Intriguing Effect of the Correlation Prior on ICD-9 Code Assignment

Zihao Yang[1,2], **Chenkang Zhang**[1,2], **Muru Wu**[1,2], **Xujin Chris Liu**[2,3],
Lavender Yao Jiang[1,2], **Kyunghyun Cho**[1,4,5,6], **Eric Karl Oermann**[2,7,8,1]
[1]Center for Data Science, New York University
[2]Department of Neurosurgery, NYU Langone Health
[3]Department of Electrical and Computer Engineering, NYU Tandon School of Engineering
[4]Courant Institute of Mathematical Sciences, New York University
[5]Canadian Institute for Advanced Research
[6]Prescient Design
[7]Department of Radiology, NYU Langone Health
[8]Neuroscience Institute, NYU Langone Health
`{gavin.yang,stephen.zhang,wm1077,chris.liu,lyj2002,kyunghyun.cho}@nyu.edu,`
`eric.oermann@nyulangone.org`

Abstract

The Ninth Revision of the International Classification of Diseases (ICD-9) is a standardized coding system used to classify health conditions. It is used for billing, tracking individual patient conditions, and for epidemiology. The highly detailed and technical nature of the codes and their associated medical conditions make it difficult for humans to accurately record them. Researchers have explored the use of neural networks, particularly language models, for automated ICD-9 code assignment. However, the imbalanced distribution of ICD-9 codes leads to poor performance. One solution is to use domain knowledge to incorporate a useful prior. This paper evaluates the usefulness of the correlation bias: we hypothesize that correlations between ICD-9 codes and other medical codes could help improve language models' performance. We showed that while the correlation bias worsens the overall performance, the effect on individual class can be negative or positive.[1] Performance on classes that are more imbalanced and less correlated with other codes is more sensitive to incorporating the correlation bias. This suggests that while the correlation bias has potential to improve ICD-9 code assignment in certain cases, the applicability criteria need to be more carefully studied.

1 Introduction

Electronic Health Records (EHRs) contain patient information in the form of clinical notes, structured data tables, and biomedical imaging and time series. For easy tracking and analysis of health data across different healthcare systems, and critically for billing purposes, hospitals and insurance companies assign codes of a standardized coding system to characterize the clinical conditions of patients. Wrong code assignments may result in billing issues that increase patients' expenses substantially, misdiagnosis, and poor tracking of population level health conditions nationally. The Ninth Revision of the International Classification of Diseases (ICD-9) is a system used worldwide to classify and code diseases, injuries, and other health conditions. There were extensive efforts studying the automated assignment of ICD-9 codes to health records and relevant documents (Yan et al., 2022).

With recent developments in NLP, there has been a focus on the use of neural networks (Yu et al., 2019; Mullenbach et al., 2018; Teng et al., 2020). One particularly recent direction is in the use of language models. Originally introduced in BERT (Devlin et al., 2019), the recipe of pretraining and finetuning of language models has shown promising performance in many tasks. Researchers have applied BERT for assigning ICD-9 codes from medical documents (Huang et al., 2022; Pascual et al., 2021; Zhang et al., 2020). However, BERT and other encoder-based language models perform poorly on ICD-9 code assignment (Yan et al., 2022).

One challenge is the extremely imbalanced distribution of ICD-9 codes. Following the distribution of medical conditions in the real world, some codes occur frequently while other codes may appear only once (Yan et al., 2022). It is difficult for models

[1] The implementation code is available on github: `https://github.com/nyuolab/text2table`

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics - Student Research Workshop, pages 109–118
July 10-12, 2023 ©2023 Association for Computational Linguistics

to correctly predict minority codes because few samples exist in the dataset (Sun et al., 2009). A proposed solution is to incorporate domain knowledge that provides useful priors for the minority codes (Bai and Vucetic, 2019; Wang et al., 2020; Zeng et al., 2019).

We hypothesize that one useful prior for ICD-9 code assignment is the correlation between ICD-9 codes and other relevant coding systems. We term other relevant coding systems auxiliary tasks because language models in our experiments predict codes from these systems in addition to ICD-9 codes. The auxiliary tasks are Current Procedural Terminology (CPT) codes and Diagnosis-Related Group (DRG) codes. This correlation prior stems from the domain knowledge that labels from other coding systems give information about ICD-9 codes. For example, patients who underwent artery bypass surgeries (CPT code 33533) are likely to have heart failures (ICD-9 code 428.0). To test our hypothesis, we investigate the effect of multitasking on correlated auxiliary tasks and encouraging similar label correlations between training labels and model predictions through regularization. We showed that 1) on average, utilizing correlations hurts language models' performance on predicting ICD-9 codes from discharge summaries, 2) for each ICD-9 code, utilizing correlations might hurt or help, 3) ICD-9 codes that are more imbalanced and less correlated with auxiliary tasks experience larger performance changes (both positive and negative) from incorporating the correlation prior. Our findings suggest that the correlation prior has the potential to improve predictions of certain ICD-9 codes, but this method suffers from instability when the main task has an imbalanced label distribution and a weak correlation with auxiliary tasks.

2 Related Work

Domain knowledge One useful prior for ICD-9 codes is its hierarchical structure. For example, a high-level code (e.g., 428.0 heart failure) encompasses its corresponding low-level codes (e.g., 428.1 left heart failure, 428.2 systolic heart failure). Tsai et al. (2019) incorporated this hierarchical prior and improved models' performance on predicting imbalanced ICD-9 codes.

CorrLoss CorrLoss is a regularization technique (Rieger et al., 2022) that encourages consistent label correlations between ground truth and predictions. Rieger et al. (2022) uses CorrLoss on the facial affect recognition task to integrate the correlation priors for facial movements. Corrloss can be used in any domain where correlation between prediction targets provides a useful signal. Thus, we adopt Corrloss to integrate information of the correlations between different kinds of diagnosis and procedure codes.

3 Methods

Task overview We formulate the task of code assignment into a multilabel text classification task because each patient has multiple codes corresponding to their discharge summaries. Each binary label in the task corresponds to a specific code. Formally, our classifier aims to approximate the probability $p(y_1, \ldots, y_n|x)$, where each y_i is an ICD-9 code and x is a discharge summary.

The Correlation Prior We hypothesize that correlations between ICD-9 and other coding systems are a useful prior for ICD-9 code assignment and choose to incorporate the prior in two ways.

First, we added the auxiliary tasks of predicting other medical codes (e.g., CPT). Formally, we train a classifier to approximate

$$p(y, z|x) = p(y|x)\, p(z|x, y), \tag{1}$$

where y is a sequence of ICD-9 codes (the main task), z is a sequence of other medical codes (the auxiliary task), and x is a discharge summary. Our domain knowledge assumes that the absolute correlation $\text{abs}(\rho(y, z)|x) > 0$, so y, z are not conditionally independent given x and $p(z|x, y) \neq p(z|x)$. This is desirable because otherwise, we are strictly increasing the difficulty of the task from learning $p(y|x)$ to learning $p(y|x)\, p(z|x)$.

There are benefits and concerns associated with Equation 1, and their trade-off is unclear *a priori*. One benefit is that extra dependency information from $p(z|x, y)$ could potentially simplify learning $p(y, z|x)$. One drawback is that the additional prediction targets z could worsen the curse of dimensionality. Whether the benefit would outweigh the drawback is difficult to determine without running a controlled experiment.

Second, we used CorrLoss to encourage similar label correlation patterns between training and predictions. Formally, we added a regularization term $c = \sum_{i \neq j} c(d_i, d_j)$. Each summation term scales with a correlation difference:

$$c(d_i, d_j) \propto |\rho(d_i, d_j)_{y_{\text{train}}} - \rho(d_i, d_j)_{\hat{y}}|, \tag{2}$$

		PROC	PROC+CPT	PROC+DRG	PROC+DIAG
ClinicalBERT	original	**0.4528**	0.397	0.3939	0.408
	CorrLoss	0.4037	0.3594	0.3272	0.363
RoBERTa	original	**0.4421**	0.4009	0.3884	0.4116
	CorrLoss	0.3736	0.3236	0.2816	0.3692
Longformer	original	**0.4712**	0.4227	0.3886	0.4219
	CorrLoss	0.4139	0.335	0.212	0.3549

Table 1: Macro F1 scores of experiments, in which procedure ICD-9 is the main task, on MIMIC-III-50 test set. For each model, the best F1 score is in bold. PROC means procedure ICD-9. DIAG means diagnosis ICD-9. PROC+CPT means that procedure ICD-9 is the main task and CPT is the auxiliary task.

where d_i, d_j are different classes, $\rho(d_i, d_j)_v$ is the correlation between class d_i and d_j in a vector v, y_{train} is the training labels, $\hat{y}$ is the predicted labels, and ρ is the Pearson correlation function.

Dataset We built two datasets from the Medical Information Mart for Intensive Care III (MIMIC-III) (Johnson et al., 2016), a database of EHRs. Our first dataset, subsequently referred to as “MIMIC-III", contains examples of each patient's discharge summary, and associated diagnosis and procedure codes (diagnosis ICD-9, procedure ICD-9, CPT, and DRG). Because this dataset is extremely imbalanced, we further select the top 50 most frequently used codes for each kind of coding system to construct a second dataset that represents a more ideal scenario. Following the convention of related literature, we call this dataset “MIMIC-III-50" (Vu et al., 2020; Luo et al., 2021; Li and Yu, 2020). Statistics of the MIMIC-III dataset are in Appendix A.

Models and Evaluation We use ClinicalBERT (Alsentzer et al., 2019), RoBERTa (Liu et al., 2019), Longformer (Beltagy et al., 2020) (justification in Appendix C). We use the macro F1 as our metric for comparison because this metric treats all classes equally, which means minority codes are as important as majority codes in evaluation (Branco et al., 2016; Sun et al., 2009; Ferri et al., 2009). Because it is an imbalanced classification, the default threshold of 0.5 is not suitable (Zhou and Liu, 2006; Zou et al., 2016). Instead, we tune the threshold according to the precision-recall curve to maximize the F1 score for each individual label.

4 Experiments

To test whether the correlation prior is useful for ICD code assignment, we incorporate multitasking (Equation 1) and CorrLoss (Equation 2) into our model and check if they improve performance. Specifically, we studied two main tasks (diagnosis ICD-9 codes and procedure ICD-9 codes). For each main task, we added one of the three auxiliary tasks: DRG codes, CPT codes, and the other ICD-9 codes (for diagnosis ICD-9 code, the auxiliary task can be procedure ICD-9 code, and vice versa). We trained both main-task-only models and multitasking models with and without CorrLoss.

5 Results

Multitasking and CorrLoss hurt performance on MIMIC-III-50 and do not significantly impact performance on MIMIC-III. Table 1 shows the macro-F1 score on procedure ICD-9 of the MIMIC-III-50 dataset. We observe two patterns for each language model. First, adding auxiliary tasks always decreases the performance of models in comparison to predicting main tasks only. Second, regularizing with CorrLoss always decreases the performance of models in comparison to not using CorrLoss. The same pattern exists for predicting diagnosis ICD-9 of the MIMIC-III-50 dataset (Appendix Table 6). However, on the full MIMIC-III dataset, multitasking and CorrLoss do not impact models' performance significantly (Appendix B).

6 Analysis

Since the macro F1 score does not show significant changes from multitasking and CorrLoss on the full MIMIC-III dataset, we investigate whether the performance changes for individual labels. Specifically, we analyzed how label imbalance (measured by Shannon entropy, defined in Appendix D.1) and label correlation (measured by the average absolute Pearson correlation coefficient between each main task label and all auxiliary task labels, as defined in Appendix D.1) affect the model's performance.

For individual ICD-9 code, incorporating the correlation prior may hurt or help. Figure 1 shows that there exist labels with both negative and positive performance changes.

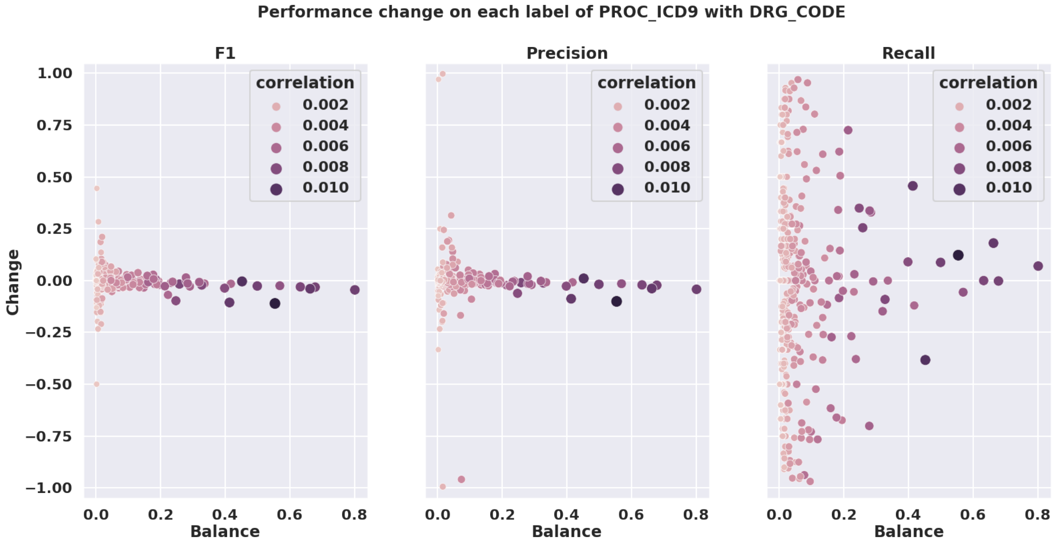

Figure 1: The plot of ClinicalBERT's performance changes (Y axis) on labels of procedure ICD-9, when DRG is added as the auxiliary task, versus the balances (X axis) of the labels, and versus the correlations (sizes and colors of the units) between each label with the whole auxiliary DRG task. CorrLoss is not included.

Labels that are more imbalanced and less correlated to auxiliary labels experience larger changes. Figure 1 shows two relationships: (1) more balanced labels (closer to the right) have less performance changes (spread of dots on the y axis), (2) labels that are more correlated with the auxiliary task (darker dots) have less performance changes (spread along the y axis). All the other plots of different tasks and setups show similar patterns (Appendix D.1).

		top50	bottom50
ClinicalBERT	+CPT	0.333	0.273
	+DRG	0.28	0.413
	+DIAG	0.3	0.387
RoBERTa	+CPT	0.4	0.3
	+DRG	0.393	0.353
	+DIAG	0.313	0.287
Longformer	+CPT	0.34	0.427
	+DRG	0.34	0.28
	+DIAG	0.347	0.307

Table 2: The percentages of positive macro F1 score changes on the top 50 most balanced procedure ICD-9 labels and on the bottom 50 least balanced procedure ICD-9 labels, with different auxiliary tasks and models. CorrLoss is not included.

In both extreme scenarios (imbalanced label, small correlation with auxiliary labels) and ideal scenarios (balanced labels, high correlation with auxiliary labels), **incorporating correlation is more likely to hurt than help.** Table 2 shows that for the top 50 most balanced labels and the bottom 50 least balanced labels, if we utilize correlations (with multitasking and CorrLoss), the percentage of positive F1 score changes is always less than 50%. Table 3 shows that for the top 50 labels that are most correlated with the auxiliary tasks and the bottom 50 labels that are least correlated with the auxiliary tasks, utilizing correlations also leads to $< 50\%$ positive F1 score change.

		top50	bottom50
ClinicalBERT	+CPT	0.333	0.327
	+DRG	0.32	0.327
	+DIAG	0.293	0.247
RoBERTa	+CPT	0.487	0.333
	+DRG	0.373	0.387
	+DIAG	0.267	0.293
Longformer	+CPT	0.433	0.327
	+DRG	0.28	0.273
	+DIAG	0.333	0.24

Table 3: The percentages of positive macro F1 score changes on the top 50 procedure ICD-9 labels that are most correlated with the auxiliary task and on the bottom 50 procedure ICD-9 labels that are least correlated with the auxiliary task, with different auxiliary tasks and models. CorrLoss is included.

7 Discussion

Since multitasking and CorrLoss worsen language models' overall performance, it contradicts our hypothesis that the correlations between ICD-9 codes and other medical codes would be a useful prior. Nevertheless, the performance changes on individual labels are more nuanced and show potential for improving prediction of certain ICD-9 codes. We wonder what characterizes the labels that benefit from incorporating the correlation prior (dots with positive changes in Figure 1). Perhaps for those labels, the additional dependency information gained from the auxiliary tasks outweigh the increased learning complexity from a larger output space. A prerequisite for a rigorous investigation would be quantifying the trade-off between the dependency information and the learning complexity.

We recognize three limitations that may influence the interpretation of our results and call for future works. First, we did not conduct a hyperparameter search for the regularization strength of CorrLoss. Second, since F1 score decreases are substantial and universal across all experiments on MIMIC-III-50, we did not run experiments multiple times with different seeds. Third, we did not provide a rigorous explanation of what caused our empirical findings. Future works can investigate the plausible hypothesis that the trade-off between the dependency information and the learning complexity causes these findings. Besides these limitations, future works can also investigate more scenarios and methods of incorporating the correlation prior.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tian Bai and Slobodan Vucetic. 2019. Improving Medical Code Prediction from Clinical Text via Incorporating Online Knowledge Sources. In *The World Wide Web Conference*, WWW '19, pages 72–82, New York, NY, USA. Association for Computing Machinery.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. ArXiv:2004.05150 [cs].

Paula Branco, Luís Torgo, and Rita P. Ribeiro. 2016. A Survey of Predictive Modeling on Imbalanced Domains. *ACM Computing Surveys*, 49(2):31:1–31:50.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

C. Ferri, J. Hernández-Orallo, and R. Modroiu. 2009. An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. PLM-ICD: Automatic ICD Coding with Pretrained Language Models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 10–20, Seattle, WA. Association for Computational Linguistics.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei 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):160035. Number: 1 Publisher: Nature Publishing Group.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Fei Li and Hong Yu. 2020. ICD Coding from Clinical Text Using Multi-Filter Residual Convolutional Neural Network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8180–8187. Number: 05.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Junyu Luo, Cao Xiao, Lucas Glass, Jimeng Sun, and Fenglong Ma. 2021. Fusion: Towards Automated ICD Coding via Feature Compression. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2096–2101, Online. Association for Computational Linguistics.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Damian Pascual, Sandro Luck, and Roger Wattenhofer. 2021. Towards BERT-based Automatic ICD Coding: Limitations and Opportunities. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 54–63, Online. Association for Computational Linguistics.

Ines Rieger, Jaspar Pahl, Bettina Finzel, and Ute Schmid. 2022. CorrLoss: Integrating Co-Occurrence Domain Knowledge for Affect Recognition. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 798–804. ISSN: 2831-7475.

Yanmin Sun, Andrew K. C. Wong, and Mohamed S. Kamel. 2009. Classification of imbalanced data: a review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04):687–719. Publisher: World Scientific Publishing Co.

Fei Teng, Wei Yang, Li Chen, LuFei Huang, and Qiang Xu. 2020. Explainable Prediction of Medical Codes With Knowledge Graphs. *Frontiers in Bioengineering and Biotechnology*, 8.

Shang-Chi Tsai, Ting-Yun Chang, and Yun-Nung Chen. 2019. Leveraging Hierarchical Category Knowledge for Data-Imbalanced Multi-Label Diagnostic Text Understanding. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 39–43, Hong Kong. Association for Computational Linguistics.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A Label Attention Model for ICD Coding from Clinical Text. volume 4, pages 3335–3341. ISSN: 1045-0823.

Ke Wang, Xuyan Chen, Ning Chen, and Ting Chen. 2020. Automatic Emergency Diagnosis with Knowledge-Based Tree Decoding. volume 4, pages 3407–3414. ISSN: 1045-0823.

Chenwei Yan, Xiangling Fu, Xien Liu, Yuanqiu Zhang, Yue Gao, Ji Wu, and Qiang Li. 2022. A survey of automated International Classification of Diseases coding: development, challenges, and applications. *Intelligent Medicine*, 2(3):161–173.

Ying Yu, Min Li, Liangliang Liu, Zhihui Fei, Fang-Xiang Wu, and Jianxin Wang. 2019. Automatic ICD code assignment of Chinese clinical notes based on multilayer attention BiRNN. *Journal of Biomedical Informatics*, 91:103114.

Min Zeng, Min Li, Zhihui Fei, Ying Yu, Yi Pan, and Jianxin Wang. 2019. Automatic ICD-9 coding via deep transfer learning. *Neurocomputing*, 324:43–50.

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, Online. Association for Computational Linguistics.

Zhi-Hua Zhou and Xu-Ying Liu. 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Quan Zou, Sifa Xie, Ziyu Lin, Meihong Wu, and Ying Ju. 2016. Finding the Best Classification Threshold in Imbalanced Classification. *Big Data Research*, 5:2–8.

A Dataset Statistics

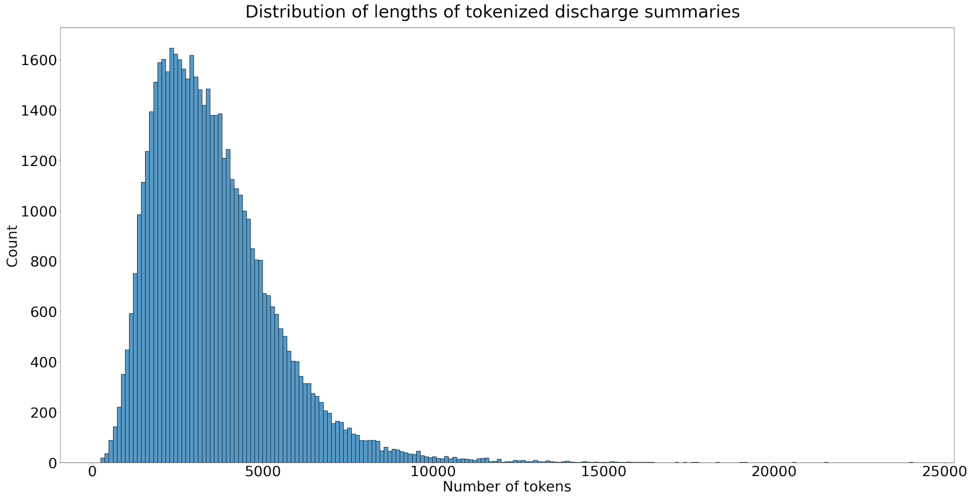

Figure 2: The distribution of lengths of tokenized discharge summaries in MIMIC-III dataset.

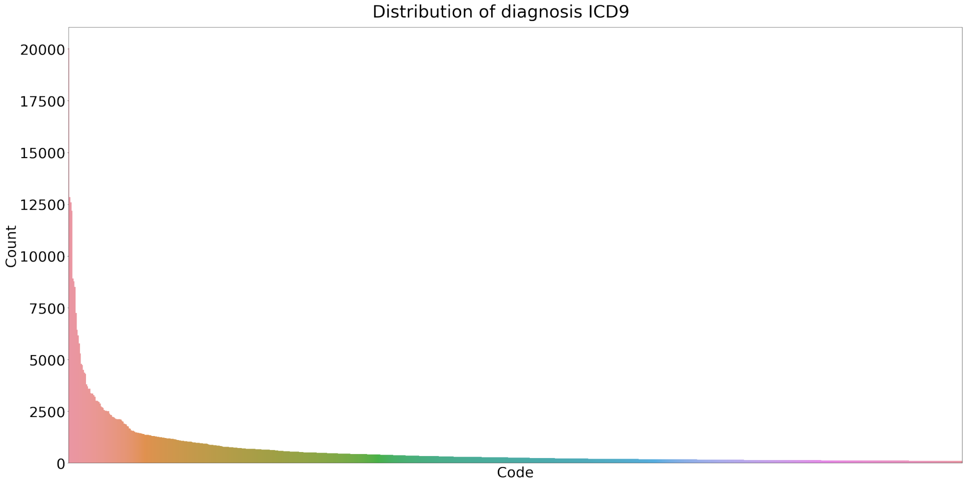

Figure 3: The distribution of diagnosis ICD-9. There are 6918 diagnosis ICD-9 codes. 6062 Codes occur less than or equal to 100 times in MIMIC-III dataset. They are not included for clarity.

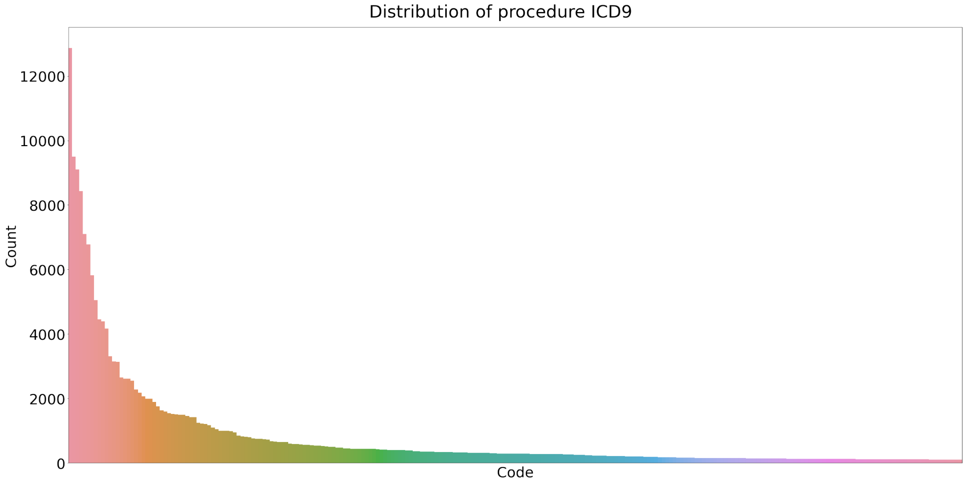

Figure 4: The distribution of procedure ICD-9. There are 2011 procedure ICD-9 codes. 1767 Codes occur less than or equal to 100 times in MIMIC-III dataset. They are not included for clarity.

B Results

		PROC	PROC+CPT	PROC+DRG	PROC+DIAG
ClinicalBERT	original	0.0098	0.0094	0.0091	0.0097
	CorrLoss	0.0102	0.0099	0.0088	0.0087
RoBERTa	original	0.0097	0.0089	0.0087	0.0088
	CorrLoss	0.0095	0.0095	0.0098	0.0089
Longformer	original	0.0088	0.0088	0.0095	0.0085
	CorrLoss	0.0094	0.0085	0.0091	0.0078

Table 4: Macro F1 scores of experiments, in which procedure ICD-9 is the main task, on full MIMIC-III test set.

		DIAG	DIAG+CPT	DIAG+DRG	DIAG+PROC
ClinicalBERT	original	0.0068	0.0066	0.0066	0.0067
	CorrLoss	0.0066	0.0069	0.0069	0.0068
RoBERTa	original	0.0069	0.0065	0.0062	0.0065
	CorrLoss	0.0071	0.0071	0.0066	0.0065
Longformer	original	0.0072	0.0069	0.007	0.0071
	CorrLoss	0.007	0.0068	0.0076	0.0071

Table 5: Macro F1 scores of experiments, in which diagnosis ICD-9 is the main task, on full MIMIC-III test set.

		DIAG	DIAG+CPT	DIAG+DRG	DIAG+PROC
ClinicalBERT	original	0.3755	0.3296	0.3351	0.3351
	CorrLoss	0.3235	0.2966	0.2947	0.2992
RoBERTa	original	0.3851	0.3255	0.3307	0.3341
	CorrLoss	0.3143	0.2822	0.2713	0.2939
Longformer	original	0.4408	0.349	0.3544	0.3552
	CorrLoss	0.3364	0.2963	0.2906	0.3027

Table 6: Macro F1 scores of experiments, in which diagnosis ICD-9 is the main task, on MIMIC-III-50 test set.

C Justifcation of Models

The variant of ClinicalBERT we use is Bio+Discharge Summary BERT model because it was further trained on discharge summaries from MIMIC-III after initialized from BioBERT (Lee et al., 2020).

We use RoBERTa because it is a variant of vanilla BERT that was trained differently to improve its performance on a range of NLP tasks.

We use Longformer because it can handle long text sequences. BERT and many BERT-based models cannnot handle text sequences longer than 512 tokens. Many tokenized discharge summaries are text sequences longer than 512 tokens and Longformer can benefit from more complete understandings of discharge summaries.

Each model represents a different improvement on top of vanilla BERT: ClinicalBERT improves through domain-specific pretraining; RoBERTa improves through tuning training setup; and Longformer improves through incorporating more information from the input. With these models, we cover a significant part of the improvement spectrum, which shows that the pattern we present is generalizable to different models.

D Analysis

D.1 Performance on Each Label

Other figures Since there are 72 experiments that have auxiliary tasks, there are 72 corresponding plots. Thus, it is unreasonable to include all of them in the appendix. You can find all plots in our github repository: https://github.com/nyuolab/text2table/tree/main/notebooks.

Shannon Entropy

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \quad (3)$$

In this equation, $H(X)$ represents the entropy of a label X with possible outcomes x_1, x_2, ..., x_n. In our context, $n = 2$ because a label only has two possible outcomes: 1 (positive) or 0 (negative). The term $p(x_i)$ represents the probability of the i-th outcome, and the logarithm is taken with base 2 to give the result in units of bits. The sum is taken over all possible outcomes of X. With only two possible outcomes, a label's Shannon entropy will be close to 1 if it is balanced, and will be close to 0 if it is imblanced.

Representation of Correlations

$$C(a, B) = \frac{\sum_{b \in B} |P(a, b)|}{card(B)} \quad (4)$$

In this equation, $C(a, B)$ represents the correlations between a label of the main task a and a set containing labels of the auxiliary task. For each label of the auxiliary task $b \in B$, $|P(a, b)|$ represents the absolute value of the Pearson correlation coefficient bettwen a and b. $card(B)$ is the cardinality of B (i.e. the number of labels in B).

D.2 Performance in Different Scenarios

		top50	bottom50
ClinicalBERT	+CPT	0.453	0.32
	+DRG	0.54	0.293
	+PROC	0.48	0.38
RoBERTa	+CPT	0.48	0.313
	+DRG	0.507	0.307
	+PROC	0.48	0.333
Longformer	+CPT	0.5	0.32
	+DRG	0.48	0.393
	+PROC	0.433	0.287

Table 7: The percentages of positive macro F1 score changes on the top 50 most balanced diagnosis ICD-9 labels and on the bottom 50 least balanced diagnosis ICD-9 labels, with different auxiliary tasks and models. CorrLoss is not included in all experiments we examine in this table.

		top50	bottom50
ClinicalBERT	+CPT	0.347	0.36
	+DRG	0.327	0.313
	+DIAG	0.273	0.28
RoBERTa	+CPT	0.32	0.32
	+DRG	0.353	0.36
	+DIAG	0.273	0.22
Longformer	+CPT	0.353	0.367
	+DRG	0.28	0.293
	+DIAG	0.307	0.26

Table 8: The percentages of positive macro F1 score changes on the top 50 most balanced procedure ICD-9 labels and on the bottom 50 least balanced procedure ICD-9 labels, with different auxiliary tasks and models. CorrLoss is included in all experiments we examine in this table.

		top50	bottom50
ClinicalBERT	+CPT	0.413	0.307
	+DRG	0.533	0.28
	+PROC	0.487	0.293
RoBERTa	+CPT	0.46	0.3
	+DRG	0.493	0.373
	+PROC	0.473	0.34
Longformer	+CPT	0.453	0.293
	+DRG	0.487	0.34
	+PROC	0.5	0.307

Table 9: The percentages of positive macro F1 score changes on the top 50 most balanced diagnosis ICD-9 labels and on the bottom 50 least balanced diagnosis ICD-9 labels, with different auxiliary tasks and models. CorrLoss is included in all experiments we examine in this table.

		top50	bottom50
ClinicalBERT	+CPT	0.467	0.32
	+DRG	0.307	0.373
	+DIAG	0.367	0.287
RoBERTa	+CPT	0.387	0.267
	+DRG	0.413	0.407
	+DIAG	0.32	0.307
Longformer	+CPT	0.427	0.367
	+DRG	0.34	0.307
	+DIAG	0.42	0.307

Table 10: The percentages of positive macro F1 score changes on the top 50 procedure ICD-9 labels that are most correlated with the auxiliary task and on the bottom 50 procedure ICD-9 labels that are least correlated with the auxiliary task, with different auxiliary tasks and models. CorrLoss is not included in all experiments we examine in this table.

		top50	bottom50
ClinicalBERT	+CPT	0.507	0.333
	+DRG	0.493	0.287
	+PROC	0.473	0.347
RoBERTa	+CPT	0.48	0.247
	+DRG	0.513	0.36
	+PROC	0.46	0.347
Longformer	+CPT	0.487	0.313
	+DRG	0.493	0.34
	+PROC	0.427	0.313

Table 11: The percentages of positive macro F1 score changes on the top 50 diagnosis ICD-9 labels that are most correlated with the auxiliary task and on the bottom 50 diagnosis ICD-9 labels that are least correlated with the auxiliary task, with different auxiliary tasks and models. CorrLoss is not included in all experiments we examine in this table.

		top50	bottom50
ClinicalBERT	+CPT	0.467	0.373
	+DRG	0.52	0.3
	+PROC	0.46	0.333
RoBERTa	+CPT	0.493	0.32
	+DRG	0.52	0.433
	+PROC	0.473	0.253
Longformer	+CPT	0.46	0.32
	+DRG	0.513	0.467
	+PROC	0.453	0.34

Table 12: The percentages of positive macro F1 score changes on the top 50 diagnosis ICD-9 labels that are most correlated with the auxiliary task and on the bottom 50 diagnosis ICD-9 labels that are least correlated with the auxiliary task, with different auxiliary tasks and models. CorrLoss is included in all experiments we examine in this table.