TraumaICD-BERT: Automating ICD10 Diagnosis Code Identification From Electronic Medical Records of Injured Patients

Stanford CS224N Custom Project

Yifu Chen
Department of Biomedical Data Science
Stanford University
yifuchen@stanford.edu

Jeff Choi MD MSc
Department of Surgery
Stanford University
jc2226@stanford.edu

Alexander Sivura
Center for Professional Development
Stanford University
asivura@stanford.edu

Abstract

Despite surging interest among trauma surgeons, machine learning-based outcome prediction tools are not used at the bedside. Most clinical outcome prediction tools require ICD10 diagnosis codes as input variables, but currently, ICD10 diagnosis codes are manually extracted weeks after a patient leaves the hospital. To meet the critical need to make ICD10 diagnosis codes available real-time, we aimed to build an NLP model that can automatically extract ICD10 diagnosis codes from unstructured free text. Our dataset comprised unstructured trauma survey notes from 3478 trauma patients treated at Stanford Hospital between 2016 and 2021. Baseline performance using Amazon Web Service Comprehend Medical (AWSCM) yielded accuracy of 0.936 and Micro-AUC of 0.760. By fine-tuning a customized deep learning biomedical language model, PubMedBERT, we achieved test set accuracy of 0.958 and Micro-AUC of 0.895. Our model also outperformed AWSCM in terms of Macro-AUC, Micro-F1, Macro-F1, Precision, and Recall scores. To out knowledge, our study is first to fine-tune and validate a deep learning language model to extract ICD10 diagnosis codes specifically for use among trauma patients. Our study explores practical ways to navigate challenge of real-world medical data mining, such as the absence of publicly-available big data, imbalanced datasets, the need for high-dimensional classification, and considering the flexibility and interpretability trade-off in model design. We believe this first step towards reliable, automated injury ICD10 diagnosis code extraction could connect the critical missing link for many prediction tools to reach the patient bedside.

1 Key Information to include

- TA mentor: Gaurab Banerjee
- External collaborators (if no, indicate "No"): No
- External mentor (if no, indicate "No"): David A. Spain, MD
- Sharing project (if no, indicate "No"): No

2 Introduction

Automating ICD10 diagnosis code extraction using NLP could meet a critical missing link for many machine learning-based prediction tools to reach the trauma patient bedside. Several challenges hinder NLP algorithms development and finetuning for bedside adoption. First, there are few publicly-available, large electronic medical record (EMR) datasets. MIMIC-III has been the training source for many medical NLP and other machine learning algorithms,[1, 2] yet comprises a distinct population: critically ill patients who were seen at a single tertiary hospital. To our knowledge, there is no publicly available dataset containing EMR text of trauma patients. Second, an individual patient can suffer multiple injuries (thus, have multiple injury ICD10 diagnosis codes) per hospitalization. ICD10 diagnoses constitute a 7-character (e.g. S04.34XAA) ontology, wherein each sequential character details increasingly-specific diagnoses (e.g. S2: injury to chest, S22: fracture of chest bone, S22.3: fracture of rib). If we aimed to predict 7-character ICD10 diagnosis codes, there would be over 10,000 injury diagnoses to predict. Unfortunately, no single hospital (or a group of hospitals that could feasibly pass a multi-institutional IRB agreement together) has large enough a trauma volume to ensure multiple instances of all 10,000 injury ICD10 diagnoses would be captured. Third, many injury ICD10 diagnosis codes are rare (e.g. code describing complete aortic dissection), while some are frequent (e.g. code describing rib fractures); ICD10 datasets are heavily-imbalanced. Last, a vast majority of surgeons do not have machine learning or data science training and have a general distrust towards "black box algorithms" that do not facilitate visual inference.

To address these challenges, we aimed to build a database of trauma patients' unstructured EMR notes. By considering how our algorithm would be implemented at the bedside, we navigate the challenge of addressing a high dimensional-classification problem using a limited size dataset through simple, practical solutions (e.g. only considering the first 4 characters of ICD10 codes to reduce the outcome dimension, as 4-character ICD10 codes would be specific enough to inform surgeon decision-making). We finetune a variant of the Bidirectional Encoder Representations from Transformers (BERT) model [3] that has been pretrained on medical text (PubMed) through a comprehensive hyperparameter-space optimization, and present attention visualizations to maximize model understanding for the surgeon audience. We developed our model with implementation and end-user in mind at the outset, to facilitate our model's adoption at the bedside and impact meaningful clinical change.

3 Related Work

Several studies have explored automating ICD10 (and the previous medical diagnosis ontology, ICD9) extraction from electronic medical records (Supplemental Table 1). [1, 2, 4, 5, 6] The input text for previous studies comprised all unstructured EMR notes written throughout a patient's hospitalization or discharge summaries (notes written on the last day of hospitalization). Models developed using such input data are not applicable for trauma patients, as critical clinical decisions usually need to be made within the first day(s) of hospitalization. Moreover, injury ICD10 diagnosis codes have not been adequately represented within previous studies; One study reported that ICD10 code classification performance was lowest among trauma patients. [6] All studies have noted the challenge of making multi-class predictions on heavily imbalanced ICD datasets; some have resorted to evaluating performance for only the top 50 or 100 ICD diagnosis codes. [2, 5, 1]

The lack of existing injury ICD10-extraction models may be attributed to the scarcity of healthcare NLP datasets, due to privacy concerns over sharing patient data. As result, previusly published state-of-the-art models – PubMedBERT, BioBERT, ClinicalBERT, SciBERT, and BlueBERT – were pre-trained on open-access data, for example, PubMed articles and MIMIC III dataset. [7, 8, 9]

Tertiary Note

1. Acute compression fracture at T7. Likely superimposed T6-T7 and T7-T8 disc injury.
2. B/l sacral fractures and right L3 and L4 transverse process fracture is better appreciated on comparison CT.
3. Grade 1 liver laceration with subcapsular hematoma.
4. Right scapular fracture

MIMIC III Discharge Summary

He does have a fracture of the T7-T8 disk. He was seen by Dr. [**Last Name (STitle) 1907**]. He had an echocardiogram on [**2211-2-7**] that showed an EF of 50 percent with right ventricular hypertrophy and mild A-V sclerosis.
He was also seen by the Pulmonary service for his snoring and his sleep apnea for which he is receiving BiPAP. The patient was fitted for a TLSO brace and had Physical Therapy.

PubMed Article

Background: There are multiple surgical treatment options for traumatic thoracic spine spondyloptosis.
Case description: A 20-year-old male presented with back deformity attributed to a fall...
Conclusion: Here we present a patient who following a fall sustained a T7/T8 spondyloptosis resulting in paraplegia treated with spondylectomy.

Figure 1: Examples of tertiary note, discharge summary, and PubMed article. All three corpus describe patients with T7-T8 vertebrae spinal cord injuries, however, texts display little similarity in terms of vocabularies and grammars.

However, there is little overlap between the vocabularies of injury tertiary notes and hospitalization notes. For example, tertiary notes describe injuries using fragmented sentences, while discharge summary describe treatments using complete sentences. Figure 1 compares injury note, MIMIC III discharge summary, and PubMed article.

Due to this lack of existing models and datasets, a custom dataset from Stanford Hospital was collected to fine-tune a pre-trained biomedical NLP model. Since pre-trained models output token embeddings rather than binary probabilities, we engineered an architecture inspired by the TransICD model [10], by concatenating the PubMedBERT Transformer encoder with fully-connected layers to produce binary classification probabilities.

4 Approach

4.1 Model Architecture

The custom NLP model, TraumaICD-BERT, was based on a pre-trained PubMedBERT model [7], a variant of BERT [3] trained using biomedical text from research articles. Figure 2 illustrates the architecture. We chose a BERT-based model as a previous study evaluating different NLP algorithms reported higher performance using BERT than RNNs. [4] Moreover, in the BLURB leaderboard (Biomedical Language Understanding and Reasoning Benchmark [7]), PubMedBERT outperformed all other biomedical BERT models, including ClinicalBERT [9], BioBERT [8], SciBERT [11], BlueBERT [12] as well as domain-general variants (BERT [3] and RoBERTa [13]). We chose PubMedBERT because it uses a custom biomedical vocabulary with pre-training-from-scratch on biomedical text. Our dataset size is large enough for training this model into an ICD10 code extractor, which we describe in the following sections.

Other approaches have been considered. One involves using UnifiedQA-T5 [14] – a version of T5 model (Text-to-Text Transfer Transformer [15]) fine-tuned on a variety of question-answering tasks (e.g., SQuAD [16]). Using T5 would be approaching the ICD10 code extraction as a sequence of question-answer or multiple-choice inferences in a decision tree manner, since each addition digit of the ICD10 code becomes more specific to the injury. The decision tree approach would allow exponentially higher compute efficiency in classifying all 10,000 7-character IC10 codes, since most branches would not be explored. However, in practical terms, only 4-character codes would be satisfactory. Sine there are only a few hundred number 4-character codes, which BERT may classify in reasonable time, we decided to forgo the decision tree approach to avoid compounding

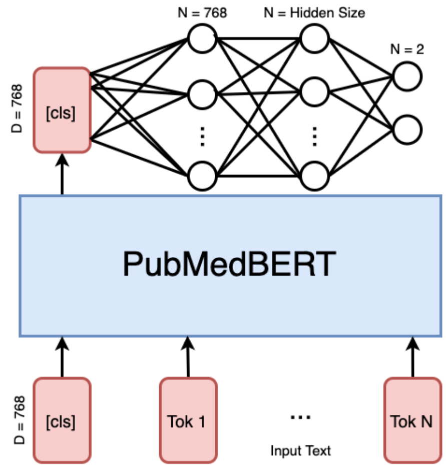


Figure 2: Illustration of the custom classification model. A pre-trained PubMedBERT was connected to a fully-connected feed-forward network (classification head) in order to convert the original PubMedBERT output into a binary probability. The classification head's hidden size is an adjustable hyper-parameter. All parameters were updated via back-propagation.

errors of non-biomedical T5 models. Another approach has been considered as well: using Amazon Web Service Comprehend Medical (AWSCM) cloud service to extract ICD10 codes. After initial experimentations, it was found that AWSCM produces a high rate of false positives. A potential solution would involve fine-tuning a BERT model to post-process the AWSCM extractions to filter out the false positive codes. A disadvantage of this approach is that the model can extract only a subset of AWSCM extractions, so our recall rate would be limited by the AWSCM recall rate. Due to the desire for high recall rates (low false negatives), we decided to fine-tune our own custom PubMedBERT model.

4.2 Data Generation & Augmentation Algorithm

To build a dataset for ICD10 diagnosis codes for real-time clinical decision making in trauma patients, we obtained trauma tertiary notes from all trauma patients who have been admitted to Stanford Hospital between January 2016 and June 2021 (2016 was the first year of ICD10 ontology implementation). Tertiary notes are written within 24 hours of hospitalization for all trauma patients nationally. Thus, to extract ICD10 codes for input into downstream prediction tools that impact real-time clinical decision-making, model development using tertiary notes is ideal.

A custom data augmentation algorithm generate training examples from each tertiary note. The algorithm is represented by the pseudo-code below, where we define the following variables: tertiary note T_i of a single patient P_i, who has a total number of N_i injuries, that has been coded as $C_{i1}^{(7)}, ...C_{iN}^{(7)}$ 7-digit injury ICD10 codes by expert humans. We truncate the injury codes to keep the first 3 or 4 characters only, resulting in $C_{i1}^{(3)}, ...C_{iN}^{(3)}$ and $C_{i1}^{(3)}, ...C_{iN}^{(3)}$ respectively. We randomly select one negative ICD10 code for each positive code as adversarial examples, $\bar{C}_{i1}^{(3)}, ...\bar{C}_{iN}^{(3)}, \bar{C}_{i1}^{(4)}, ...\bar{C}_{iN}^{(4)}$. Each positive and negative code is paired with the ICD10 injury description $D_{C_{i\cdot}^{(\cdot)}}$ that is associated with the code. The pseudocode algorithm is available in the Appendix Algorithm 1. Two concrete examples are shown in Figure 3.

We focused on predicting only the first 4 characters of injury ICD10 diagnosis codes. Trauma prediction tools only require inputting 4-character ICD10 diagnosis codes, as further detailed injury description is unnecessary for real-time clinical use. Limiting ICD10 diagnosis codes to the first 4 characters yields only 729 ICD10 codes to predict.

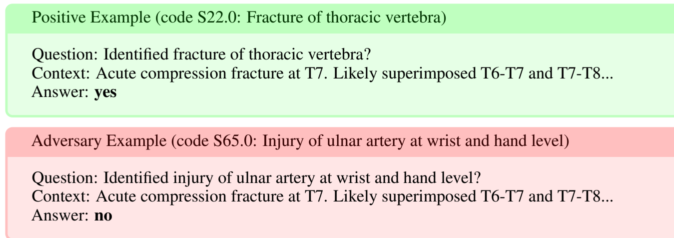


Figure 3: Two training examples that were use to fine-tune PubMedBERT to extract injury-specific codes. The positive example is generated by retrieving the description of patient's ICD10 code, while the negative, adversary example is randomly sampled from all possible ICD10 injury descriptions.

Note that several other settings were explored, including different difficulties and ratios of adversarial examples, and formatting of inputs (prompting). When more than one adversarial example was generated per positive example, the model performance decreased. Since we want to keep all the positive examples (full set of training labels), we maintained a 1-to-1 ratio between positive and negative codes. We also found that generating especially difficult adversarial examples (e.g. codes that describe very similar injuries but were incorrect) did not boost model performance. For example, when we generated negative code S23 (dislocation and sprain of joints and ligaments of thorax) for the positive code S22 (fracture of ribs, sternum and thoracic spine), it introduced noise that confuses model. Lastly, we explored splitting the conjuctions in the description into separate parts (e.g. "fracture of ribs," "fracture of sternum," "fracture of thoracic spine").

4.3 Baseline

As a baseline comparison, we evaluated model performance using Amazon's Comprehend Medical (AWSCM) inference API.[17]. AWSCM is widely-used, and hence Amazon would be able to improve its model using the large amounts of training data. Therefore, AWSCM is a strong benchmark to compare our own model against. Additional baseline models, including GPT-3 or T5 with zero-shot learning was considered. GPT-3 was an infeasible option due to our data privacy concerns over sending patient data to OpenAI's hosted inference APIs. We found that T5 underperformed AWSCM in recall while showing similar precision. Since T5 outputs texts whose logits are difficult to extract, we decided to only compare models that output probabilities, leaving only AWSCM.

4.4 Hyper-Parameter Optimization

Due to the novel architecture, we performed an extensive hyper-parameter to obtain the highest accuracy possible. The hyper-paramter sweep was conducted in parallel across multiple V100 and P100 GPU instances (agents). All instances are orchestrated by a centralized sweep controller, which generates a set of hyper-parameter configurations to be explored. [18] After the model was trained for 5 epochs (potentially early-stopped by the Hyperband algorithm [19]), the controller dispatches another set of configurations to the GPU instance and wait for it to train another model. The agents automatically log the following: training loss, validation loss, validation accuracy, number of steps, number of batches, the hyper-parameters, and model checkpoints. The sweep process is repeated until the validation performance converges. The best-performing model is selected based on the highest validation accuracy, and was finally evaluated on an unseen holdout test set. We use the sweep results to obtain various additional metrics, such as a hyper-parameter's correlation and importance with respect to validation accuracy and loss.

4.5 Probing and Interpretability Methods

In order to facilitate interpretability of PubMedBERT outputs, we adapted a BERT visualization library, BertViz to produce the visualized attention maps of tokens at each layer and attention head. [20] The visualizations help surgeons understand and interpret the PubMedBERT model qualitatively.

5 Experiments

5.1 Data

We queried our institution's trauma registry data for adults (aged 18 years) who were admitted to the Stanford trauma service between January 2016 and June 2021. To capture patients most likely to have trauma surveys, we included patients who were admitted after trauma 97 and trauma 99 activations (the most serious two of three activation levels) and were hospitalized 2 days. As PubMED BERT is limited 520-token inputs, we extracted portions of the tertiary notes most likely to dictate ICD10 diagnoses (imaging reports and the injury list/impression).

After splitting our data ($N = 3513$ notes) into train-validation-test sets, we found that 197 of the 729 ICD10 codes had > 5 instances within the training dataset. We report report performance based only on specified ICD10 codes with > 5 instances, as including ICD10 codes with fewer instances may yield too high a variance for reliable generalizability.These ICD10 codes reflect injuries seen at one of the busiest trauma centers in the country; thus, we felt our findings would provide an appropriate foundation for future work.

A total of 3513 patients met the inclusion criteria (Appendix Supplemental Tables), among whom 3478 had both tertiary survey notes written and injury ICD10 diagnoses. The median character length of these notes ("imaging reports" and "list of injuries/impression" section) was 2280. The 3478 patients had a total of $16,090$ injury ICD10 codes assigned by trauma registrars (median 4 injury diagnoses per patient, Appendix Figure 7; total 572 unique injury ICD10 codes).

Using algorithm in Appendix Algorithm 1, we generated 55789 training examples from all the ICD10 codes associated with 3478 notes. On average, each patient was affected by 4.6 unique injuries. For each injury we generated 4 input-output examples (2 positive, 2 adversarial) using the ICD10 description and tertiary note. The training set contained 39009 examples, which was used to fine-tune the custom PubMedBERT model.

5.2 Evaluation method

Considerations were taken in choosing the evaluation metrics. While it is time-consuming to look up ICD10 diagnosis code associated with an injury, a trauma surgeon can easily assess whether a list of predicted ICD10 outputs are appropriate. Thus, minimizing false negatives is more important than minimizing false positives. Our most important evaluation metric was thus recall. Specifically, we evaluated recall@5 and recall@10, as up to 10 recommendations were deemed practical for trauma surgeons to quickly evaluate (i.e. we considered future model implementation). To provide a balanced understanding of model performance, we also evaluated precision@5, precision@10, micro-AUC, macro-AUC, micro-F1, and macro-F1, whose calculations adhere to the standard formulae (e.g. the micro-F1 scores are calculated by micro-averaging F1 scores).

In addition to primary analysis, we conducted sensitivity analysis by evaluating model performance on injury ICD10 diagnosis codes excluding those describing superficial injuries (third character "0"). Superficial injuries (e.g. minor bruises and cuts on the skin), though coded for billing purposes, are not of interest to trauma surgeons, as these do not meaningfully impact any patient outcome or affect clinical decisions. As such, superficial injuries are rarely detailed in trauma tertiary surveys (they are detailed in other parts of the EMR and thus transcribed into ICD 10 codes). As lack of input data that could reasonably derive superficial injury ICD10 codes could unfairly deflate model performance, we evaluated model performance on the subset of ICD10 codes excluding these superficial injuries.

To quantify the importance of input data availability, we compared model performance on the 50 most frequent ICD10 codes in concordance to previous studies. [10]

5.3 Experimental details

The dataset of 3513 patients were split into training, validation, and test sets with ratio of 70-15-15 (note that each patient has on average approximately 15 training examples associated with their tertiary note, making the total number of data points large enough to justify the 70-15-15 split). The training and validation dataset facilitates fine-tuning and hyperparameter-optimization. The holdout test dataset was used only after the model selection is complete, and was used to establish the performance of the model in unseen data, and to compare the model to AWSCM.

We fine-tuned the following parameters to minimize binary Cross-Entropy Loss: batch size, percent dropout in the dropout layer, learning rate, size of fully connected layers, and the number of training warmup steps. [21] We trained multiple models in parallel across multiple GPU instances to efficiently find the optimal hyperparameters. The hyper-parameter sweep was conducted through randomized search, which researchers found to be more efficient and fault-proof than grid-based hyper-parameter search. [22] The optimization objective of the sweeps are to minimize validation loss. Then the performance metrics were analyzed with interpretable visualizations of how each hyperparameter value affects model performance to demonstrate our model fine-tuning for the surgeon audience.

5.4 Results

After a number of initial experimental runs, we performed a comprehensive random sweep of the hyper-parameter space specified in Table 1. Notably, we randomly sample *warmup steps* and *FC layer size* in log-scale distributions to maximize performance (capturing the a wide range of possibilities) while preserving sweep efficiency (log-scale sampling). FC dropout and FC hidden size control the dropout rate and number of hidden neurons in the custom fully-connected classification head.

Hyper-Parameter	Range of Values	Sampling Distribution	Example Value
warmup steps	[100, 10000]	Log-Uniform	1000
learning rate	$[10^{-7}, 10^{-5}]$	Uniform	0.0000005
batch size	{1,4,16}	Uniform	4
FC hidden size	[16, 8192]	Log-Uniform	1024
FC dropout	{0, 0.15, 0.30, 0.45}	Uniform	0.30

Table 1: The hyper-parameter space on which the model optimization was performed. The space was chosen based on preliminary results from several initial experiments.

After extensive validation of various models (), the best model was selected based on high validation accuracy and low validation loss. The model then performed inference on the test set tertiary notes, and accuracy metrics were calculated using standard formulae. The micro and macro-scores are calculated by micro-averaging (example-level) and macro-averaging (code-level) the performance metrics. We report the performance of the fine-tuned PubMedBERT below.

Metric	All Codes		Non-Sup. Codes		Top-50 Codes		Top-10 Codes	
Metric	Ours	AWSCM	Ours	AWSCM	Ours	AWSCM	Ours	AWSCM
Accuracy	**0.958**	0.936	**0.969**	0.934	**0.923**	0.885	0.835	**0.842**
AUC_{macro}	**0.838**	0.684	**0.873**	0.709	**0.883**	0.764	**0.878**	0.798
AUC_{micro}	**0.895**	0.760	**0.920**	0.796	**0.897**	0.807	**0.887**	0.832
$F1_{macro}$	**0.188**	0.174	**0.206**	0.189	**0.383**	0.363	0.524	**0.529**
$F1_{micro}$	**0.341**	0.285	**0.406**	0.293	**0.465**	0.411	0.559	**0.569**
Precision@5	**0.331**	0.259	**0.334**	0.251	**0.328**	0.286	**0.244**	0.226
Precision@10	**0.220**	0.194	**0.225**	0.184	**0.217**	0.190	**0.136**	**0.136**
Recall@5	**0.469**	0.392	**0.562**	0.452	**0.658**	0.590	**0.926**	0.857
Recall@10	**0.579**	0.532	**0.705**	0.595	**0.811**	0.710	**1.000**	**1.000**

Table 2: Performance metrics of fine-tuned TraumaICD-BERT compared AWS Comprehend Medical (CM), grouped in four sets: 1) all 197 codes, 2) all 170 non-superficial codes, 3) the top 50 most frequent codes in the dataset, and 4) the top 10 most frequent codes.

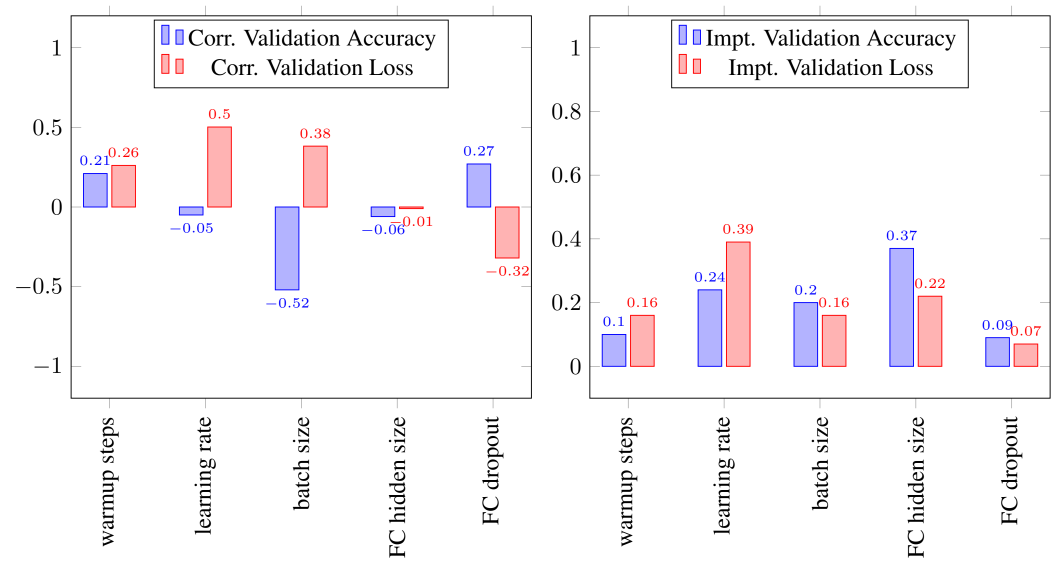


Figure 4: The correlation (corr.) and importance (impt.) scores of each hyper-parameter with respect to validation performance, averaged across all experiments. Note: the impt. metric was obtained by calculating the decrease in node impurity of a constructed hyper-parameter decision tree. [23]

6 Analysis

Your report should include *qualitative evaluation*. That is, try to understand your system (e.g. how it works, when it succeeds and when it fails) by inspecting key characteristics or outputs of your model.

The copious amount of domain-specific training data was perhaps one advantage of our method that allowed us to outperform AWSCM. Our labels were curated by expert ICD10 code reviewers, and the notes written by clinicians adhere to a consistent use of ontology and abbreviations. Using the tertiary notes from 3478 patients, we generated a over 39,000 input-output training examples, which helped PubMedBERT to robustly adapt to injury-specific reading-comprehension.

We analyzed how the hyper-parameters affect validation performances. Figure 6 shows the correlation and importance scores between validation performance and hyper-parameter configurations. There is moderate correlation between batch size and validation performance, with larger batch sizes leading to higher loss – this was expected since the validation loss is the average of the sum of individual losses in a batch. We also realized that picking a good learning rate and the hidden layer size for fully-connected classification head are the two most important metrics. A low learning rate prevents the model from over-fitting to quickly. Interestingly, although the hidden layer was an important metric, it showed no correlation with performance; One possible explanation is that changing the sizes of hidden layer increases model variance. More information is available in Figure 9 of the Appendix.

To facilitate interpretability and explainability, a critical need for clinical applications, we used visualization tools to interpret the model attentions. The PubMedBERT model's attention outputs can be difficult to interpret due to the sheer complexity. For example, the PubMedBERT model has 14 layers $\times$ 12 heads $=$ 168 unique attention structures for each token. An example is shown in Appendix. We attempt to analyze attentions visually for common patterns. Figure 5 shows an example input text "Question: Identified Traumatic subdural hemorrhage? Context: Injury List: Subdural hematoma 3 mm in right lateral convexity without midline shift."

In order to understand potential error cases, we display several tertiary notes, the model-extracted codes, and the ground truth in Appendix Table 11. Those examples are reviewed by the surgeon and engineers in order to plan to improve the model in future works.

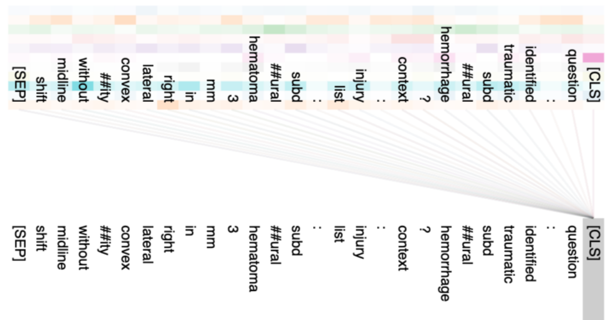


Figure 5: A head-wise visualization of PubMedBERT encoder-encoder self-attention, where each color represents one of PubMedBERT's 12 attention heads. Note this shows the attention of "[CLS]" because our custom fully-connected layer converts the 786-dimensional "[CLS]" embedding into a binary probability

7 Conclusion

A novel biomedical NLP dataset was collected from 3478 trauma patients. A dataset generation and augmentation algorithm yielded a classification dataset size of 55789 examples. A fully-connected classification head was added to a pre-trained PubMed model, whose hyper-parameter were extensively explored. The best model, selected based on validation performance, was evaluated on an unseen, holdout test-set. Our TraumaICD-BERT outperformed an industry baseline, Amazon Web Service Comprehend Medical, in terms of higher Micro-AUC, Macro-AUC, Micro-F1, Macro-F1, Precision, and Recall scores. Although future work is needed to process longer sequences of texts, this work is one step towards automating the extraction of ICD10 diagnosis codes for injury patients, with utility in providing on-time a set of input variables for downstream machine learning prediction tools for injury prognosis.

8 Acknowledgements

We thank our project mentor, Gaurab, for providing feedback and debugging tips throughout this project.

References

[1] Jinmiao Huang, Cesar Osorio, and Luke Wicent Sy. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. 177:141–153.

[2] Brent Biseda, Gaurav Desai, Haifeng Lin, and Anish Philip. Prediction of ICD codes with clinical BERT embeddings and text augmentation with label balancing using MIMIC-III.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding.

[4] Pei-Fu Chen, Ssu-Ming Wang, Wei-Chih Liao, Lu-Cheng Kuo, Kuan-Chih Chen, Yu-Cheng Lin, Chi-Yu Yang, Chi-Hao Chiu, Shu-Chih Chang, and Feipei Lai. Automatic ICD-10 coding and training system: Deep neural network based on supervised learning. 9(8):e23230. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

[5] Biplob Biswas, Thai-Hoang Pham, and Ping Zhang. Transicd: Transformer based code-wise attention model for explainable icd coding. In *International Conference on Artificial Intelligence in Medicine*, pages 469–478. Springer, 2021.

[6] Ssu-ming Wang, Yu-hsuan Chang, Lu-cheng Kuo, Feipei Lai, Yun-nung Chen, Fei-yun Yu, Chih-wei Chen, Zong-wei Li, and Yufang Chung. Using deep learning for automatic icd-10

classification from free-text data. 16(1). Publisher: International Journal of Medical Research & Health Sciences.

[7] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, January 2022. arXiv: 2007.15779.

[8] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, page btz682, September 2019. arXiv: 1901.08746.

[9] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv:1904.05342 [cs]*, November 2020. arXiv: 1904.05342.

[10] Biplob Biswas, Thai-Hoang Pham, and Ping Zhang. TransICD: Transformer Based Code-Wise Attention Model for Explainable ICD Coding. In Allan Tucker, Pedro Henriques Abreu, Jaime Cardoso, Pedro Pereira Rodrigues, and David Riaño, editors, *Artificial Intelligence in Medicine*, Lecture Notes in Computer Science, pages 469–478, Cham, 2021. Springer International Publishing.

[11] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv:1903.10676 [cs]*, September 2019. arXiv: 1903.10676.

[12] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *arXiv:1906.05474 [cs]*, June 2019. arXiv: 1906.05474.

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

[14] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. UnifiedQA: Crossing Format Boundaries With a Single QA System. *arXiv:2005.00700 [cs]*, October 2020. arXiv: 2005.00700.

[15] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. page 67.

[16] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv:1606.05250 [cs]*, October 2016. arXiv: 1606.05250.

[17] Parminder Bhatia, Busra Celikkaya, Mohammed Khalilia, and Selvan Senthivel. Comprehend medical: a named entity recognition and relationship extraction web service.

[18] Hyperparameter Tuning - Weights and Biases.

[19] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *arXiv:1603.06560 [cs, stat]*, June 2018. arXiv: 1603.06560.

[20] Jesse Vig. A Multiscale Visualization of Attention in the Transformer Model. *arXiv:1906.05714 [cs]*, June 2019. arXiv: 1906.05714.

[21] Weights & biases.

[22] James Bergstra and Yoshua Bengio. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13(10):281–305, 2012.

[23] Don Coppersmith, Se June Hong, and Jonathan R.M. Hosking. Partitioning Nominal Attributes in Decision Trees. *Data Mining and Knowledge Discovery*, 3(2):197–217, June 1999.

[24] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning. *arXiv:2004.11362 [cs, stat]*, March 2021. arXiv: 2004.11362.

A Appendix

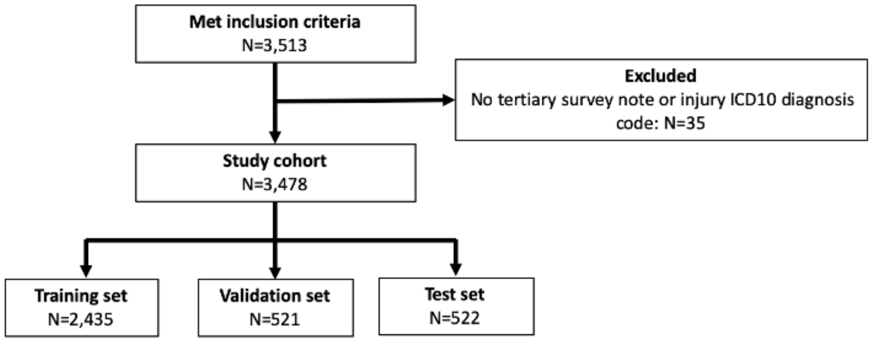
Supplemental table 1: Previous studies evaluating ICD10 outcomes

Author	Data source	ICD 9/10
Wang et al.	All unstructured text within the electronic medical record during a patient's hospitalization	ICD10
Chen et al.	All unstructured text within the electronic medical record during a patient's hospitalization	ICD10
Biseda et al	All unstructured text within MIMIC-III database	ICD9 (100 most frequent)
Biwas et al	Discharge summary within MIMIC-III database	ICD9 (50 most frequent)
Huang et al	Discharge summary within MIMIC-III database	ICD9 (50 most frequent)

Supplemental table 1: Study cohort characteristics. SMD= standardized mean difference

	Training set N=2429	Validation set N=521	Test set N=522	p
Age, median (IQR), years	62.00 [42.00, 78.00]	62.00 [39.00, 78.00]	63.00 [44.00, 78.00]	0.769
Male, No. (%)	1521 (62.6)	340 (65.3)	331 (63.4)	0.521
Race, No. (%)				0.398
Asian/Pacific Islander	325 (13.4)	74 (14.2)	61 (11.7)	
African American	58 (2.4)	12 (2.3)	10 (1.9)	
Native American	3 (0.1)	1 (0.2)	1 (0.2)	
Not specified	18 (0.7)	2 (0.4)	5 (1.0)	
Other	639 (26.3)	131 (25.1)	146 (28.0)	
Caucasian	1371 (56.4)	296 (56.8)	289 (55.4)	
Cause of Injury, No. (%)				0.318
ASSAULT	60 (2.5)	13 (2.5)	18 (3.4)	
ATV	5 (0.2)	3 (0.6)	0 (0.0)	
BIKE	287 (11.8)	53 (10.2)	59 (11.3)	
CUT	3 (0.1)	1 (0.2)	1 (0.2)	
FALL	1141 (47.0)	237 (45.5)	246 (47.1)	
FIREARM	2 (0.1)	2 (0.4)	3 (0.6)	
GSW	20 (0.8)	6 (1.2)	0 (0.0)	
MCC	178 (7.3)	39 (7.5)	38 (7.3)	
MV	156 (6.4)	30 (5.8)	40 (7.7)	
MVC	328 (13.5)	82 (15.7)	73 (14.0)	
O_BLUNT	71 (2.9)	14 (2.7)	7 (1.3)	
O_PEN	16 (0.7)	6 (1.2)	2 (0.4)	
PED	121 (5.0)	27 (5.2)	23 (4.4)	
SCOOTER	9 (0.4)	0 (0.0)	1 (0.2)	
STAB	31 (1.3)	8 (1.5)	11 (2.1)	
UNK	1 (0.0)	0 (0.0)	0 (0.0)	
ISS (median [IQR])	10.00 [8.00, 17.00]	10.00 [8.00, 17.00]	10.00 [5.00, 16.00]	0.114
LOS (median [IQR])	3.00 [1.00, 6.00]	3.00 [1.00, 6.00]	3.00 [1.00, 6.00]	0.467

Supplemental figure 1: Flow diagram showing study cohort selection



Supplemental figure 2: Length of characters in input tertiary survey notes

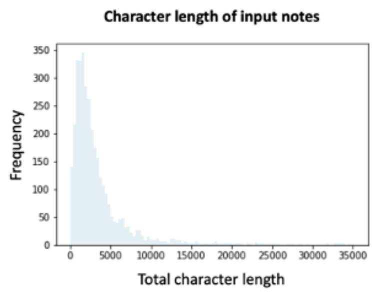


Figure 6: Supplemental Flow diagram showing study cohort selection

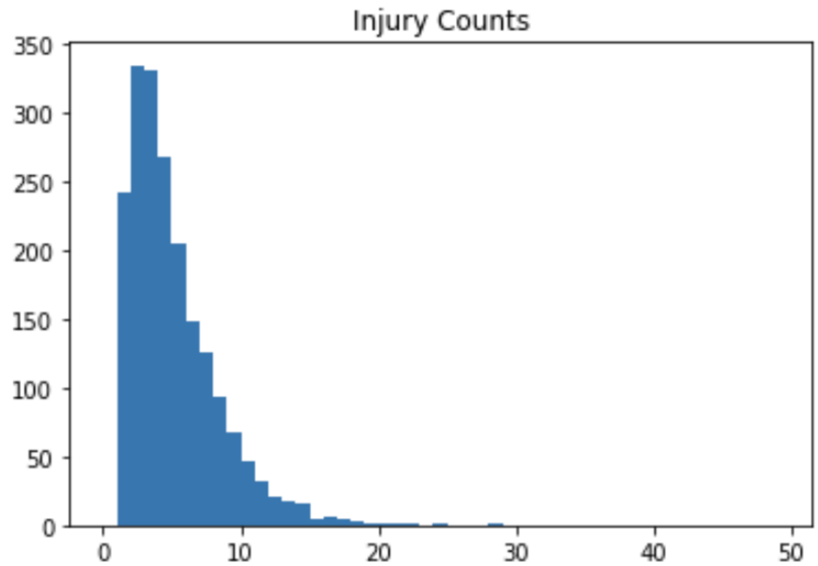


Figure 7: The number of patients who has a specific number of injuries that experts extracted ICD10 codes from their tertiary notes.

Algorithm 1: Algorithm for generation and augmentation of ICD10 code-wise classification dataset. In summary, The algorithm compiles the positive ICD10 codes from each note, then randomly generates the same number of negative codes as adversary examples (a technique resembling supervised contrastive learning [24]).

foreach $P_i \in \{P_1, ..., P_{3478}\}$ **do**
    initialize Q as an empty list of input-output pairs
    foreach $C_{ij}^{(7)} \in \{C_{i1}^{(7)}, ..., C_{iN_i}^{(7)}\}$ **do**
        $C_{ij}^{(3)} \leftarrow$ *first 3 digits of* $C_{ij}^{(7)}$ ▷ *E.g. S22* $\leftarrow S22.000A$
        Input $\leftarrow$ *"Question: Identified* $D_{C_{ij}^{(3)}}$ *? Context:* T_i*"*
        Output $\leftarrow$ *yes*
        $Q \leftarrow Q \cup \{(Input, Output)\}$
        $C_{ij}^{(4)} \leftarrow$ *first 4 digits of* $C_{ij}^{(7)}$ ▷ *E.g. S22.0* $\leftarrow S22.000A$
        Input $\leftarrow$ *"Question: Identified* $D_{C_{ij}^{(4)}}$ *? Context:* T_i*"*
        Output $\leftarrow$ *yes*
        $Q \leftarrow Q \cup \{(Input, Output)\}$
        $\bar{C}_{ij}^{(3)} \leftarrow$ *randomly sample a 3-digit injury code* $\notin \{C_{i1}^{(3)}, ..., C_{iN_i}^{(3)}\}$
        Input $\leftarrow$ *"Question: Identified* $D_{C_{ij}^{(3)}}$ *? Context:* T_i*"*
        Output $\leftarrow$ *no*
        $Q \leftarrow Q \cup \{(Input, Output)\}$
        $\bar{C}_{ij}^{(4)} \leftarrow$ *randomly sample a 4-digit injury code* $\notin \{C_{i1}^{(4)}, ..., C_{iN_i}^{(4)}\}$
        Input $\leftarrow$ *"Question: Identified* $D_{C_{ij}^{(4)}}$ *? Context:* T_i*"*
        Output $\leftarrow$ *no*
        $Q \leftarrow Q \cup \{(Input, Output)\}$
    end
end

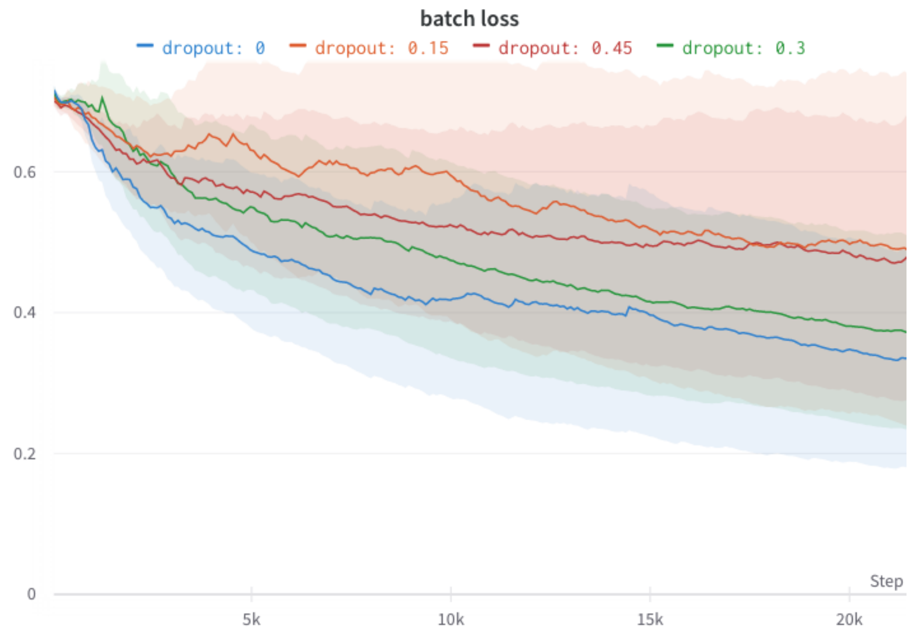


Figure 8: Batch loss during training, with respect to different dropout rates, with shaded area indicating the standard errors. Although dropout rate of 0 achieves the lowest training loss, dropout rates of 0.3 or higher resulted in lower validation loss – a sign that dropout is necessary to prevent overfitting

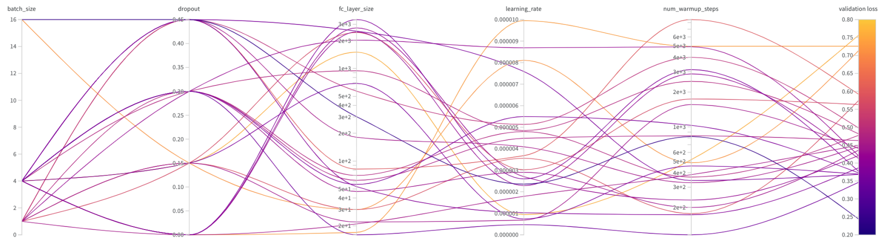


Figure 9: A parallel coordinates chart visualizing the various hyper-parameters with respect to validation loss. Each line represents one configuration of learning rate, batch size, and etc. Since the objective is to minimize validation loss, the darker lines are desired.

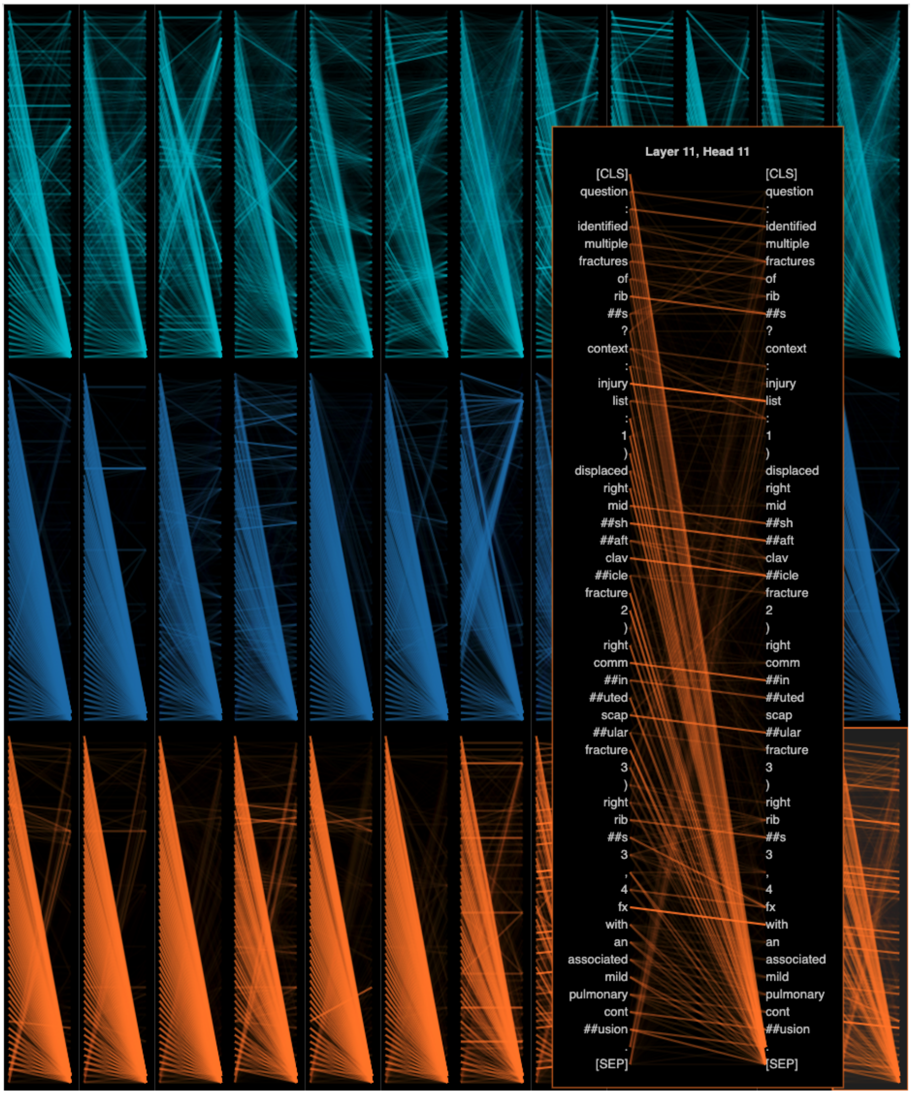


Figure 10: A model-wise attention visualization of PubMedBERT on the input text "question: Identified Multiple fractures of ribs? context: INJURY LIST: 1) Displaced right midshaft clavicle fracture 2) Right comminuted scapular fracture 3) Right ribs 3, 4 fx with an associated mild pulmonary contusion.". The Y-Axis corresponds to different layers, while the X-Axis correspond to attention heads. Note that only 3 out of 12 layers' attentions are shown to fit on the page.

Tertiary Notes	Extractions (sorted by decreasing probability)	Ground Truth
1. Mild compression fractures within the T11, T12, and L1 vertebral bodies with close to 20% loss of vertebral body height in the L1 vertebral body. 2. Nondisplaced fracture of the T12 spinous process 3. Midline c spine pain. CT Abd Pelvis: IMPRESSION: 1. No traumatic visceral injury in the abdomen and pelvis. 2. Mild compression fractures within the T11, T12, and L1 vertebral bodies with close to 20% loss of vertebral body height in the L1 vertebral body. Nondisplaced fracture of the T12 spinous process. See dedicated CT T and L-spine for further findings. CTA Chest: IMPRESSION: 1. Superior endplate compression fractures of T11, T12 and L1 with approximately 20% vertebral body height loss of L1. No retropulsed fracture fragments. 2. Nondisplaced fracture of the T12 spinous process. 3. Cortical irregularity of the anterior and posterior aspects of the manubrium which is felt to be related to streak artifact given that no surrounding hematoma. However, may represent an acute fracture. Correlate with point tenderness. CT Head Cervical Spine IMPRESSION: 1. No acute intracranial abnormality. 2. Small left frontal scalp contusion. 3. No traumatic injury within the cervical spine. CT Lumbar Spine: IMPRESSION: 1. Mild compression fractures in the T11, T12, and L1 vertebral bodies with close to 20% height loss at L1, in addition to a nondisplaced spinous process fracture at T12. No retropulsion or epidural hematoma. CT Thoracic Spine: IMPRESSION: 1. Mild compression fractures in the T11, T12, and L1 vertebral bodies with close to 20% height loss at L1, in addition to a nondisplaced spinous process fracture at T12. No retropulsion or epidural hematoma. CXR: IMPRESSION: 1. No acute cardiopulmonary disease. Pelvis XR: IMPRESSION: 1. No radiographically visible fracture. If there is concern for fracture, cross-sectional imaging with CT or MRI would be recommended.	S00.0,Superficial injury of scalp S01.0,Open wound of scalp S32.0,Fracture of lumbar vertebra S24.1,Other and unspecified injuries of thoracic spinal cord S00.2,Other and unspecified superficial injuries of eyelid and periocular area S30.8,Other superficial injuries of abdomen, lower back, pelvis and external genitals	S22.0,Fracture of thoracic vertebra, S32.0,Fracture of lumbar vertebra, S00.0,Superficial injury of scalp
INJURY LIST: 1. Pelvic fractures- comminuted and mildly displaced L superior/inferior pubic ramus, possible non-displaced R inferior pubic ramus fracture, nondisplaced fracture L sacral ala -No operative intervention warranted per Ortho - WBAT, PT, pain control - F/u with Dr. Stephanie Pun in Redwood City Orthopaedic Surgery Clinic in 4-6 weeks (please place outpatient referral at discharge 2. Retropubic hemorrhage, possible active venous hemorrhage. -No need for embolization per IR. H&H downtrending to 7.1 this AM and plan for 2 unit(s) pRBC today. 3. L 5th proximal phalanx fracture. CT Pelvis Cysto 8/12 1. No evidence of a bladder injury. 2. Redemonstration of the numerous pelvic fractures as described previously with extraperitoneal hemorrhage predominantly involving the retropubic space. Evaluation for active extravasation cannot be done in the absence of IV contrast. CT Pelvis 8/12 1. Mildly comminuted and displaced fractures of the left superior and inferior pubic rami extending to pubic symphysis with extraperitoneal hemorrhage and evidence of active extravasation in the retropubic space (may be venous origin). 2. Delayed imaging demonstrates no evidence of bladder leak; however the bladder is not well distended with contrast and hence an underlying bladder injury cannot be completely excluded. If there is persistent concern for bladder injury, a CT cystogram should be considered. 3. Nondisplaced fracture of the left sacral ala, and possible nondisplaced fracture of the right inferior pubic ramus. 4. Mild circumferential thickening of the distal stomach, which is nonspecific but can be seen in the setting of gastritis. CT Head 8/11 1. No acute intracranial hemorrhage. XR L wrist 8/11 1. Comminuted mildly displaced fracture of the base of the fifth proximal phalanx with possible extension into the MCP joint space. 2. Mild soft tissue swelling around the wrist with no underlying fracture or malalignment at the wrist. 3. Osteoarthritis of the hand. XR L hand 8/11 1. Comminuted mildly displaced fracture of the base of the fifth proximal phalanx with possible extension into the MCP joint space. 2. Mild soft tissue swelling around the wrist with no underlying fracture or malalignment at the wrist. 3. Osteoarthritis of the hand. CXR 8/11 1. No significant interval change in persistent retrocardiac opacity, compatible with atelectasis or infection. XR Pelvis 8/11 1. Mildly displaced left superior and inferior pubic rami fractures. 2. Healing left intertrochanteric fracture status post internal fixation without evidence of acute hardware complication.	S32.5,Fracture of pubis	S32.5,Fracture of pubis, S62.6,Fracture of other and unspecified finger(s), S60.2,Contusion of wrist and hand, S32.1,Fracture of sacrum
INJURY LIST and plan 1. Left first and third through sixth rib fractures -pain control -pulmonary toilet, IS 2. Left pneumothorax status -pigtail catheter placed with resolution of PTX -changed to water seal today 3. Comminuted fracture of the left scapula involving the scapular body and glenoid neck. -orthopedic surgery consult -NWB LUE -full ROM LUE as tolerated -no need for sling immobilization. 06/17/2017 CXR -Serial radiographs of the chest demonstrate interval placement of left-sided pigtail pleural catheter with complete radiographic resolution of the left pneumothorax. -Acute mildly displaced fractures of four contiguous left-sided ribs. -Comminuted fracture of the left scapula. 06/17/2017 CT head IMPRESSION: 1. No acute intracranial abnormality. 2. Punctate foci of debris along the surface of the forehead. 06/17/2017 CT chest/abd/pelvis IMPRESSION: 1. Left first and third through 7th rib fractures with associated trace left hemopneumothorax status post left pleural pigtail catheter placement. 2. Comminuted fracture of the left scapula involving the scapular body and glenoid neck. 3. No thoracic vascular injuries. 4. No acute intra-abdominal injury. 5. 1.4 cm enhancing lesion in the hepatic dome appears isodense on delayed images and may represent a vascular shunt, but is incompletely evaluated. 06/17/2017 left shoulder 2V XR IMPRESSION: 1. Comminuted fracture of the left scapular body and neck. 2. Well-corticated ossicle adjacent to the acromion may be related to prior trauma versus basi-acromial type of os acromiale. 3. Multiple left-sided rib fractures with a left pleural pigtail drainage catheter in place. 06/18/2017 left shoulder 2V XR IMPRESSION: 1. And overall mildly increased displacement of a small fragment of the comminuted left scapular fracture just below the glenoid. Otherwise no significant interval change	S22.4,Multiple fractures of ribs, S27.3,Other and unspecified injuries of lung, S27.0,Traumatic pneumothorax, S42.1,Fracture of scapula, S20.3,Other and unspecified superficial injuries of front wall of thorax, S27.2,Traumatic hemopneumothorax	S42.1,Fracture of scapula, S27.0,Traumatic pneumothorax, S22.4,Multiple fractures of ribs
INJURY LIST: 1) Displaced right midshaft clavicle fracture 2) Right comminuted scapular fracture 3) Right ribs 3, 4 fx with an associated mild pulmonary contusion. CTA Chest: 1. Mildly displaced fractures of the lateral right third rib and posterior right fourth rib with mild pulmonary contusion. There is no pneumothorax or evidence of acute vascular injury. 2. Redemonstration of comminuted fractures of the right clavicle and right scapula. CT Thoracic Spine: 1. Focal cortical step-off involving the left lateral aspect of the T9 vertebral body may represent a small minimally displaced fracture. Correlate with point tenderness. 2. Partial visualization of right third rib head fracture as well as right scapular and clavicular fractures. CT Head and Cervical Spine: 1. No CT evidence of acute intracranial traumatic injury. 2. Cervical spondylosis without evidence of acute cervical osseous injury. 3. Partial visualization of right clavicular, scapular, and right third rib head fractures, please refer to concurrent CT chest. Findings were discussed with Dr. Zho by Dr. Kuhn at 6:12 PM on 6/7/2017. There are no substantial differences between the preliminary results and the impressions in this final report. "Physician to Physician Radiology Consult Line: (650) 736-1173" Signed XR Chest: 1. Displaced right midshaft clavicular fracture. Recommend dedicated radiographs of the clavicle for further characterization. 2. Minimally displaced fractures of the right third and likely fourth ribs. No pneumothorax. XR Right Clavicle: 1. Displaced fracture of the right midclavicle. 2. Mild widening of the acromioclavicular joint, concerning for separation. XR Pelvis: 1. No fracture or malalignment. XR Shoulder 2 Views Right: 1. Mildly comminuted and displaced mid shaft clavicular fracture. 2. Comminuted scapular fracture, better demonstrated on same-day CT chest dated 6/7/2017. 3. Minimally displaced fractures of the right lateral third and fourth ribs, better demonstrated on same-day CT chest.	S22.4,Multiple fractures of ribs, S27.3,Other and unspecified injuries of lung, S42.0,Fracture of clavicle	S22.4,Multiple fractures of ribs, S06.0,Concussion, S22.0,Fracture of thoracic vertebra, S27.3,Other and unspecified injuries of lung, S42.1,Fracture of scapula, S42.0,Fracture of clavicle
1) Subdural hematoma 3 mm in right lateral convexity without midline shift 2) Intraventricular hemorrhage in left occipital horn 3) Possible focal SAH at left sylvian fissure 4) T11 superior endplate fracture, age indeterminate. Ct Thoracic Lumbar Spine Wo Iv Contrast ¬† Result Date: 10/24/2020 IMPRESSION: 1. Age-indeterminate superior endplate compression fracture of T11 with approximately 10% height loss affecting the left lateral aspect of the vertebral body with no retropulsion or paraspinal hematoma. Recommend correlation for point tenderness in the setting of trauma. No prior examinations are available for comparison. 2. Multilevel degenerative changes of the thoracolumbar spine with severe degenerative facet joint osteoarthritis with no high-grade bony spinal canal or foraminal narrowing as described above. 3. Greater than 20 solid bilateral pulmonary nodules and left upper lobe pulmonary mass, concerning for metastatic disease, possibly with lung primary. Further evaluation with a dedicated chest CT would likely provide a better evaluation for suspected metastatic disease. These findings were discussed over the phone by Dr. Bates with Dr. Adkar at 2141 hours on 10/24/2020. For over 60 years: Some imaging findings are common, even in normal, pain-free volunteers. Among people over the age of 60 who do not have back pain, a CT scan will find that about: - 9 in 10 have disk degeneration - 9 in 10 have disk signal loss or desiccation - 8 in 10 have disk height loss - 8 in 10 have a disk bulge - 4 in 10 have a disk protrusion - 4 in 10 have an annular fissure - 4 in 10 have facet degeneration - 3 in 10 have spondylolisthesis I have personally reviewed the images for this examination and agree with the report transcribed above. Signed""Final report"" ¬† Ct Head Wo Iv Contrast ¬† Result Date: 10/25/2020 IMPRESSION: 1. Stable slightly redistributed right lateral high convexity subdural hematoma with no mass effect on underlying brain parenchyma, midline shift, or herniation. 2. Unchanged trace intraventricular hemorrhage layering in the left occipital horn of the lateral ventricles. 3. Unchanged punctate hyperdensity in the anterior body of the left corpus callosum which again is of uncertain clinical significance and given the stability on subsequent imaging is favored to represent a cavernoma, however, intraparenchymal hemorrhage is still not completely excluded. This can be followed in subsequent imaging. Stable to decreased conspicuity of the tubular hyperdensity in the anterior left sylvian fissure which is again favored to represent a partially calcified vessel. I have personally reviewed the images for this examination and agree with the report transcribed above. Signed""Final report"" ¬† Ct Head Wo Iv Contrast ¬† Result Date: 10/24/2020 IMPRESSION: 1. 3 mm high right lateral convexity subdural hemorrhage measuring up to 3 mm in maximal thickness with no significant mass effect on the underlying brain parenchyma. This subdural hemorrhage is new/more conspicuous when compared to the previous CT. 2. New/increase in conspicuity minimal intraventricular hemorrhage layering dependently in the left occipital horn. 3. Stable tubular hyperdensity in the anterior left sylvian fissure which is favored to represent a partially calcified vessel given its unchanged appearance. Stable hyperdensity in the anterior body of the left corpus callosum which again is of uncertain clinical significance and could represent a cavernoma however, in the setting of trauma with multiple intracranial hemorrhages, intraparenchymal hemorrhage is not excluded. These findings were discussed over the phone by Dr. Bates with Dr. Adkar at 2055 hours on 10/24/2020. I have personally reviewed the images for this examination and agree with the report transcribed above. Signed""Final report"" ¬† Ct Head Cervical Spine Wo Iv Contrast Trauma ¬† Result Date: 10/24/2020 IMPRESSION: 1. Somewhat tubular-appearing hyperdensities in the left sylvian fissure and in the superior aspect of the left lateral ventricle. Given the patient's history of fall and the location of the soft tissue swelling, this could represent an atypical appearance of subarachnoid or intraventricular hemorrhage. However these could also just represent	S06.5,Traumatic subdural hemorrhage, S06.6,Traumatic subarachnoid hemorrhage, S06.3,Focal traumatic brain injury, S00.0,Superficial injury of scalp, S06.8,Other specified intracranial injuries, S24.1,Other and unspecified injuries of thoracic spinal cord	S06.5,Traumatic subdural hemorrhage, S06.8,Other specified intracranial injuries

Figure 11: Five examples comparing model's ICD10 code extractions to ground truth codes.