An exploratory data analysis: the performance differences of a medical code prediction system on different demographic groups

Heereen Shim[1,2,3], Dietwig Lowet[3], Stijn Luca[4] and Bart Vanrumste[1,2]
[1]Campus Group T, e-Media Research Lab, KU Leuven, Leuven, Belgium
[2]Department of Electrical Engineering (ESAT), STADIUS, KU Leuven, Leuven, Belgium
[3]Philips Research, Eindhoven, the Netherlands
[4]Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium
{heereen.shim, bart.vanrumste}@kuleuven.be
{dietwig.lowet}@philips.com
{stijn.luca}@ugent.be

Abstract

Recent studies show that neural natural processing models for medical code prediction suffer from a label imbalance issue. This study aims to investigate further imbalance in a medical code prediction dataset in terms of demographic variables and analyse performance differences in demographic groups. We use sample-based metrics to correctly evaluate the performance in terms of the data subject. Also, a simple label distance metric is proposed to quantify the difference in the label distribution between a group and the entire data. Our analysis results reveal that the model performs differently towards different demographic groups: significant differences between age groups and between insurance types are observed. Interestingly, we found a weak positive correlation between the number of training data of the group and the performance of the group. However, a strong negative correlation between the label distance of the group and the performance of the group is observed. This result suggests that the model tends to perform poorly in the group whose label distribution is different from the global label distribution of the training data set. Further analysis of the model performance is required to identify the cause of these differences and to improve the model building.

1 Introduction

Medical coding is the process of assigning standard codes, such as The International Classification of Diseases (ICD) codes, to each clinical document for documenting records and medical billing purposes. Even though medical coding is an important process in the healthcare system, it is expensive, time-consuming, and error-prone (O'malley et al., 2005).

Researchers have investigated approaches for automated ICD coding systems and there has been great progress with neural network architectures (Kalyan and Sangeetha, 2020). However, current state-of-the-art models still suffer from data imbalance issues: since the benchmark dataset is imbalanced in terms of assigned ICD codes, the model performances differ across ICD codes (Mullenbach et al., 2018; Li and Yu, 2020; Kim and Ganapathi, 2021; Vu et al., 2021; Ji et al., 2021). Moreover, a recent study argues that the performances of models tend to decrease when the ICD codes have fewer training instances (Ji et al., 2021).

Based on this observation from the literature (i.e., imbalanced ICD code distribution results in the performance imbalance between the ICD codes), the goal of this paper is to investigate the effect of the imbalance of different demographic groups in the training data set on the performances of the demographic groups. More specifically, we study the following questions: 1) Is a benchmark dataset for medical code prediction imbalance in terms of the data subject's demographic variables (i.e., age, gender, ethnicity, socioeconomic status)?; 2) If so, would it result in performance differences between demographic groups? To answer these questions, we analyse the benchmark dataset, reproduce one of the state-of-the-art models (Li and Yu, 2020), and analyse the performance of the model. To the best of our knowledge, this is the first attempt to study the demographic imbalance of the medical code prediction benchmark dataset and analyse the performance differences between demographic groups.

Our contribution is three-fold. Firstly, we analysed the medical code prediction benchmark dataset to investigate the underlying imbalance in the dataset (Section 4.1) and reproduced one of the state-of-the-art medical code prediction models proposed by Li and Yu (2020). Secondly, we propose sample-based evaluation metrics (Section. 3.4) to identify potential biases inside a model and potential risk of the bias (Section. 4.2). Thirdly, we propose a simple label distance metric to quantify the

Proceedings of the 4th Clinical Natural Language Processing Workshop, pages 93 - 102

differences in the label distribution between each group and the global data (Section. 3.2) and found that the label distance metric is strongly correlated with the performance negatively (Section. 4.3). We expect that these analytic results could provide a valuable insight to the natural language processing (NLP) research community working for clinical applications.

2 Data

This section includes the information on the benchmark dataset used and the details of pre-processing steps taken for preparing data for the experiments. Note that we followed the previous approach to reproduce the result from the literature. More details are explained in the following subsections.

2.1 MIMIC-III dataset

We used Medical Information Mart for Intensive Care (MIMIC-III v1.4.) dataset (Johnson et al., 2016)[1] for the experiments. MIMIC-III is the benchmark dataset that has been widely used to build a system for automated medical code prediction (Shi et al., 2017; Mullenbach et al., 2018; Li and Yu, 2020; Kim and Ganapathi, 2021). For medical code prediction, discharge summary texts[2] are used as inputs and corresponding ICD-9 codes[3] are used as output of a system. In other words, the medical code prediction is formulated as a multi-label classification where the ground truth of the given input includes one or more ICD-9 codes.

For benchmarking purposes, Mullenbach et al. (2018) provides script codes that pre-process the discharge summary text data and splits the dataset by patient IDs into training, validation, and testing sets[4]. Also, Mullenbach et al. (2018) creates two benchmark sets, with full ICD codes as well as with the top 50 most frequent ICD codes, which are denoted as `MIMIC-III full` and `MIMIC-III 50`, respectively. The `MIMIC-III full` dataset contains 52,728 discharge summaries with 8,921 unique ICD codes and the `MIMIC-III 50` dataset contains 11,368 discharge summaries with 50 unique ICD codes.

In this paper, we only consider the `MIMIC-III 50` dataset. Following the previous works (Li and Yu, 2020; Kim and Ganapathi, 2021; Vu et al., 2021), we used Mullenbach et al. (2018)'s scripts to split the data which results in 8,066 discharge summaries for training, 1,573 for validation, and 1,729 for testing. Additionally, we extracted patients' demographic information from the MIMIC-III dataset, including gender, age, ethnicity, and insurance type as a socioeconomic proxy.

2.2 Data pre-processing

Discharge Summary texts One of our objectives is to reproduce the results by Li and Yu (2020) and analyse the performance. Therefore, we followed the Li and Yu (2020)'s pre-processing steps which are the same as the work by Mullenbach et al. (2018). Data cleaning and pre-processing include the following steps: the discharge summary texts were tokenized, tokens that contain no alphabetic characters were removed, and all tokens were lowercased. All documents are truncated to a maximum length of 2500 tokens. More details can be found in the original paper (Mullenbach et al., 2018).

Demographic data In the MIMIC-III dataset, each unique hospital visit for a patient is assigned with a unique admission ID. Therefore we used admission ID to extract the demographic information of patients. The following steps were taken to pre-process the demographic data: firstly, age values are computed based on the date of birth data and the admission time data[5]. Secondly, the four most frequent values in ethnicity data, including 'WHITE', 'BLACK', 'ASIAN', 'HISPANIC', are being kept, whereas the remaining values are combined into one group and labelled as 'OTHER'. Thirdly, the three most frequent values in insurance type data, including 'Medicare', 'Private', 'Medicaid', are being kept, whereas the other values are combined into one group 'Other'.

3 Methods

3.1 Data analysis

We analysed the size, as well absolute as relative, of each group and investigated relationships between variables. Also, we analysed the length of discharge summary notes and the number of assigned ICD codes per note to investigate relation-

[1] https://physionet.org/content/mimiciii/1.4/

[2] A discharge summary is a note that summarises information about a hospital stay

[3] MIMIC-III dataset includes both diagnoses and procedures which occurred during the patient's stay

[4] https://github.com/jamesmullenbach/caml-mimic

[5] The date of birth data of patients older than 89 have been shifted and the original values cannot be recovered. Therefore, we assigned the same age value of 90 to all patients who are older than 89.

ships between the length of notes and demographic variables and between the number of ICD codes per note and demographic variables. We also calculate the differences in the ICD code label distributions between the entire data and each group.

3.2 Label distribution distance metric

To calculate the differences in the ICD code label distributions between the entire data and each group, we used cosine distance[6] between ICD code label representations, each of which is a multi-hot vector $\mathbb{R}^{1\times 50}$. Specifically, we compute the average distances between the globally averaged label vector and the label vector of each data point in groups, which is defined as:

$$D_k = \frac{1}{N_k}\sum_{i}^{N_k} 1 - \frac{\boldsymbol{u}\cdot\boldsymbol{v_i}}{||\boldsymbol{u}||_2||\boldsymbol{v_i}||_2} \quad (1)$$

where $\boldsymbol{u}$ is the globally averaged label vector of the entire data and $\boldsymbol{v_i}$ is a label vector of a single data point in the group k that contains N_k of data points. A low distance score means the group contains patients whose label set is close to the global label distribution of the entire data.

3.3 Medical code prediction model

In this study, we study one of the state-of-the-art medical code prediction models proposed by Li and Yu (2020). There are three important architectural details in Li and Yu (2020)'s model: firstly, it uses a convolutional layer with multiple filters where each filter has a different kernel size (Kim, 2014). This multi-filter convolutional layer allows a model to capture various text patterns with different word lengths. Secondly, residual connections (He et al., 2016) are used on top of each filter in the multi-filter convolutional layer. This residual convolutional layer enlarges the receptive field of the model. Thirdly, the label attention layer (Mullenbach et al., 2018) is deployed after the multi-filter convolutional layer. More details on the model architecture can be found in the original paper (Li and Yu, 2020). For implementation, we re-trained a model by using a script[7] and followed the same hyperparameter setting except the early-stopping setting: we used a macro-averaged F1 score as an early-stopping criterion with a patience value 10.

[6]We used cosine distance because it is widely used to calculate the similarity between high-dimensional vectors and the distance is always normalised between 0 and 1.

[7]https://github.com/foxlf823/Multi-Filter-Residual-Convolutional-Neural-Network

3.4 Evaluation metrics

Performance metrics To evaluate the model's performance, micro-and macro-averaged F1 scores are widely used in the literature (Shi et al., 2017; Mullenbach et al., 2018; Li and Yu, 2020). Micro-averaged scores are calculated by treating each <text input, code label> pair as a separate prediction. Macro-averaged scores are calculated by averaging metrics computed per label. For recall, the metrics are computed as follows:

$$\text{Micro-R} = \frac{\sum_{l=1}^{L}\text{TP}_l}{\sum_{l=1}^{L}\text{TP}_l + \text{FN}_l} \quad (2)$$

$$\text{Macro-R} = \frac{1}{|L|}\sum_{l=1}^{L}\frac{\text{TP}_l}{\text{TP}_l + \text{FN}_l} \quad (3)$$

where TP_l and FN_l, denote true positive examples and false negative examples for a specific ICD-9 code label l, respectively. Since we use `MIMIC-III 50` dataset, $|L|$ equals 50

Since we focus on performance differences in terms of data subject's demographics, we additionally use sample-averaged F1 scores. Sample-averaged scores are calculated by computing scores at the instance level and averaging over all instances in the data set. For sample-averaged recall, the metric is computed as follows:

$$\text{Sample-R} = \frac{1}{|N|}\sum_{n=1}^{N}\frac{|\boldsymbol{y}_n \cap \hat{\boldsymbol{y}_n}|}{|\boldsymbol{y}_n|} \quad (4)$$

where $\boldsymbol{y}_n$ and $\hat{\boldsymbol{y}_n}$ denote the ground truth labels and the predicted labels for the n-th test example, respectively and N denotes the total number of test samples. Precision is computed in a similar manner.

For statistical analysis, we conducted the Kruskal-Wallis tests to investigate differences between the average performance scores of each group. Also, we computed the Pearson correlation coefficient and p-value for testing the correlation between the training data size of the group and the model performance on the group and between label distance of the group and the model performance on the group. All statistical tests were done by using sample-F1 scores.

Error metrics Following previous studies (Hardt et al., 2016; Chouldechova, 2017), we consider two metrics to quantify the error of a trained model: false negative rate (FNR) and false positive rate

	Count (n)	Percentage (%)
Total	8066	
Gender		
F	3593	44.5
M	4473	55.5
Age		
0-17	440	5.5
18-29	300	3.7
30-49	1148	14.2
50-69	2931	36.3
70-89	2817	34.9
90+	430	5.3
Ethnicity		
WHITE	5651	70.1
OTHER	1097	13.6
BLACK	799	9.9
HISPANIC	311	3.9
ASIAN	208	2.6
Insurance		
Medicare	4440	55.0
Private	2636	32.7
Medicaid	709	8.8
Other	281	3.5

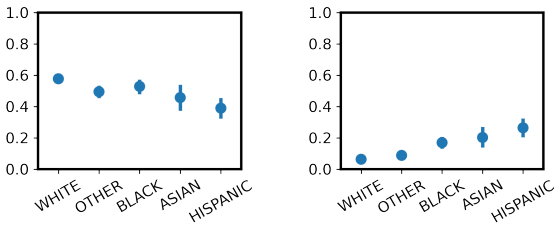
Table 1: Sample size (absolute and relative) of the groups of gender, age, ethnicity, and insurance type.

(FPR) in the sample level. FNR is the fraction of ICD codes that are failed to be predicted by a system but included in a ground truth label set. FPR is the fraction of ICD codes that are erroneously predicted by a system but not included in a ground truth label set. High FNR scores imply low recall scores and high FPR implies low precision scores. Two metrics are computed as follows:

$$\text{FNR} = \frac{1}{|N|}\sum_{n=1}^{N} 1 - \frac{|\boldsymbol{y}_n \cap \hat{\boldsymbol{y}}_n|}{|\boldsymbol{y}_n|} \quad (5)$$

$$\text{FPR} = \frac{1}{|N|}\sum_{n=1}^{N} 1 - \frac{|\boldsymbol{y}_n \cap \hat{\boldsymbol{y}}_n|}{|\hat{\boldsymbol{y}}_n|} \quad (6)$$

To assess the risk of errors, we use the worst-case comparison method (Ghosh et al., 2021). Also, we conducted Mann–Whitney U tests to investigate the differences between the error scores of the best and the error scores of the worst models.



(a) Percentage of Medicare within each ethnic group

(b) Percentage of Medicaid within each ethnic group

Figure 1: Relationship between insurance and demographic variables. 95% confidence intervals are illustrated by lines.

4 Results

4.1 Data analysis results

Table 1 summarizes the sample sizes of the data set. It is shown that only gender variables are well-balanced. For age groups, patients who are 50-89 take up to 71.2% of the data. Also, the data set includes more White patients than patients from other ethnic groups. Also, more than half of the entire patients in the data set are patients with Medicare insurance and only 8.8% of patients are with Medicaid insurance.

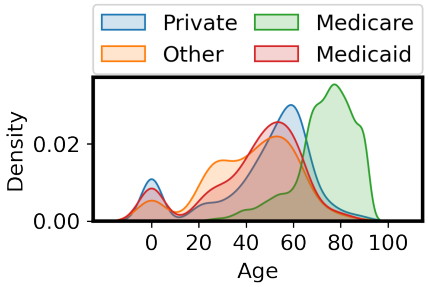
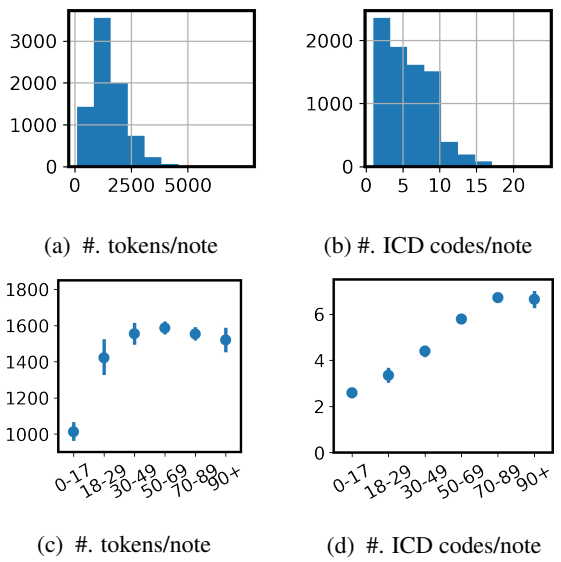


Figure 2: Kernel density estimate plot for visualising the age distribution of each insurance type

Figure 1 shows the relationship between insurance types, Medicare and Medicaid, and ethnicity variables. It is observed that insurance type has a certain relationship with the patient's race: 57.7% of White patients are paying with Medicare, whereas 38.9% of Hispanic patients are paying with Medicare. On the other hand, 26.4% of Hispanic patients are paying with Medicaid, whereas only 0.63% of White patients are paying with Medicaid.

Figure 2 illustrate the age distribution of each insurance type. Medicare and Medicaid are two separate, government-run insurance in the United States. Medicare is available for people age 65 or above and younger people with severe illnesses and



(a) #. tokens/note (b) #. ICD codes/note

(c) #. tokens/note (d) #. ICD codes/note

Figure 3: The distribution of the length of a discharge summary note (a) and the number of ICD codes assigned per note (d). Relationship between the length of notes and age groups (c) and between the number of ICD codes per note and age groups (d). X-axes indicate the average number of tokens in a note (a, c) and the average number of ICD codes per note (b,d). 95% confidence intervals are illustrated by lines.

	Distance
Gender	
F	0.613 (0.137)
M	0.615 (0.133)
Age	
0-17	0.737 (0.097)
18-29	0.746 (0.111)
30-49	0.684 (0.133)
50-69	0.610 (0.129)
70-89	0.564 (0.116)
90+	0.560 (0.118)
Ethnicity	
WHITE	0.610 (0.135)
OTHER	0.607 (0.131)
BLACK	0.633 (0.135)
HISPANIC	0.646 (0.135)
ASIAN	0.626 (0.143)
Insurance	
Medicare	0.579 (0.124)
Private	0.653 (0.135)
Medicaid	0.658 (0.136)
Other	0.691 (0.139)

Table 2: Average label distribution distances between each group and the global data. Standard deviations are added in parentheses.

Medicaid is available to low-income individuals under the age of 65 and their families. Because of the eligibility criteria for Medicare, Medicare includes more older patients compared to other insurance types, as we can see from the Figure 2.

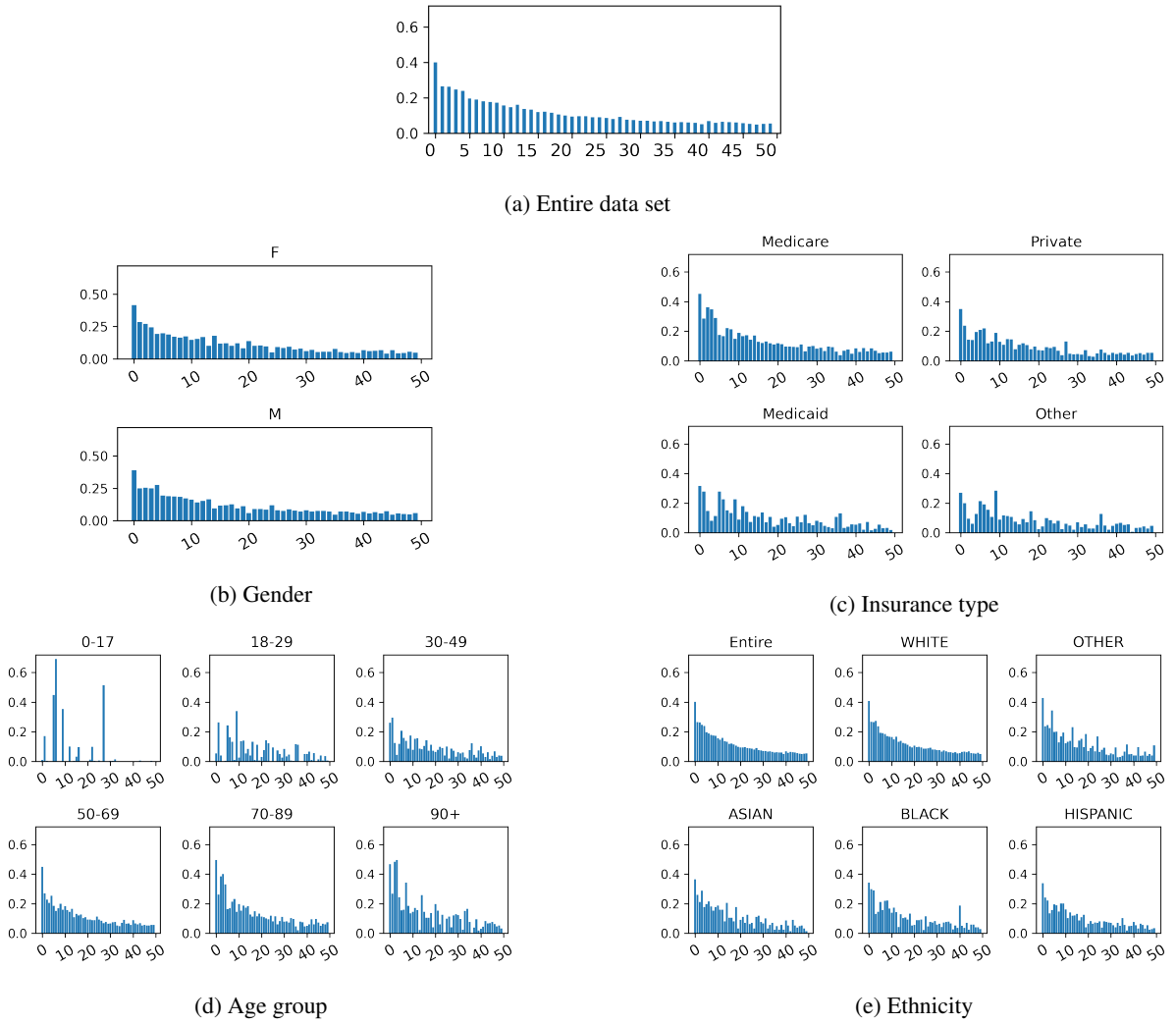
Figure 3a and Figure 3b show the distribution of the length of a discharge summary note and the number of ICD codes assigned per note, respectively. The average length is 1529.7 (std=754.9) and the average number of codes per note is 5.7 (std=3.3). Figure 3c and Figure 3d illustrate relationship between patients age and the length of note and the number of codes per note, respectively. From Figure 3c, it is observed that the length of note tends to increase until age group 50-69 and starts to decrease afterwards. From Figure 3d, positive correlations between age and the number of ICD codes per note are observed. Other noticeable patterns are not observed in other demographic variables (i.e., gender, insurance, ethnicity) with the respect to the length of a discharge summary note and the number of ICD codes assigned per note.

Figure 4 illustrates ICD code distributions. Figure 4a shows the entire data set has long-tail distribution. Between female and male patient groups, no noticeable difference between the label distributions is not observed. In terms of insurance type and ethnicity, each group shows slightly different ICD code distributions. Clear differences are observed between age groups: patients whose ages are younger than 30 (0-17, 18-29) show less spread ICD code distributions with fewer ICD codes than other age groups. The label distribution distances between each group and the global data are summarised in Table 2. Similar to the observations from Figure 4, age groups 0-17 and 18-29 have the bigger distance scores.

4.2 Performance & error analysis results

Table 3 summarises the prediction results on the test set. It is observed that a re-trained model slight underperforms compared to the original model (Li and Yu, 2020). The different early-stopping settings might cause this difference. Both models achieve higher scores in micro-averaged metrics than macro-averaged metrics, which means the model's performance on rare labels is worse than on frequent labels. The sample-averaged metrics are higher than macro-averaged metrics but lower than micro-averaged metrics.

Noticeable performance differences are observed between age groups, especially between patients



(a) Entire data set

(b) Gender

(c) Insurance type

(d) Age group

(e) Ethnicity

Figure 4: ICD code distribution. X-axis indicates the sorted ICD code class label and Y-axis indicate the percentage of labels observed in the training set.

younger than 30 years (18-29) and older than 90 (90+). The percentages of both groups in the training set are low but patients younger than 30 years get distinctively worse predictions in terms of all F-1 scores. Between different ethnic groups, it is observed that Hispanic and Asian patients get worse predictions compared to other patients. Between insurance types, it is also observed that patients with other types of insurance and Medicaid insurance get worse predictions compared to patients with Medicare and Private insurance in sample-averaged F-1 scores.

As the result of the Kruskal-Wallis test, we found statistically significant differences in sample-averaged F1 scores according to age group ($H(4)=46.57$, $p<0.001$) and insurance type ($H(3)=18.58$, $p<0.001$), separately. Close to being statistically significant is found according to gender ($H(1)=3.65$, $p=0.056$) and no statistically significant difference is found according to ethnicity ($H(4)=2.657$, $p=0.657$).

Error metrics per group are summarised in Table 4. Error metrics between groups show a similar trend as the performance metrics: differences between age groups are the most pronounced. It is observed that FNR scores tend to decrease as age increases. However, the largest difference between age groups is not significant ($p=0.06$). FPR also tends to increase as the age increases in the age groups under 90 and the largest difference between the younger group (18-29) and the older group (70-89) is significant ($p<0.001$). Patients with other types of insurance take significantly worse scores compared to Medicare patients in terms of FNR scores. Interestingly, FPR shows different patterns. For example, patients with Medicare get the worst FPR scores and patients with Private insurance get the best FPR scores.

	F-1 (%)		
	Micro	**Macro**	**Sample**
Total			
Li and Yu (2020)	67.3†	60.8†	-
Reproduced	64.4	59.2	60.6
Gender			
F (44.5)	63.2	58.1	59.7
M (55.5)	**65.3**	**59.4**	**61.4**
Age			
18-29 (3.7)	53.9	36.1	48.2
30-49 (14.2)	58.9	58.2	52.4
50-69 (36.3)	64.2	57.7	60.9
70-89 (34.9)	65.6	**59.2**	63.6
90+ (5.3)	**67.1**	55.9	**65.0**
Ethnicity			
WHITE (70.1)	64.3	59.2	60.8
OTHER (13.6)	64.3	**60.9**	60.7
BLACK (9.9)	**66.2**	60.2	**61.7**
HISPANIC (3.9)	62.0	54.6	56.0
ASIAN (2.6)	64.7	51.2	59.3
Insurance			
Medicare (55.0)	**65.3**	58.4	**62.5**
Private (32.7)	63.4	58.8	59.0
Medicaid (8.8)	62.9	**59.3**	57.8
Other (3.5)	56.0	49.3	50.5

Table 3: Performances on the MIMIC-III 50 test set. † indicates performances reported in the paper by Li and Yu (2020). Other results are obtained from a reproduced model. The percentage of training samples (%) is added in parentheses after the group labels. Best performances are boldfaced and worst performances are underlined.

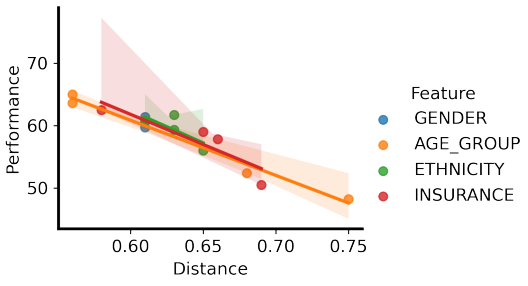


Figure 5: Label distance of each group and the model performance on each group. Linear relationships are illustrated by lines determined through linear regression.

4.3 Correlation test result.

As the result of correlation tests, we found a weak positive correlation (0.43, p=0.09) between training set size and performance. This result shows that even though the model performs well for groups

	FNR (%)	FPR (%)
Total	40.6	3.8
Gender		
F (44.5)	39.7	4.3
M (55.5)	**38.0**	**4.2**
largest diff. (↓)	1.7	0.1
smallest ratio (%) (↑)	95.8	98.2
Age		
18-29 (3.7)	46.2	**2.9**
30-49 (14.2)	45.9	3.3
50-69 (36.3)	39.5	3.9
70-89 (34.9)	35.7	5.0
90+ (5.3)	**34.1**	4.4
largest diff. (↓)	12.2	2.1***
smallest ratio (%) (↑)	73.7	57.7
Ethnicity		
WHITE (70.1)	38.7	4.2
OTHER (13.6)	39.3	4.5
BLACK (9.9)	**37.0**	4.2
HISPANIC (3.9)	42.5	4.2
ASIAN (2.6)	40.3	**3.8**
largest diff. (↓)	5.4	0.8
smallest ratio (%) (↑)	87.2	83.3
Insurance		
Medicare (55.0)	**37.0**	4.7
Private (32.7)	40.7	**3.4**
Medicaid (3.5)	41.0	3.6
Other (8.8)	46.9	4.2
largest diff. (↓)	9.8*	1.3***
smallest ratio (%) (↑)	79.0	71.5

Table 4: Errors on the MIMIC-III 50 test set. The percentage of training samples (%) is added in parentheses. Best performances are boldfaced and worst performances are underlined. * and *** indicate the error of the worst model is greater than the error of the best with statistical significance of p=0.05 and p=0.001 (Mann–Whitney U test), respectively.

with more training data in general, the relationship is not statistically significant. Contrary to this result, we found a very strong negative correlation (-0.95, p<0.001) between label distance and performance. This result implies that the model performs poorly in the groups containing many patients whose label set is different from the global label distribution of the entire data. The group-specific correlations between label distances and the performances are illustrated in Figure 5. It is observed that the negative correlation is much more pronounced between different age groups than in other groups.

5 Discussion

Impact of the study. The MIMIC-II dataset for medical code prediction provides opportunities to develop and benchmark models and facilitates natural language processing research in the clinical domain. Since it is one of the most frequently used benchmark datasets for medical code prediction, it has a huge impact on the quality of the developed models. For example, previous studies (Mullenbach et al., 2018; Li and Yu, 2020; Kim and Ganapathi, 2021; Vu et al., 2021; Ji et al., 2021) have shown that the ICD code distribution in the MIMIC-III dataset is imbalanced and it results in performance differences between ICD codes. In this study, we investigated the data imbalance of the MIMIC-III 50 data, in terms of the data subject's demographic factors, and its effect on the model performance for ICD code prediction.

Evaluation metrics for fairness. In this paper, we proposed metrics that can correctly evaluate the model's performance in terms of individual patients' benefits and potential harms. Especially, we formulated the medical code prediction task as a multi-label classification task. From a machine learning perspective, sample-based metrics and label-based metrics are used to evaluate the performance of a model in a multi-label classification task (Zhang and Zhou, 2013). Sample-based and label-based metrics focus on different aspects of model performance, one in sample-wise performance and the other in label-wise performance. However, label-based metrics are more frequently used in the literature (Xiao et al., 2018; Mullenbach et al., 2018; Li and Yu, 2020; Kim and Ganapathi, 2021; Vu et al., 2021; Ji et al., 2021). Considering a healthcare application setting where all patients are expected to receive an equal quality of service, we argue that using sample-based metrics is required to evaluate the model performance. Also, we propose to use disaggregated metrics (Barocas et al., 2021), which are metrics evaluated on each group of data, to ensure that a model is equally accurate for patients from different demographic groups (Rajkomar et al., 2018; Gichoya et al., 2021).

Correlation between demographic variables We analysed the MIMIC-III dataset to identify the underlying data imbalance of demographic variables. Our data analysis results show that the MIMIC-III dataset is imbalanced in terms of the data subject's demographics. However, we also found a correlation between demographic variables. For example, age is correlated with insurance type: patients older than 65 are likely to be insured with Medicare. This confounding factor across demographic variables makes it complicated to interpret the main effects of the data subject's demographics on the model performance.

Correlation between label distance and performance Based on the previous study arguing the performances of models tend to decrease when the ICD codes have fewer training samples (Ji et al., 2021), we hypothesised that the performance of the model on a demographic group is correlated with the number of data of that group in the training data set. However, the analysis results do not support this hypothesis: even though the performance differences are observed across some demographic groups (i.e., across age groups and insurance types), the correlation between the number of training data of the group and the performance of the group is weak. Instead, we found that the label distance of the group is negatively correlated with the performance of the group. This result suggests that when the group contains patients whose label set is different from the global label distribution of the entire data, it is likely that the model performs poorly in that group.

In terms of machine learning perspective, this issue can be seen as a label shift: the train and test label distribution is different while the feature distribution remains the same (Lipton et al., 2018; Guo et al., 2020). To address this issue, one interesting area for future work may be in re-training the classifier with adjusted training sample weights (Lipton et al., 2018) or adapting the predictions of a pre-trained classifier (Saerens et al., 2002; Du Plessis and Sugiyama, 2014; Alexandari et al., 2020).

Limitations and future directions There are several limitations to this study. Firstly, we used a subset of MIMIC-III data with the top 50 most frequent ICD codes to simplify the analysis. Since the full MIMIC-III dataset contains more than 47,000 ICD codes, further study is required. Secondly, we only studied the model proposed by Li and Yu (2020). One potential direction is to investigate the performance of models using pre-trained language models (Zhang et al., 2020; Ji et al., 2021). Thirdly, we found an issue of confounding across demographic variables, which makes it complicates the interpretation of the main effects of the data

subject's demographic factors on the model performance. To address this issue, further analysis of multiple intersectional groups or causal analysis is required. In future work, we will also investigate how to build a model that can perform equally well on across all demographic groups.

6 Conclusion

In this study, we performed an empirical analysis to investigate the data imbalance of the MIMIC-III 50 dataset and its effect on the model performance for ICD code prediction. We found that demographic imbalance exists in the MIMIC-III 50 dataset and a medical code prediction model performs differently across some demographic groups. Interestingly, the correlation between the number of training data of the group and the performance of the group is weak. Instead, we found a negative correlation between the label distance of the group and the performance of the group. This result suggests that the model tends to perform poorly in the group whose label distribution is different from the global label distribution. Potential future research direction includes further analysis of the main effects of the data subject's demographic factors on the model performance and investigation of building a robust and fair model that can perform equally well across demographic groups with different label distributions.

Acknowledgements

We thank anonymous reviewers for providing valuable feedback on this work. Data processing activities were conducted by the first author and the other authors did not involve in the raw data processing. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 766139. This article reflects only the author's view and the REA is not responsible for any use that may be made of the information it contains.

References

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. 2020. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning*, pages 222–232. PMLR.

Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W Duncan Wadsworth, and Hanna Wallach. 2021. Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 368–378.

Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Marthinus Christoffel Du Plessis and Masashi Sugiyama. 2014. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks*, 50:110–119.

Avijit Ghosh, Lea Genuit, and Mary Reagan. 2021. Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, pages 22–34. PMLR.

Judy Wawira Gichoya, Liam G McCoy, Leo Anthony Celi, and Marzyeh Ghassemi. 2021. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health & Care Informatics*, 28(1).

Jiaxian Guo, Mingming Gong, Tongliang Liu, Kun Zhang, and Dacheng Tao. 2020. Ltf: A label transformation framework for correcting label shift. In *International Conference on Machine Learning*, pages 3843–3853. PMLR.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Shaoxiong Ji, Matti Hölttä, and Pekka Marttinen. 2021. Does the magic of BERT apply to medical code assignment? A quantitative study. *Computers in Biology and Medicine*, 139.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Katikapalli Subramanyam Kalyan and Sivanesan Sangeetha. 2020. Secnlp: A survey of embeddings in clinical natural language processing. *Journal of biomedical informatics*, 101:103323.

Byung-Hak Kim and Varun Ganapathi. 2021. Read, Attend, and Code: Pushing the Limits of Medical Codes Prediction from Clinical Notes by Machines. *Proceedings of Machine Learning Research*, 149:1–12.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Fei Li and Hong Yu. 2020. ICD coding from clinical text using multi-filter residual convolutional neural network. In *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, volume 34, pages 8180–8187. AAAI press.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. 2018. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 1101–1111.

Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. 2005. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. 2002. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P Xing. 2017. Towards automated icd coding using deep learning. *arXiv preprint arXiv:1711.04075*.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2021. A label attention model for icd coding from clinical text. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3335–3341.

Cao Xiao, Edward Choi, and Jimeng Sun. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428.

Min-Ling Zhang and Zhi-Hua Zhou. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.

Zachariah Zhang, Jingshu Liu, and Narges Razavian. 2020. Bert-xml: Large scale automated icd coding using bert pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 24–34.