

Using natural language processing for automated classification of disease and to identify misclassified ICD codes in cardiac disease

Maarten Falter ^{1,2,3,*}, Dries Godderis⁴, Martijn Scherrenberg ^{1,2,5},
Sevda Ece Kizilkilic ^{1,2,6}, Linqi Xu ^{1,2}, Marc Mertens⁷, Jan Jansen⁷,
Pascal Legroux⁷, Hanne Kindermans ¹, Peter Sinnaeve ³, Frank Neven ⁴,
and Paul Dendale ^{1,2}

¹Faculty of Medicine and Life Sciences, Hasselt University, Agoralaan gebouw D, 3590 Diepenbeek, Hasselt, Belgium; ²Heart Centre Hasselt, Jessa Hospital, Stadsomvaart 11, 3500 Hasselt, Belgium; ³Department of Cardiology, KULeuven, Faculty of Medicine, Herestraat 49, 3000 Leuven, Belgium; ⁴Data Science Institute, Hasselt University, Agoralaan gebouw D, 3590 Diepenbeek, Hasselt, Belgium; ⁵Faculty of Medicine and Health Sciences, Antwerp University, Universiteitsplein 1, 2610 Antwerp, Belgium; ⁶Faculty of Medicine and Health Sciences, Ghent University, Corneel Heymanslaan 10, 9000 Gent, Belgium; and ⁷Department of Information and Communications Technology, Jessa Hospital, Stadsomvaart 11, 3500 Hasselt, Belgium

Received 30 August 2023; revised 30 January 2024; accepted 5 February 2024; online publish-ahead-of-print 9 February 2024

Aims

ICD codes are used for classification of hospitalizations. The codes are used for administrative, financial, and research purposes. It is known, however, that errors occur. Natural language processing (NLP) offers promising solutions for optimizing the process. To investigate methods for automatic classification of disease in unstructured medical records using NLP and to compare these to conventional ICD coding.

Methods and results

Two datasets were used: the open-source Medical Information Mart for Intensive Care (MIMIC)-III dataset ($n = 55,177$) and a dataset from a hospital in Belgium ($n = 12,706$). Automated searches using NLP algorithms were performed for the diagnoses 'atrial fibrillation (AF)' and 'heart failure (HF)'. Four methods were used: rule-based search, logistic regression, term frequency-inverse document frequency (TF-IDF), Extreme Gradient Boosting (XGBoost), and Bio-Bidirectional Encoder Representations from Transformers (BioBERT). All algorithms were developed on the MIMIC-III dataset. The best performing algorithm was then deployed on the Belgian dataset. After preprocessing a total of 1438 reports was retained in the Belgian dataset. XGBoost on TF-IDF matrix resulted in an accuracy of 0.94 and 0.92 for AF and HF, respectively. There were 211 mismatches between algorithm and ICD codes. One hundred and three were due to a difference in data availability or differing definitions. In the remaining 108 mismatches, 70% were due to incorrect labelling by the algorithm and 30% were due to erroneous ICD coding (2% of total hospitalizations).

Conclusion

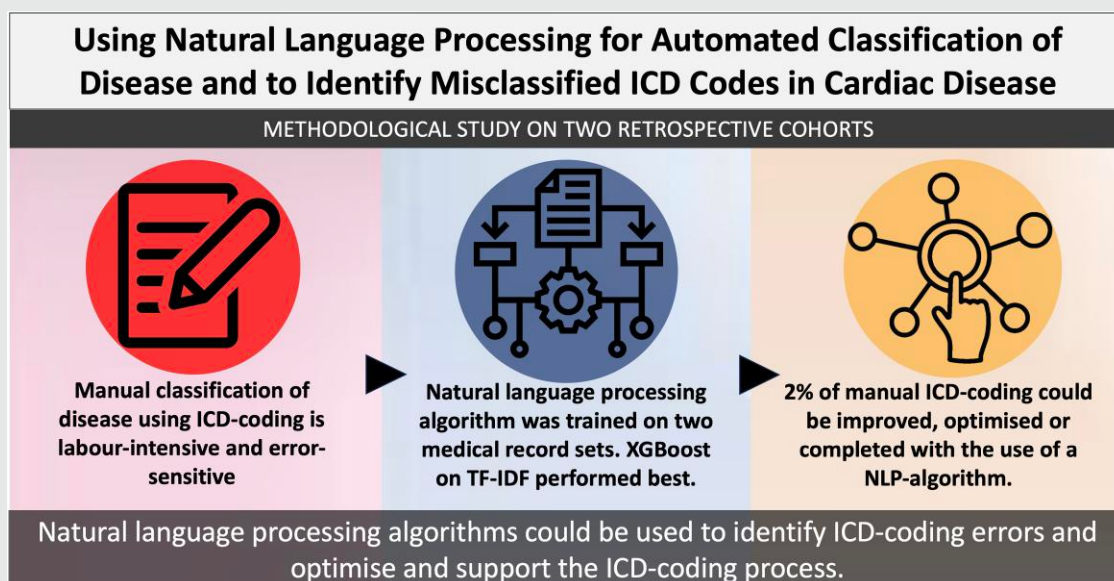
A newly developed NLP algorithm attained a high accuracy for classifying disease in medical records. XGBoost outperformed the deep learning technique BioBERT. NLP algorithms could be used to identify ICD-coding errors and optimize and support the ICD-coding process.

* Corresponding author. Tel: +3211 37 35 65, Email: maarten.falter@jessazh.be

© The Author(s) 2024. Published by Oxford University Press on behalf of the European Society of Cardiology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Graphical Abstract



Keywords

Atrial fibrillation • Heart failure • ICD codes • International classification of disease • Natural language processing • Machine learning • Deep learning • XGBoost • BioBERT

Introduction

Background

The International Statistical Classification of Diseases and Related Health Problems (ICD) is a healthcare classification system that is published by the World Health Organization (WHO).¹ The system is used worldwide for classifying disease in medical organizations. The main purpose of routine use is reporting diseases treated in hospitals to governments for epidemiological and financial reasons. More recently, the collected data is increasingly reused for clinical research purposes and quality of care assessment.

An advantage of using the ICD codes in research and quality assessment is the ease and speed in which large-scale retrospective analyses can be performed. However, it is known that the accuracy of such analyses suffers from human coding errors.^{2,3}

The importance of correct ICD coding is not to be underestimated in current times, and will even grow in importance over the coming years.

First, in its main use today the ICD-coding system is used for correct medical billing from the hospital to the patient and from the hospital to governments and healthcare insurance instances. It is known that medical coding errors result in billing inaccuracies that lead to large losses of revenue for healthcare systems.⁴

Second, quality of care is becoming a crucial and well-defined aspect in the current healthcare system. Increasingly, financial incentives are created to let healthcare systems aim for value-based healthcare. Often, these quality measures are based on bulk data of a health system/hospital, which is in turn based on medical coding through ICD codes.⁵

Third, ICD codes are already today being used for large-scale clinical research within hospitals and hospital networks. Capacities for

sharing data in a data-safe manner are increasing, and thus capacities for true big data research even across country borders is increasing. The accuracy of research findings then fully depends on correctness of ICD coding.

Lastly, ICD codes are often being used, such as in our study, to train new artificial intelligence algorithms meant to recognize certain diagnoses or that could aid as decision support systems. The ICD code mostly serves as a gold standard in the training and validation process. If systematic errors occur, systematic biases could seep through into clinical practice directly impacting quality of care.

Natural language processing (NLP) is a research discipline that has developed a wide range of automated techniques for classifying structured as well as unstructured text into classes. Such techniques offer promising solutions for aiding in the often still manual and labour-intensive process of ICD coding, and for correcting ICD-code related errors.⁶

Objectives

The aim of this study was to investigate methods for automatic classification of disease in unstructured medical records using NLP and to compare these to conventional ICD coding.

Methods

Study design

This study is a methodological study on automatic coding of unstructured and labelled medical text performed on two retrospectively collected cohorts.

Datasets

Two datasets containing full lengths textual hospitalization reports were used.

The open-source Medical Information Mart for Intensive Care (MIMIC)-III dataset⁷ was used for algorithm development. The MIMIC-III dataset is a dataset comprising of English deidentified health-related data of 55,177 patients. The patients stayed in critical care units of the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. It included demographics, vital sign measurements, laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality information. The focus for this study were discharge summaries which contained ICD-code information associated with these hospitalizations. The ICD codes used in this dataset follow the structure of the ninth revision of the ICD as published by the WHO (ICD-9, published 1978).

A novel dataset was extracted from a single hospital in Belgium. The dataset was extracted from the hospital data warehouse. The medical records of all patients receiving a percutaneous coronary intervention (PCI) between 2012 and 2020 were extracted in folders containing all the reports that were created during the hospitalization in which the PCI was performed. All reports were written in Dutch and extraction was done to Portable Document Format (pdf). The documents contained plain text with semistructured information (semistructured through use of subtitles).

The dataset contained information on 12,706 hospitalization. Discharge letters and all other reports were used with the exclusion of laboratory result reports. A paired dataset with ICD-10 codes (10th revision of the ICD as published by the WHO) associated with these hospitalizations was separately extracted in a structured manner for the period 2016–20. The ICD codes were extracted from a database that is routinely kept by the hospital for administrative purposes and that is generated by specialized administrative personnel that is trained to reread the hospitalization reports to produce the associated ICD coding.

Dataset preprocessing

In the MIMIC-III dataset various steps were performed as preprocessing. Reports were separated from files labelled as 'addendums', and only reports were used for the final analysis. Identical ICD-9 codes associated with the same hospitalization were deleted. Of patient reports containing multiple hospitalizations only the latest discharge note per patient was kept in order to avoid multiple analyses on the same patient, which would introduce bias into the algorithm (i.e. the algorithm would re-analyse the record of the same patient as a new patient record).

In the Belgian dataset, a first step before further analysis was automated pseudonymization. A combination of rule-based de-identification and deep-learning-based de-identification was used. For the deep-learning-based de-identification the Bidirectional Long Short-Term Memory (BiLSTM)-fast model as published by Trienes *et al.* was used.⁸

Definition of medical conditions

Two highly prevalent conditions that are often comorbidities in patients hospitalized in an intensive care unit and in patients with ischaemic heart disease were selected: atrial fibrillation (AF) and heart failure (HF) of all types. An advantage in the use of these two conditions is the fact that there are clear ICD-9 and ICD-10 codes and that the codes are routinely used by the ICD-coding personnel, i.e. they are not considered rare diagnoses. The ICD-9 and ICD-10 codes that were used are depicted in [Supplementary material online, Table S1](#).

Training, validation, and test datasets

For each of the two independent binary classification tasks matching the two cardiovascular conditions (AF and HF) datasets were created. The response variable indicated whether the discharge note was labelled as having the disease (1) or not (0). For the MIMIC-III dataset, a downsampled dataset with an equal amount of both response classes was created. Next, the complete dataset was split into a training (60%), validation (20%), and test (20%) dataset. The split was performed with stratification on the response variable (presence of disease). The training set is used to fit the model whereas the test set provides an evaluation of the model fit. When neural network-based models were used, there was a need for a validation set to provide an

evaluation of the model fit while tuning the model hyperparameters. For regular machine learning tasks, *k*-fold cross-validation was applied on a set consisting of the training and validation set in search of optimal hyperparameters.

For the Belgian dataset, a training dataset (75%) and a test dataset (25%) were created. In order to search for optimal hyperparameter settings, a *k*-fold cross-validation strategy was set up on the training dataset. A final XGBoost model was then fit on this training dataset using these found hyperparameters. Finally, accuracy metrics were calculated on the test dataset. No validation dataset was used in order to maximize the amount of available training observations on a relatively small dataset.

Information extraction methods

Multiple methods for information extraction and classification were deployed on the MIMIC-III dataset. The best performing algorithm was then deployed on the Belgian dataset.

Rule-based approach

A baseline model was constructed using a rule-based approach. Words and word stems were defined that could be recognized by the algorithm to label a certain medical condition. The condition was predicted as being present when at least one of the keywords was found within the report. The used keywords for each of the conditions are depicted in [Supplementary material online, Table S2](#).

Term frequency-inverse document frequency (TF-IDF)

Term frequency-inverse document frequency (TF-IDF) is a numerical statistic that is built around the counts of words within the documents in which they appear. The term frequency (TF) is the number of times a term appears in a document divided by the total number of words in that document. The inverse document frequency (IDF) is a measurement of how frequent a term occurs in the total number of documents, and can be seen as a proxy of the informativeness of that term.

An extra count was added to every document frequency applied to prevent zero divisions. A logarithm was applied to avoid large numbers in the case of a large corpus. Stemming was applied to reduce words to their base stem. A list was used to remove predefined stop words. Punctuation was removed from all tokens.

Logistic regression

A logistic regression was applied on the TF-IDF matrix. An elastic net penalized regression method was applied to reduce variance.⁹

Extreme gradient boosting (XGBoost)

Extreme gradient boosting (XGBoost), a type of boosted trees model (which is considered an improvement on decision trees), was applied on the TF-IDF matrix.¹⁰ The optimal hyperparameters for each classification task as determined by a randomized search approach on the MIMIC-III dataset were as follows (atrial fibrillation/heart failure): max_depth 6/6; n_estimators 200/300; learning_rate 0.05/0.05; gamma 1/2. Further in-depth explanation about the hyperparameters used can be found in [Supplementary material online, Table S3](#).

Bio-Bidirectional Encoder Representations from Transformers (BioBERT)

The Bidirectional Encoder Representations from Transformers (BERT) for Biomedical Text Mining (BioBERT) language model as developed by Lee *et al.* was used in this study.¹¹ Specifically BioBERT v1.1 was deployed on the MIMIC-III dataset. A limitation of the model is a maximum of 512 input tokens. The following hyperparameter values were defined: batch size: 64; Adam learning rate: 5e-5, 4e-5, 3e-5, 2e-5; number of epochs: 4.

Performance metrics and further analysis

Precision (or positive predictive value, PPV), recall (or sensitivity), and accuracy were calculated for each model. The best-performing algorithm within the MIMIC-III dataset was then deployed on the Belgian dataset.

In the Belgian dataset a consequent manual classification step was then performed on those hospitalization reports in which the ICD code and the algorithm differed. In this step, an investigator with a medical background (MF) performed a manual search through the patients' medical records to determine the final diagnosis. The reasons for mismatch between algorithm and ICD code were then investigated.

Results

Data exploration

The MIMIC-III dataset contained 55,177 reports. There were 13,142 ICD-code-based diagnoses of AF and 14,016 ICD-code-based diagnoses of HF. The dataset was filtered on the latest admission per patient retaining a total of 39,344 reports. After downsampling to retain an equal number of positive and negative diagnoses, an AF dataset was retained with 18,468 reports and a HF dataset was retained with 17,738 reports. Preprocessing pipeline for the MIMIC-III dataset is depicted in [Supplementary material online, Figure S1](#).

The Belgian dataset contained a total of 12,706 reports. After filtering on the latest admission per patient a dataset of 5780 reports was retained. A dataset with ICD codes associated with a subset of these reports was delivered for the period 2016–20 containing ICD codes for a total of 1438 hospitalizations. Preprocessing pipeline for the Belgian dataset is depicted in [Supplementary material online, Figure S2](#).

Performance of NLP algorithms

Results for the different NLP models are depicted in [Table 1](#). All models were first applied on the MIMIC-III dataset to train an optimal algorithm for later deployment on the Belgian dataset. A baseline model using a rule-based method demonstrated high precision (PPV) in both the AF and HF dataset. However, recall (sensitivity) was low, and accuracy was 0.72 and 0.75, respectively. Overall performance was better when using the logistic regression on the TF-IDF matrix. Even better overall performance was achieved when using the XGBoost on the TF-IDF matrix. The BioBERT method was only deployed on the AF dataset and demonstrated overall lower performance compared to the latter two methods.

The best performing algorithm, being the XGBoost on TF-IDF matrix was then deployed on the Belgian dataset. For both diagnoses precision (PPV) and recall (sensitivity) were lower compared to the MIMIC-III dataset. However, accuracy was still 0.94 and 0.92 for AF and HF, respectively.

Manual reclassification

False positives and false negatives for both diagnoses were identified with false positives being defined as the XGBoost on TF-IDF matrix algorithm labelling a report as positive where the ICD code is negative for this diagnosis and false negatives being defined as the XGBoost on TF-IDF matrix algorithm labelling a report as negative where the ICD code is positive for this diagnosis.

For the Belgian dataset a manual reclassification step was performed as an in-depth analysis of all misclassified reports. In this dataset a total of 211 misclassifications was found, meaning a mismatch between the reported ICD code and the label that was generated by the algorithm.

In 103 reports the misclassification was due to either different data availability to the algorithm and the ICD coders (e.g. algorithm included a report of an outpatient visit directly after the hospitalization which was not available to the ICD coders) or due to a difference in definitions (e.g. algorithm used a diagnosis mentioned in the medical history section of the patient which was intentionally not used by ICD coders as their focus is on the acute diagnoses during the relevant hospitalization and thus on the conclusion part of the report). In these cases, neither

the algorithm nor the ICD coders could be deemed right or wrong, and these cases were excluded from the further analysis.

The results for the reclassification step for the remaining 108 reports is depicted in [Table 2](#). In total, 70% of these remaining misclassifications were due to incorrect labelling by the algorithm. Examples of algorithm mistakes leading to false positives include positive labelling of a negation (e.g. 'no atrial thrombus but best excluded with transoesophageal echocardiography' resulted in positive finding due to the fact that the negation ('no') was not detected by the algorithm) or medication-based coding (e.g. the term 'amiodarone' was highly associated with AF; however, it is also used in other arrhythmias such as ventricular tachycardia which occur more rarely and which the algorithm often mislabelled).

In contrast, 30% of the misclassifications were due to erroneous or incomplete ICD coding, meaning that a diagnosis did occur but was missed by the ICD coders, or that a diagnosis was coded that did not occur in patients. When comparing to the total of 1438 hospitalizations analysed, 2% had a wrong or incomplete ICD coding that could be improved by an NLP-algorithm.

Discussion

Key results

In this study, multiple algorithms were developed for automated extraction of diagnoses from unstructured medical reports in two databases. Algorithms were developed using the first database (the MIMIC-III dataset). The logistic regression on TF-IDF matrix algorithm and the XGBoost on TF-IDF matrix algorithm demonstrated the highest overall performance, with the latter being the highest between these two. The XGBoost on TF-IDF matrix was then deployed on the second dataset (the Belgian dataset) also demonstrating high accuracy.

Further manual reclassification demonstrated that out of the false positive and false negative results that were able to be reclassified from the second dataset, up to 30% were due to erroneous or incomplete ICD coding.

It should be noted that the ratio of false positives to false negatives for algorithm errors (2.8 [56/20]) is higher compared to the ratio of false positives to false negatives for ICD-coding errors (1.3 [18/14]). This indicates that the algorithm is more likely to result in false positives than false negatives (i.e. it has a high sensitivity), which is beneficial in the presented use case: false positives can be easily manually reviewed by an ICD coder, while false negatives are likely to remain missed if not detected by the manual ICD-coding process.

Limitations

This study has certain limitations. First, the Belgian dataset that was used was first constructed using all patients that received a PCI intervention in a certain period for other purposes, after which it was reused for the scope of this study. The fact that patients underwent a PCI was not directly relevant for the research question. In future studies a dedicated data extraction step could be considered.

Second, there were limitations in the availability of ICD codes in the Belgian hospital before 2016. As such, while full reports were available for the period of 2012–20, ICD codes were available only for the period of 2016–20, resulting in a relatively small sample size. Still, due to the use of an external dataset for training purposes, relevant analyses could be performed.

Third, the BioBERT algorithm had limitations such as a limitation of 512 input tokens and no development on scientific biomedical language without finetuning on clinical medical language. Recent advances in the deep learning field (e.g. large language models) could possibly perform analyses with more accurate performance. Recent research has demonstrated that the use of generative fine-tuning could further

Table 1 Precision (positive predictive value), recall (sensitivity), and accuracy for different methods in atrial fibrillation and heart failure

	Atrial fibrillation			Heart failure		
	Precision	Recall	Accuracy	Precision	Recall	Accuracy
<i>MIMIC-III</i>						
Rule-based	0.98	0.45	0.72	0.93	0.53	0.75
Logistic regression on TF-IDF matrix	0.97	0.89	0.93	0.88	0.84	0.86
XGBoost on TF-IDF matrix	0.96	0.92	0.94	0.88	0.87	0.87
BioBERT	0.91	0.74	0.84	—	—	—
<i>Belgian dataset</i>						
XGBoost on TF-IDF matrix	0.83	0.87	0.94	0.79	0.77	0.92

Table 2 Manual reclassification of misclassifications

	Atrial fibrillation		Heart failure		Total (n = 108)
	FP (n = 33)	FN (n = 9)	FP (n = 41)	FN (n = 25)	
Algorithm incorrect	23 (70%)	3 (33%)	33 (80%)	17 (68%)	76 (70%)
ICD code incorrect or incomplete	10 (30%)	6 (67%)	8 (20%)	8 (32%)	32 (30%)

FP, false positive; FN, false negative.

improve coding accuracy also in medical records where the ICD code is not explicitly mentioned in the medical note.¹² The use of generative models including large language models can be part of future research.

Fourth, for data privacy and data security reasons, all of the analyses described were performed on the local, secured servers of our hospital network using devices with limited computational power. The BioBERT algorithm training step was a step needing large capacities of computational power during an elaborate length of time. Due to this reason, it was chosen to only perform it on one of the two selected diagnosis (i.e. AF) and not to repeat it on the HF dataset. For the same reason, the analyses on the MIMIC-III dataset were performed on a down-sampled dataset with an equal number of positive and negative diagnoses for both AF and HF.

Fifth, the study is a proof-of-concept study in which only two diagnoses were selected. For routine applicability a more elaborate set of algorithms would be needed to directly support the ICD-coding process.

Sixth, for identifying misclassifications, only mismatches between the automated and manual classification were analysed. In those files in which there was a match between the automated and manual classification, it should be assumed that a small percentage is incorrectly identified by both methods, thus resulting in a match. These cases were not identified by the method used in this study.

Seventh, the specific aim of the algorithm was classification of disease and not identification of misclassification. Specifically detecting misclassifications could potentially be achieved by the use of confidence scores on the document level. This could be the subject of future research.

Eighth, it should be noted that different data availability and difference in definitions used (as mentioned in the section 'manual reclassification') was only detected when a manual review of a section of the files was performed and could not be detected in the bulk dataset in advance.

Ninth, previous research has demonstrated that language models such as BioBERT are known to mitigate the effect on negation words

and could in a negation scenario thus outperform the XGBoost algorithm.¹³ No specific analysis on negation was performed in our study. However, this could be the potential area of future work.

Tenth, worldwide many governments and/or insurance companies reimburse hospitals through a diagnosis-related group system. Changing ICD codes for a patient would impact the specific diagnosis-related group for the patient and thus the amount that can be reimbursed to the hospital or patient. Due to the complexity of this system, and the differences per country, no specific cost calculation could be performed based on the 2% improvement in ICD-coding accuracy.

Finally, due to the nature of the study only diagnoses were selected that were already coded routinely with the use of ICD codes. The scope of NLP goes further than extraction of diagnostic information and could be expanded to elements that are not typically coded by hospital administrative departments (detailed information about diagnoses, risk factors, comorbidities, etc.). This could be used in the scope of future studies.

Interpretation

NLP has been used as a tool for research in past studies with various techniques being applied.¹⁴ Only limited groups have specifically focused on using the technology to improve accuracy of ICD coding with varying accuracy.^{15–17} Many of these studies focus on deep learning techniques. Surprisingly, in our study the XGBoost on TF-IDF matrix algorithm was superior to the deep learning BioBERT algorithm. One hypothesis to explain this finding is that the BioBERT model was trained on scientific articles and abstracts, of which the vocabulary does not necessarily translate to clinically used language.

Also, while our findings demonstrate insufficient accuracy of all algorithms to fully automate or replace the ICD-coding process, the results simultaneously demonstrate that automated algorithms could optimize the current ICD-coding process. In our Belgian dataset it was found that up to 30% of the real inconsistencies between the NLP algorithm and the ICD coding was due to an erroneous or incomplete ICD-coding

process. This translates as 2% of the total dataset in which the ICD coding could be improved, optimized or completed with the use of a newly developed NLP algorithm.

With the growth of applications of ICD coding (financial, quality of care, clinical, and epidemiological research) also the importance of fully correct ICD coding is growing. There is thus enormous potential for algorithms that could improve the accuracy and increase the speed of the ICD-coding process.

Conclusion

A newly developed NLP algorithm was demonstrated to attain relatively high accuracy for classifying disease in medical records. NLP algorithms could be used to identify ICD-coding errors and optimize and support the ICD-coding process.

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health*.

Funding

P.D., H.K., and S.E.K. received funding through the Horizon 2020 CoroPrevention project, project number 848056. M.F. received funding through the Flanders Research Foundation FWO, file number 1SE1222N.

Conflict of interest: None declared

Data availability

The data underlying this article are available on request.

References

1. Hirsch JA, Nicola G, McGinty G, Liu RW, Barr RM, Chittle MD, et al. ICD-10: history and context. *AJNR Am J Neuroradiol* 2016;**37**:596–599.
2. Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, et al. Systematic review of discharge coding accuracy. *J Public Health (Oxf)* 2012;**34**:138–148.
3. Bosco-Lévy P, Duret S, Picard F, dos Santos P, Puymirat E, Gilleron V, et al. Diagnostic accuracy of the international classification of diseases, tenth revision, codes of heart failure in an administrative database. *Pharmacoepidemiol Drug Saf* 2019;**28**:194–200.
4. Champagne SJ. Medicare loses billions to billing errors (September 5, 2019). In: *Proceedings of the Ninth International Conference on Engaged Management Scholarship*, 2019. doi:10.2139/ssrn.3454108.
5. Portuondo JL, Harris AHS, Massarweh NN. Using administrative codes to measure health care quality. *JAMA* 2022;**328**:825–826.
6. Berman AN, Biery DW, Ginder C, Hulme OL, Marcusa D, Leiva O, et al. Natural language processing for the assessment of cardiovascular disease comorbidities: the cardio-canary comorbidity project. *Clin Cardiol* 2021;**44**:1296–1304.
7. Johnson AEW, Pollard TJ, Shen L, Lehman L, Wei H, Feng M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;**3**:160035.
8. Trienes J, Trieschnigg D, Seifert C, Hiemstra D. Comparing rule-based, feature-based and deep neural methods for de-identification of Dutch medical records. *CEUR Workshop Proc* 2020;**2551**:3–11.
9. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 2005;**67**:301–320.
10. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. p.785–794.
11. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**:1234–1240.
12. Yang Z, Kwon S, Yao Z, Yu H. Multi-Label few-shot ICD coding as autoregressive generation with prompt. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. p. 5366–5374.
13. van Aken B, Trajanovska I, Siu A, Mayrdorfer M, Budde K, Loeser A. Assertion Detection in Clinical Notes: Medical Language Models to the Rescue? *Association for Computational Linguistics (ACL)*; 2021. p.35–40. doi:10.18653/v1/2021.nlpmc-1.5.
14. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform* 2019;**7**:e12239.
15. Chen PF, Wang SM, Liao WC, Kuo LC, Chen KC, Lin YC, et al. Automatic ICD-10 coding and training system: deep neural network based on supervised learning. *JMIR Med Inform* 2021;**9**:e23230.
16. Diao X, Huo Y, Zhao S, Yuan J, Cui M, Wang Y, et al. Automated ICD coding for primary diagnosis via clinically interpretable machine learning. *Int J Med Inform* 2021;**153**:104543.
17. Zhou L, Cheng C, Ou D, Huang H. Construction of a semi-automatic ICD-10 coding system. *BMC Med Inform Decis Mak* 2020;**20**:67.