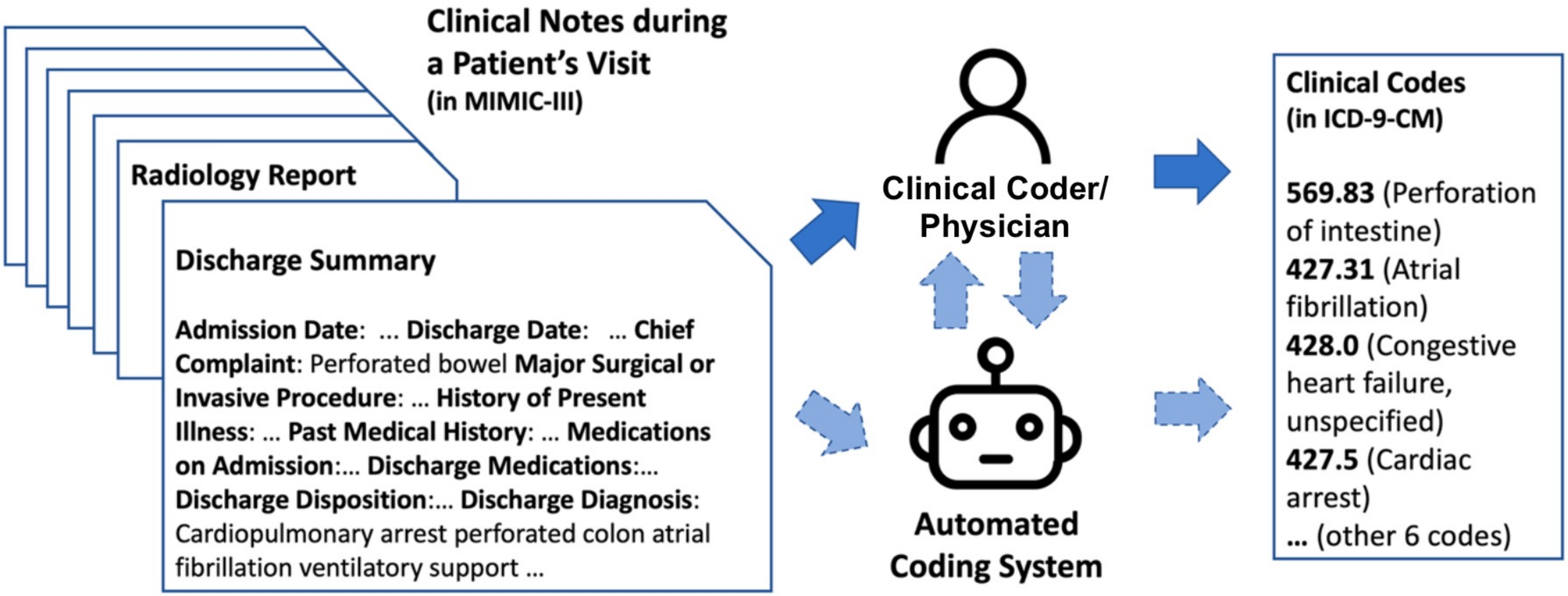
Automatic ICD Code Generation Using Discharge Summaries

Team19 - Sriharsha Devarapu, Akhil Perumal Reddy and Xianshi Yu

2023 May 03

Automate International Classification of Diseases (ICD) code assignment



[Dong et al 2022]

Automate ICD code assignment--Motivation

- Manual coding is **time-consuming**
  - Large code pool: 14,025 (68,069) for ICD-9 (ICD-10) diagnosis codes
  - Coder in NHS Scotland codes about 60 cases a day (7–8min for each case)

- Manual coding may be **prone to error**
  - Subjectivity in choosing the diagnosis codes
  - Data entry errors

- Accurate code assignment is important
  - ICD codes used in billing and reimbursements
  - Health condition monitoring & policy decisions
  - Risk prediction modeling

Related Work & Challenges

- Challenges
  - Variation in length of discharge summaries, informal structure, and notation
  - Discharge summary’s length can be very long; difficult to fit into standard pre-trained BERT models
  - Majority of codes rarely observed due to the dimensionality of the label space
  - Label size is extensive (14,025 ICD-9 and 68,069 for ICD-10 for diagnosis codes only)
  - Every discharge summary is mapped to a set of multiple ICD codes – multi-class classification

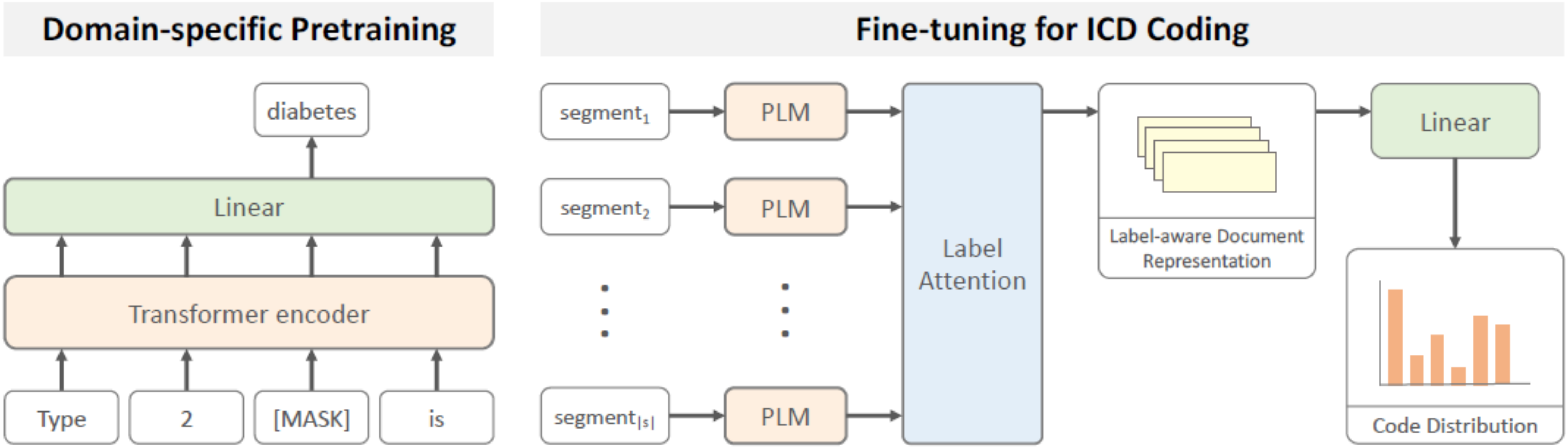
- Variants of CNNs and LSTMs models have shown significant potential
- Most BERT-based approaches still do not outperform CNN-based methods, except for the PLM-ICD model [ClinicalNLP, July 2022]

PLM-ICD Model

Architecture has three main components

- Domain-specific training:
  - General-purpose PLMs can’t understand the medical text.
  - Utilize a pre-trained PLM (Bio+Discharge Summary BERT)
- Segment Pooling:
  - Segment the whole document into shorter segments
  - Encode the segments with PLM and then aggregate
- Label-Aware Attention:
  - Obtain a label-wise attention weight matrix: A
  - Use matrix A to compute a weighted sum of Hidden representation (H), which generates the label-specific document D
  - Apply the sigmoid function on the inner product between D and label weights (L)

PLM-ICD Architecture



$$Z = \tanh(VH)$$
$$A = (\mathrm{softmax}(WZ))^T$$
$$D = HA$$

$$p_i = \sigma(\langle L_i, D_i \rangle)$$

sizes:
Z, H=[h,t]
A=[t,c]
D=[h,c]
P=[c]

[Huang et al 2022]

Reimplementation & Extensions

Dataset & Reimplementation

MIMIC-III (Medical Information Mart for Intensive Care III) Clinical Database

- Health-related data form Beth Israel Deaconess Medical Center
- Over **40,000** patients who stayed in critical care units between 2001 and 2012

Pre-processing

- All text converted to *lowercase, tokenized, and filtered* to include only non-numeric characters
- Number of word tokens in each input text/document ranges from 9 to 7,504 (with a mean of **1,338**)
- Total number of unique ICD codes are **8,994** (only a few codes occur very frequently)
- Train/Dev/Test split: 80%, 10%, 10% of discharge summaries (one for each admission)

Independent implementation (for Top-50 ICD Codes)

Model	Macro-F1	Micro-F1
PLM-ICD	64.9	69.3
PLM-ICD-re	59.52	64.6

Reasons for difference

- Epochs (5 versus 20)
- Batch size (4 v/s 8)
- Learning rate scheduler

Using texts in code definitions

- Code definitions publicly available at cms.gov*

```
005.0  Staphylococcal food poisoning
005.1  Botulism food poisoning
005.2  Food poisoning due to Clostridium perfringens (C. welchii)
```

- If word in code definition appears in discharge summary, more like code should be assigned

Two approaches

1. Pretrain PLM-ICD model using code definitions
2. Add code definition into training data (use definition text to predict code)

	Macro-F1	Micro-F1	Macro-Auc	Micro-Auc
PLM-ICD (vanilla)	11.2	59.2	92.5	98.9
Method 2	12.1	**59.5**	94.3	99.1
Methods 1 & 2	**12.8**	59.4	94.5	99.1
RAC [PLMR 2021]	12.7	58.6	**94.8**	**99.2**

* Centers for Medicare & Medicaid Services

Modified C-HMCNN (h)[3]

- Exploits the **hierarchy information** to produce predictions coherent with the hierarchy constraint (The primary goal is to enhance the performance for low-frequent classes). (Macro F1 for Phenotype code prediction is 30.4)

- Two main updates to the current PLM-ICD model:-
  - A **constraint layer** is built on top of the existing network, ensuring that the predictions are coherent by construction
  - Updated **loss function** for exploiting the prediction on the higher class (parent(h_B) – example: 401) in the hierarchy to make predictions on the lower one (child (h_A) – example: 401.9)

$$Loss_A = -y_A \ln(MCM_A) - (1 - y_A)\ln(1 - MCM_A)$$

- Changes compared to the reference C-HMCNN (h):
  - Used **min constraint module (MCM)** instead of max constraint module
    - $MCM_A = Min(h_A, h_B),\ MCM_B = h_B$

  - For improved efficiency (in speed), we have used **mapping** of child to parents, instead of n x n mask for MCM & loss function layer calculations

Modified C-HMCNN (h) - Results

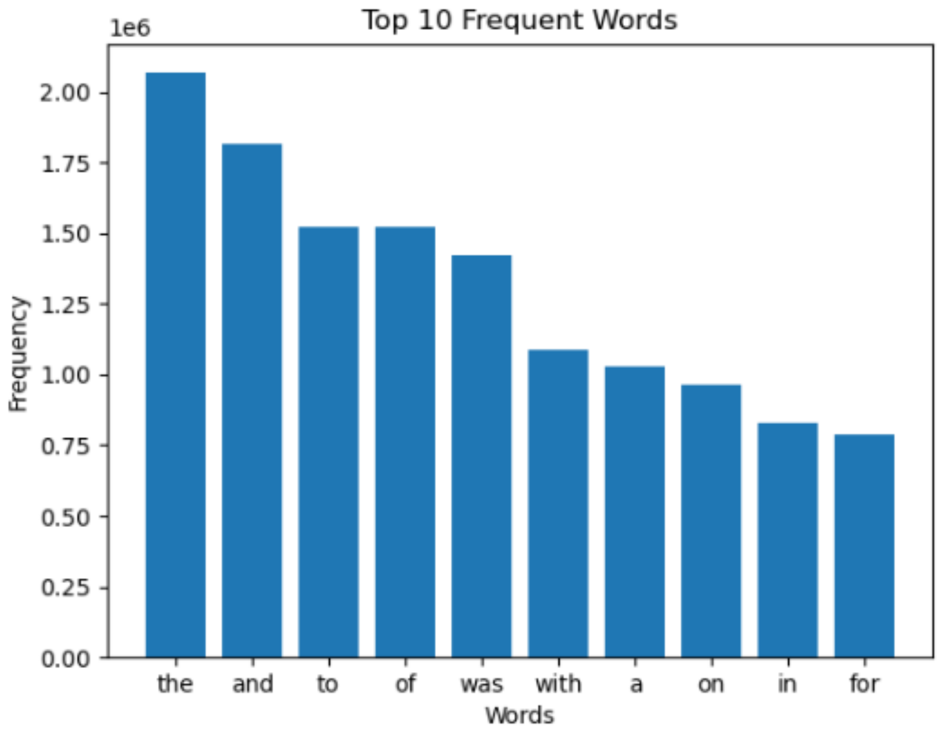
Comparison of performance between PLM ICD and modified version (by incorporating C-HMCNN) with the limited target label set (Top-50 ICD codes) (%)

Model	Macro-F1	Micro-F1
PLM-ICD-re	59.52	**64.6**
PLM-ICD-re-C-HMCNN	**59.9**	64.5

Removal of Unnecessary Words

- Removed unnecessary words from dataset which don't play important role in making ICD code predictions (e.g., "the", "of", "was", etc.)
- After removing 25% of the text content (from 30 most frequent words), model accuracy remained very similar to original model.
- More frequent words can be examined further to remove based on medical context

	Macro-F1	**Micro-F1**	**Macro-Auc**	**Micro-Auc**
PLM-ICD (vanilla)	11.2	59.2	92.5	98.9
Method 2	12.1	59.5	94.3	99.1
Methods 1 & 2	12.8	59.4	94.5	99.1
Method 1 & 2 & remove freq words	**13.0**	**59.6**	94.4	99.0
RAC [PLMR 2021]	12.7	58.6	**94.8**	**99.2**



Findings and Next-steps

- Code definition texts helpful for automated code assignment; simple method treating definition text as training data results in considerable improvement; will try put different weights on two types of observations (original & code definition) in loss function and use dev data to choose best weight

- Our findings indicate that including hierarchy information during modeling/prediction leads to a slight improvement in performance. However, further exploration of this approach could potentially yield even better results

- We can incorporate the class-weights information in the BCE loss function based on the below equations. This ensures the model to perform better when there is high imbalance present in the data.

$$dL_i = -[w_p * y\_true_i * \log\left(y\_pred_i\right) + w_n * \left(1 - y\_true_i\right) * \log\left(1 - log(y\_pred_i)\right)]$$

References

[1] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and MatthewB. A. McDermott. 2019. Publicly available clinicalbert embeddings. arXiv preprint arXiv:1904.03323.

[2] Hang Dong, Matúš Falis, William Whiteley, BeatriceAlex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen,and Honghan Wu. 2022. Automated clinical coding:what, why, and where we are? NPJ digital medicine,5(1):159.

[3] Eleonora Giunchiglia and Thomas Lukasiewicz.2020. Coherent hierarchical multi-label classificationnetworks. Advances in neural information processingsystems, 33:9662–9673.

[4] Chao-Wei Huang, Shang-Chi Tsai, and Yun-NungChen. 2022. Plm-icd: automatic icd codingwith pretrained language models. arXiv preprintarXiv:2207.05289.

[5] Patrick Lewis, Myle Ott, Jingfei Du, and VeselinStoyanov. 2020. Pretrained language models forbiomedical and clinical tasks: understanding and extendingthe state-of-the-art. In Proceedings of the3rd Clinical Natural Language Processing Workshop,pages 146–157.

[6] James Mullenbach, Sarah Wiegreffe, Jon Duke, JimengSun, and Jacob Eisenstein. 2018. Explainableprediction of medical codes from clinical text.

[7] Colin Raffel, Noam Shazeer, Adam Roberts, KatherineLee, Sharan Narang, Michael Matena, YanqiZhou, Wei Li, and Peter J Liu. 2020. Exploringthe limits of transfer learning with a unified text-totexttransformer. The Journal of Machine LearningResearch, 21(1):5485–5551.

[8] Chandan Sen, Bin Ye, Jawad Aslam, and Amir Tahmasebi.2021. From extreme multi-label to multiclass:A hierarchical approach for automated icd-10coding using phrase-level attention. arXiv preprintarXiv:2102.09136.

[9] Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen.2020. A label attention model for icd coding fromclinical text. pages 3335–3341. Main track.

[10] Wei-Qi Wei, Lisa A Bastarache, Robert J Carroll,Joy E Marlo, Travis J Osterman, Eric R Gamazon,Nancy J Cox, Dan M Roden, and Joshua C Denny.2017. Evaluating phecodes, clinical classificationsoftware, and icd-9-cm codes for phenome-wide associationstudies in the electronic health record. PloSone, 12(7):e0175508.

[11] Patrick Wu, Aliya Gifford, Xiangrui Meng, XueLi, Harry Campbell, Tim Varley, Juan Zhao, LisaBastarache, Joshua C. Denny, Evropi Theodoratou,and Wei-Qi Wei. 2018. Developing and evaluatingmappings of icd-10 and icd-10-cm codes to phecodes.bioRxiv.

[12] Byung-Hak Kim and Varun Ganapathi. 2021. Read, at- tend, and code: Pushing the limits of medical codes prediction from clinical notes by machines. In *Machine Learning for Healthcare Conference*, pages 196–208. PMLR.

Appendix

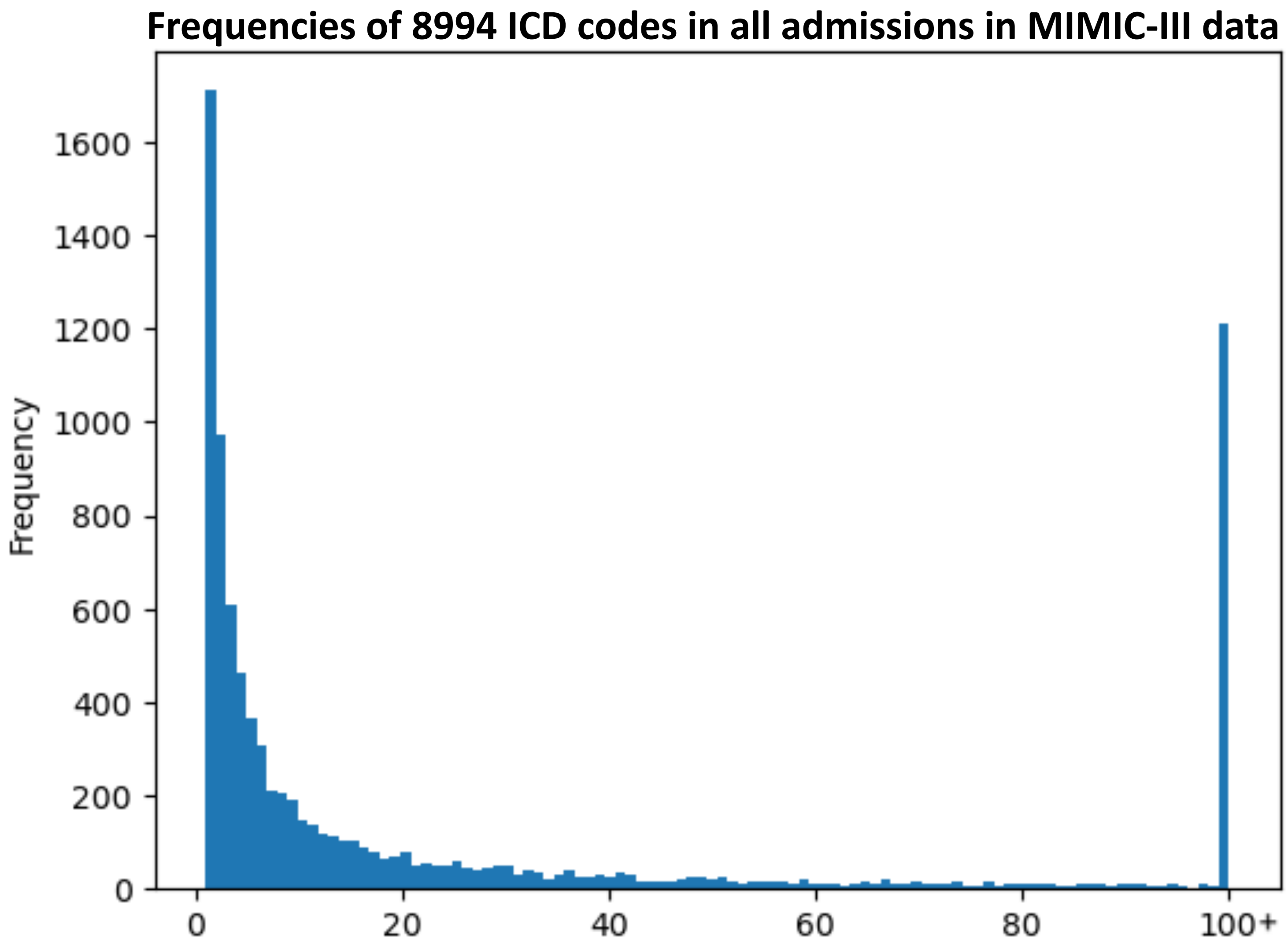
$$\mathrm{F1} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$ (harmonic mean of precision and sensitivity)

$$\text{Macro F1} = \frac{\sum_{1 \le i \le m} \mathrm{F1}_i}{m}$$

$$\text{Micro F1} = \frac{\text{Net } TP}{\text{Net } TP + \frac{1}{2}(\text{Net } FP + \text{Net } FN)}$$

Similarly, Macro (Micro) Auc is calculated using Macro (Micro) sensitivity and (Micro) specificity.

Appendix


Frequencies of 8994 ICD codes in all admissions in MIMIC-III data
Frequency
0
200
400
600
800
1000
1200
1400
1600
0
20
40
60
80
100+

Appendix

# Distinct Words	# Distinct Removed words	# Total words	# Total removed words
150 k	30	80 million	22 million

Additional Methods tried (data preprocessed by independently developed code)

	Macro-F1	Micro-F1	Macro-Auc	Micro-Auc
PLM-ICD (vanilla)	11.5	59.7	93.9	99.1
Method 1	12.6	59.8	94.1	99.1
Method 2	12.1	**59.8**	95.1	99.2
Combine Methods 1 & 2	**12.8**	59.6	95.5	**99.3**
Group-MinMax	12.6	55.2	**95.7**	99.2

Appendix

Performance of **phecode prediction** (1411 phecodes) using discharge summaries with the PLM-ICD model.

No. of epochs	Macro-F1	Micro-F1
3	12.9	57.9
10	28.0	63.1
20	30.4	61.8