá MMdSe, Salgado GA. Predicting multiple icd-10 codes from Brazilian-Portuguese clinical notes. In: Brazilian Conference on Intelligent Systems. Springer; 2020. p. 566–80.
- Li F, Yu H. Icd coding from clinical text using multi-filter residual convolutional neural network. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34; 2020. p. 8180–7.
- Mullenbach J, Wiegrefe S, Duke J, Sun J, Eisenstein J. Explainable prediction of medical codes from clinical text. In: Walker MA, Ji H, Stent A, editors. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1–6, 2018, Volume 1 (Long Papers). Association for Computational Linguistics; 2018. p. 1101–11.