ED Radiology Report Label Extraction

Stanford CS224N Custom Project

Jenny Shi
Department of Chemistry
Stanford University
jshi7@stanford.edu

Iris Xia
Department of Statistics
Stanford University
ixxia@stanford.edu

Serena Zhang
Department of Computer Science
Stanford University
serena2z@stanford.edu

Abstract

Machine learning has great potential to automate the interpretation and processing of medical imaging data. We focus on labelling written CT and X-Ray scan reports, which are doctor-written summaries of the scans. Label extraction is challenging because there are many ways to describe observations and observations are mentioned both in the context of presence, absence, and uncertainty. Existing labellers only can label X-ray and CT chest data. We aim to predict presence and location of relevant emergency medicine labels (specific diseases and symptoms) in 4 distinct datasets of radiology reports. We first applied and modified interpretable, rule-based methods (Sarle-Rules and CheXpert), and created our own Regex methods to automatically extract labels from each dataset. We also modified and fine-tuned the transformer-based CheXbert (with three different pretrained models) to extract labels. We generated new generalizable labels for ED-relevant CT abdomen and CT head reports, and labelled over 100,000 reports with accuracy comparable to current state-of-the-art methods. We also found that pretraining on ClinicalBERT generates the most accurate labels, and that training on two Rules-Based method labels increases accuracy. Our models/ methods can be used to extract these labels from new datasets.

1 Key Information to include

- Mentor: Abhinav Garg
- External Collaborators (if you have any): Professor David Kim, MD, PhD.
- Sharing project: No

2 Introduction

In recent years, the COVID-19 pandemic has put incredible strain on the emergency department to operate far above its normal capacity. Crowding of hospitals has results in increased wait times and even higher mortality. The previous capacity prior to COVID-19 was 80k patients, but it has since been raised to 150k with an average admitted patient staying for 2.3 days. It is possible to reduce the average patient stay time and overall patient number in hospitals by analyzing patient emergency department (ED) radiology reports, which detail the severity of a patient's injuries. If we are able to extract critical information from radiology reports, we can optimize resources in the ED for more efficient patient care.

In general, radiological label extraction is important for many clinical applications, from improving medical workflow prioritization to the training of medical imaging models. Because of their unstructured nature, NLP techniques are a promising tool to tackle classification of clinical observations within radiology reports. Current methods for this task rely on powerful language models, such as transformers, that utilize contextual information to improve the understanding of medical text and capture intricate relationships between medical concepts. However, these models require a large amount of labeled training data to achieve high accuracy, and hiring experts to annotate medical reports is costly and time-inefficient.

To resolve the issue of manual labeling, our approach fine-tunes a pre-trained BERT model on labels generated from existing rules-based labelers. We then manually label a portion of our data to evaluate our rules-based labelers and BERT models. Our approach improves on the previous work of models like CheXbert by generalizing to all types of radiology reports, including head, abdomen, and chest CT scans, not just chest x-rays. We also experiment with different pre-trained transformer models to compare performance. Creating this improved model will allow us to identify important information needed to optimize ED resources in Stanford hospitals and beyond, leading to better patient care and outcomes.

3 Related Work

Prior to the invention of powerful language models, various rules-based methods were the main choice for extracting labels from radiology reports. Previous groups have used advanced RegEx systems, like NegEx, or dependency parsing methods, like NegBio, to detect pathologies in chest x-ray reports (Chapman et al., 2002; Peng et al., 2017). In particular, CheXpert expands upon NegBio by extracting “mentions” of 14 conditions commonly found in chest x-rays and classifying them as uncertain, negative, or positive (Irvin et al., 2019). These "mentions" are manually curated by a group of expert radiologists. CheXpert uses dependency parsing to infer the structure of the sentences, then uses parsing-based rules to infer negation and uncertainty in the sentence (the remaining sentences are labelled as positive).

Other methods, like Sentence Analysis for Radiology Label Extraction (SARLE), utilize extensively customized rules, incorporating negation detection as well as “normality detection” to detect labels in a sentence. Sarle specifically mentions that it can better handle the minority of sentences that include both a normal and an abnormal statement (e.g., “the heart is enlarged without pericardial effusion”). It does this by detecting where a negation phrase is (ex. without), then deleting a segment before or after to remove the label (ex. pericardial effusion) (Rachel Draelos, 2021). Overall, rules-based methods offer an advantage of interpretability over other models. While these methods have generated considerable results, they are heavily dependent on rules that can be generalized to match text variations. As a result, they are unable to capture the full diversity of complexities and ambiguities of natural language in the context of radiology reports. Moreover, they require extra work and expert annotation to craft the rules and monitor the edge cases.

The development of powerful large language models that can be adapted to medical text offers a new approach for extracting radiology labels. Early models trained convolutional neural networks (CNNS) with glove embeddings, as well as recurrent neural networks (RNNS) with attention mechanisms (Bustos et al., 2019). More recently, transformer based models are being used as end-to-end solutions for radiology report labeling. Drozdov et al. (2020) utilized BERT and XLNet classifiers, while Wood and Lynch (2020) developed a classifier called ALARM using BioBERT models, both of which were used in the analysis of radiologist-labeled reports. Although large models do not require advanced rules or feature engineering, they do require massive amounts of manually labeled data for training. Models like CheXbert address this concern by using a clinically pretrained BERT model first trained on the CheXpert auto-labeled dataset and then fine-tuned on a dataset of expert annotations augmented with automated back-translation to achieve more accurate automated radiology report labeling (Smit et al., 2020). CheXbert has a statistically significant improvement over CheXpert (0.743) with a F1 score of 0.798.

4 Approach

Our approach is to adapt existing rules-based and transformer-based labellers to label new types of datasets. We chose to first apply rules-based methods because they can label large datasets. We evaluate the accuracy of these interpretable labels on a manually labelled set of reports. Then, we feed our labelled datasets into transformer models with the goal of increasing accuracy of our labelling.

4.1 Baselines

We ran CheXpert, CheXbert, and Sarle on XR chest on their existing labels and computed F1 score to evaluate the accuracy on the existing 14 labels (not entirely the same as our labels) (Figure 1).

4.2 Preprocessing and Manual Labelling

Our dataset has four distinct types of radiology reports, namely XR chest, CT head, CT abdomen, and CT chest scans. Data was separated by dataset type (ex. CT chest) and sorted by CSN number, a unique identifier for each radiology report. We removed duplicate reports (by CSN) and converted reports to all lower case. To establish a gold standard for evaluation, we manually labelled the first 100 reports from each dataset (400 reports total).

4.3 Rules-based Labelling

We labelled our 160000 reports with Regex (our own code), CheXpert (modified), and Sarle-Rules (modified).

Regex. To retrieve labels for the 4 different types of radiology reports, we implemented custom regular expressions. We carefully examined the reports and identified recurring patterns or phrases associated with each label. In our code, each sentence in a report is considered separately. We adopted a negation-first strategy, where we first identified if there were mentions of the label, then excluded instances of negative occurrence (ex. **no evidence** of pneumonia or **no acute** intracranial hemorrhage). Each label has a specific set of negation patterns. Regex is a simple but powerful baseline approach, and since it is customized to our data, we thought it could outperform existing models. For example, CheXpert has been optimized only for XR chest but not for CT chest, CT head, and CT abdomen.

Sarle-Rules. Sarle-Rules is a rules-based labeler designed to extract 83 structured labels from only chest CT and XR reports. Each label uses a list of pre-optimized negation patterns (either CT-specific or XR-specific patterns). Sarle-Rules is able to extract modifiers (such as whether an observation is in the left/right lung), so we also labelled the locations of postive observations for CT chest and XR chest data. We made several changes to the code. First, we added new negation patterns based on examination of our data that were missing from Sarle-Rules. We also added new labels and synonyms of these labels, as Sarle did not cover all of our chest labels and covered none of our head/abdomen labels. These changes allow us to extract new labels from all our datasets. Note that Sarle-Rules outputs only positive or negative for each label, while Regex generates positive, negative, and no finding.

CheXpert. CheXpert extracts 14 labels tailored to XR chest data. While most of these labels aligned with our XR chest and CT chest labels, we coded additional observations such as "Aspiration" and "Infection". There were no overlapping labels in our CT abdomen and CT head datasets with CheXpert, so we introduced entirely new CT abdomen and CT-head specific labels as "mention" phrases in the code. We did not modify the negation patterns.

4.4 Transformer-Based Labelling

We selected the best performing rules-based labeler (Regex) and used the labeled data as input to train a BERT transformer model. We fine-tuned three pre-trained BERT models - BERT (modified), ClinicalBERT (modified), and CheXbert (modified) using our labelled datasets to determine the optimal model for our dataset. We chose to use BERT-based models because they have been proven to perform well on language labeling tasks (cite CheXbert here maybe). We also chose to use models specifically pretrained on clinical data similar to our input.

Our model architecture (Figure 1 below) is adapted from CheXbert. using the pretrained BERT model, we added linear heads for each label (i.e. "pneumonia", "pneumothorax", etc.) according to the type of scan we were training on. Each linear head predicted one of three classes: positive, meaning that the condition was present in the report, negative, meaning it was found to be not present, and no mention. This framework can be generalized for any dataset with the labels being the conditions we're trying to detect.

Using this architecture, we trained our data on three new BERT-based models:

BERT. For our BERT baseline model, we built linear heads on top of the BERT-base-uncased model, the original BERT model trained on a large English language corpus.

CheXbert. The CheXbert model has been trained to extract labels from radiology reports from X-ray chest scans. In this model, we used a CheXbert checkpoint to train the BERT model with our new labels, generalizing this work to CT chest, CT abdomen, and CT head radiology reports.

ClinicalBERT. We also built linear heads on top of the ClinicalBERT model (Alsentzer et al., 2019), which was initialized using BioBERT then trained on all MIMIC notes from the MIMIC III database.

We trained each of the four datasets on the three models and evaluated their results. Reports are tokenized (the maximum number of tokens was 512), and concatenated with a CLS token, which we feed into the hidden layer for identification.

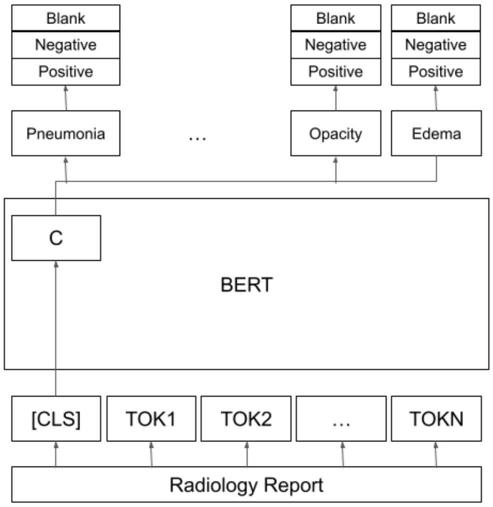


Figure 1: BERT model architecture. (adapted from CheXbert paper) We added 8-11 linear heads, depending on the size of the dataset, each with three classes. The BERT model was either BERT-base-uncased, ClinicalBERT, or ChexBert Checkpoint.

We used cross entropy loss (multiclass because we have 3 outcomes - positive, negative and blank) in our model.

$$\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

We also investigated whether combining different autolabeller data in the training set would improve BERT performance. We trained on both Sarle-Rules and Regex labelled data (CombinedBERT) with the hope that since Sarle-Rules and Regex get different edge cases wrong, that the BERT could noisily learn the edge cases slightly better and it might improve accuracy. Finally, we investigated whether fine-tuning on small amounts of manual data would increase model accuracy.

5 Experiments

5.1 Data

We have a csv file with 160,000 radiology reports. The radiology reports are separated into 4 categories: (1) X-Ray (XR) chest, (2) CT chest scans, (3) CT head scans, and (4) CT abdomen scans. Each report has a unique CSN and a couple sentences summarizing the results of the scan. For each dataset, our task is to extract a specific set of interpretable labels from the reports that can be used to assess the conditions and their severity.

5.2 Evaluation method

To evaluate the performance of our baseline, we manually labeled 100 radiology reports to serve as our "silver standard" (see above). We calculated the F1 score across all classes of labels as our evaluation metric, as F1 score is resistant to class imbalances. The equation for F1 score is as follows, where TP and FP are the true and false positive rates, respectively.

$\texttt{precision} = \frac{TP}{TP+FP}$ $\texttt{recall} = \frac{TP}{TP+FN}$
$F1 = \frac{2 \cdot \texttt{precision} \cdot \texttt{recall}}{\texttt{precision} + \texttt{recall}}$

We used the "weighted" averaging for F1 score in scikit (weighted by the support for each classification). The average F1 score is the mean of all F1 scores (not weighted by occurence of each label).

5.3 Experimental details

Rules-based methods. We used our code and existing code to label all the data for each dataset. F1 scores and average F1 scores were calculated and tabulated in Table 2.

BERT. The training and development set sizes varied depending on the dataset, but we had an average of 20,000 reports used in the training set and 10,000 in the dev set, as well as additional manually labeled data used for evaluation. Our train/dev split was achieved using Python library functions and we used a 70/30 split while excluding the manually labeled data.

For each experiment, we used a learning rate of 2e-5, batch size of 18, Adam Optimizer, Multi-class Cross Entropy Loss, and trained for 10 epochs (or until convergence). Training took an average time of 1.6 hours for each experiment.

For the CombinedBERT (Sarle + Regex) experiment, we combined both labelled CT abdomen datasets into one file and used the same method as above to split the training and dev sets. We excluded the 100 manually labelled reports from both datasets.

Lastly, we finetuned our models on additionally manually labeled data. For these experiments, we focused on CT Abdomen data and ran checkpoints from each of our models on a training set of 120 manually labeled entries and a test set of 1000 regex entries. We used a learning rate of 1.5e-5, batch size of 10, and trained for 10 epochs. These results are included in Table 3.

5.4 Results

Category	CheXpert	CheXbert	Sarle-Rules
Cardiomediastinum	0.989	0.989	1.0
Cardiomegaly	0.989	0.989	1.0
Lung Opacity	0.884	0.879	0.947
Lung Lesion	0.969	0.979	1.0
Edema	0.958	0.967	1.0
Consolidation	0.964	0.957	1.0
Pneumonia	0.826	0.832	0.963
Atelectasis	0.935	0.99	1.0
Pneumothorax	0.986	1.0	1.0
Pleural Effusion	0.950	0.967	0.968
Pleural Other	0.988	0.988	1.0
Fracture	1.0	1.0	1.0
Support Devices	0.972	0.991	1.0
No Finding	0.866	0.897	-
Average	0.948	0.959	0.991

Table 1: F1 scores of **baseline** models across all labels - naive CheXpert, CheXbert, and SarleRules

In our baseline results, we found that Sarle performed the best across all 14 categories. It is important to note, however, that Sarle only labels positive and negative indications, while CheXpert and

CheXbert also consider "possible" indications. The inclusion of this more challenging task of identifying the possibility of a condition may explain the lower F1 scores of the two models.

Dataset	Category	BERT	Regex	Sarle-Rules	CheXpert
XR chest	Pneumonia	0.930	0.981	0.963	0.826
	Pneumothorax	0.979	1.0	1.0	0.986
	Pleural Effusion	0.969	0.989	0.968	0.950
	Pulmonary edema	0.939	0.978	0.978	0.958
	Rib fracture	0.959	0.960	0.976	1.0
	Infection	0.828	0.979	0.979	0.921
	Aspiration	0.923	0.989	0.978	0.975
	Cardiomegaly	0.989	0.989	0.989	0.989
	Opacities	0.938	0.910	0.947	0.884
	Atelectasis	0.845	0.990	0.979	0.935
Average		0.930	0.977	0.976	0.942
CT head	Intracranial hemorrhage	0.664	0.960	1.000	-
	Subarachnoid hemorrhage	1.0	1.0	0.989	-
	Subdural hemorrhage	1.0	1.0	0.970	-
	Epidural hemorrhage	1.0	1.0	1.0	-
	Intraparenchymal hemorrhage	1.0	1.0	1.0	-
	Intraventricular hemorrhage	1.0	1.0	1.0	-
	Skull fracture	0.942	0.964	0.985	-
	Stroke	1.0	1.0	1.000	-
	Cerebral edema	0.970	0.973	0.988	-
	Diffuse axonal injury	1.0	1.0	1.0	-
Average		0.953	0.990	0.993	-
CT abdomen	Appendicitis	0.965	0.968	0.989	-
	Cholecystitis	0.940	0.977	0.989	-
	Abdominal Aortic Aneurysm	1.0	1.0	0.985	-
	Small bowel obstruction	0.919	0.962	0.978	-
	Pancreatitis	0.984	1.0	1.0	-
	Splenic laceration	1.0	1.0	1.0	-
	Liver laceration	1.0	1.0	1.0	-
	Colitis	0.980	1.0	0.990	-
	Pyelonephritis	1.0	1.0	0.989	-
	Nephrolithiasis	1.0	1.0	1.0	-
	Malignancy	0.923	0.981	0.890	-
Average		0.974	0.990	0.982	-
CT chest	Pneumonia	0.967	0.995	0.978	-
	Pneumothorax	0.964	0.980	0.978	-
	Pleural Effusion	0.990	0.990	0.916	-
	Pulmonary edema	0.970	0.970	0.960	-
	Rib fracture	0.937	0.924	0.909	-
	Pericardial effusion	0.992	0.992	0.992	-
	Aortic dissection	0.959	0.982	0.985	-
	Malignancy	0.957	0.967	0.945	-
Average		0.967	0.975	0.957	-

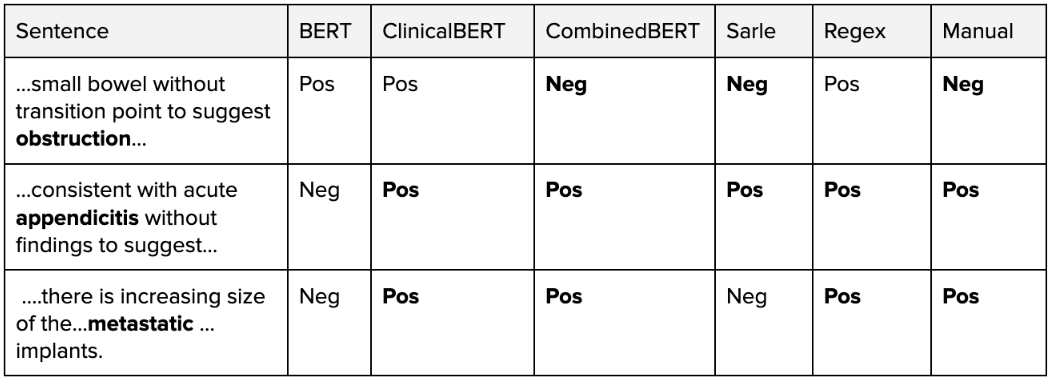
Table 2: F1 scores of **experimental** models across all labels

We did a series of experiments to distinguish between the different BERT models. Table 3 includes F1 scores compared to our manually labeled data, and Table 4 compares the models on their performance when evaluated on regex data. We also finetuned the BERT models with additional manually labeled data for CT Abdomen.

Pretrain Model	XR chest	CT head	CT abdomen	CT chest
BERT	0.930	0.953	0.974	0.967
CheXbert	0.930	0.953	0.972	0.968
ClinicalBERT	0.930	0.954	0.974	0.967
CombinedBERT (Sarle + Regex)	-	-	0.979	-
BERT (finetuned on manual data)	-	-	0.990	-
ClinicalBERT (finetuned on manual data)	-	-	0.993	-

Table 3: average F1 score of different BERT-pretrain regimes (refer to appendix for label F1 scores)

6 Analysis



Sentence	BERT	ClinicalBERT	CombinedBERT	Sarle	Regex	Manual
...small bowel without transition point to suggest **obstruction**...	Pos	Pos	**Neg**	**Neg**	Pos	**Neg**
...consistent with acute **appendicitis** without findings to suggest...	Neg	**Pos**	**Pos**	**Pos**	**Pos**	**Pos**
....there is increasing size of the...**metastatic** ... implants.	Neg	**Pos**	**Pos**	Neg	**Pos**	**Pos**

Figure 2: Qualitative examples of CT abdomen reports and associated labels from our models

6.1 Rules-Based Approaches

Rules-based approaches work well and have several advantages, because they don't require manually labelled data, and are much faster. However, there are edge cases so it is hard to get 100% accurate extraction. For example, for the report " ...mildly dilated **small bowel** without transition point to suggest **obstruction**..", Regex classifies "small bowel obstruction" as positive (it is actually negative). This example fails because it is less commonly seen phrase, so we didn't have this phrase pattern in Regex.

Besides the patterns, another issue that arose was finding synonyms of a specific label. For example, "malignancy" is a term that could be related to other cancer terms such as "metastasis." Rule based methods have no way of knowing that these terms are similar. Less commonly used phrases were not included in the CT abdomen and CT head labels for both Sarle-Rules and Regex, which resulted in lower accuracy. The scope of labels must also be decided by doctors. For example, our labels had a narrower scope ("skull fracture" instead of the original CheXpert and Sarle label "fracture" and "pulmonary edema" instead of "edema"). These can be easily tuned by adding or removing specific vocabulary and rules to our methods.

We labelled uncertain reports such as "findings are indeterminate for acute cholecystitis" as positive. This is a clear advantage of CheXpert and CheXbert over our labels, as the granularity could be helpful in diagnosis or to determine further scans for the patient. These types of "uncertain" cases would be extremely difficult to code in a purely rules-based method because the diversity of ways to express uncertainty is higher than that of negation (which might be why Sarle is only binary). However, we did not use CheXbert for our BERT models because it's F1 scores were low and it was missing very obvious phrases, such as labelling "without evidence of X" as a positive occurence (this may be a bug in their code). However, CheXpert did successfully label uncertain cases such as "may represent atelectasis" or "unchanged cardiomediastinal silhouette." Because of its low accuracy upon inspection, we generated labels for CT abdomen data and coded the customized labels for our other datasets but did not proceed with evaluating the F1 scores.

Despite these concerns, the new labels we created have high accuracy against our manual set, comparable to our baseline F1 scores.

6.2 Pre-trained BERT Models

Our BERT models trained on just Regex labels performed very well when evaluated against regex data, and as well as the regex labels when evaluated against manual data (see table in appendix for evaluations with regex). ClinicalBERT did especially well compared to other models, but all BERT models were comparable in quality. The better understanding of medical terminology in ClinicalBERT may allow it to label more accurately than BERT (second row in table) For example, in row 2 of Fig 2., BERT labels "appendicitis" as negative. It is possible that the BERT recognizes the "without" after the word and assumes it is negative, but with a better understanding of medical terminology in ClinicalBERT, it is able to understand context and underlying sentence structure better.

Our CombinedBERT model (Regex + Sarle) has comparable average accuracy to Sarle, higher accuracy than ClinicalBERT, and performs better than Regex or Sarle on some observations (but worse on others). For example, it is able to get the uncommon phrase not coded in Regex (1st row in Fig 2.) while it identifies a synonym for the label "maligancy" which was not coded in Sarle (3rd row in Fig 2.). Regex and Sarle contain different rules, so they can disagree on less common negation phrases. Thus using both auto-labels as training data for BERT can allow it to synthesize the two and can lead to more accurate labels. Since auto-labelling is fast and requires no medical knowledge, this method can be easily applied to increase accuracy of models when there is no access or ability to manually label data. Finally, we fine-tuned our models on very little manual data (120 reports), which increased accuracy to above that of Sarle and Regex, making it the best model for CT abdomen. This strategy of auto-labelling most of the data, then adding a small amount of manually labelled data, can save costs for creating new, accurate models.

7 Conclusion

In this study, we leverage NLP methods for extracting clinical labels from radiology reports. We utilize the method proposed in Smit et al. (2020) to fine-tune several pre-trained BERT models on labels generated from rules-based labelers. We experiment with 3 different rules-based labelers: Regex, Sarle, and CheXpert. Regex outperforms the other labelers and exhibits an average F1 score in the range of 0.97-0.99 on our datasets. Therefore, we use regex to train our BERT models. Our best BERT model, ClinicalBERT, achieves the highest F1 score across all our 4 datasets. We then trained CombinedBERT using Sarle and Regex, which achieves the highest F1 score of 0.979 on the CT abdomen dataset, and we also finetuned our models on manually labelled data, achieving an F1 score of 0.993, higher than other Bert-based or Rules-based methods. Our models can be used to extract important labels from radiology report datasets relevant to how critical the condition of a patient is. Improving these models can help accelerate the analysis of ED reports and lead to life saving improvements in the medical system.

One of the main limitations of our models is that only 100 reports were manually labelled for evaluation. For example, the CheXbert paper labelled 500 reports to accurately evaluate model performance, as some labels appear very sparsely. Another issue with the BERT models is that the Regex labels used for training are not a gold standard. To address this limitation, we finetuned our BERT models on manually labeled data, but our manual labels are sparse.

For future work, we plan to run CombinedBERT on our additional datasets, namely XR Chest, CT Head, and CT Chest, to evaluate its performance and generalizability compared to the other BERT models. We also plan to obtain more gold-standard manual labels and use them to fine-tune the current models that were initially trained on regex labels. This will provide more precise input to the models than relying solely on regex labels, potentially resulting in improved performance in accurately identifying medical observations.

References

Emily Alsentzer, John Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *arXiv.org*.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. 2019. Padchest: A large chest x-ray image dataset with multi-label annotated reports. In *arXiv.org*.

Wendy Chapman, Will Bridewell, Paul Hanbury, Gregory Cooper, and Bruce Buchanan. 2002. A simple algorithm for identifying negated findings and diseases in discharge summaries. In *Journal of Biomedical Informatics*.

Ignat Drozdov, Daniel Forbes, Benjamin Szubert, Mark Hall, Chris Carlin, and David Lowe. 2020. Supervised and unsupervised language modelling in chest x-ray radiological reports. In *PLOS One*.

Jeremy Irvin, Pranav Rajpurka, and Michael Ko. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *arXiv.org*.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2017. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. In *arXiv.org*.

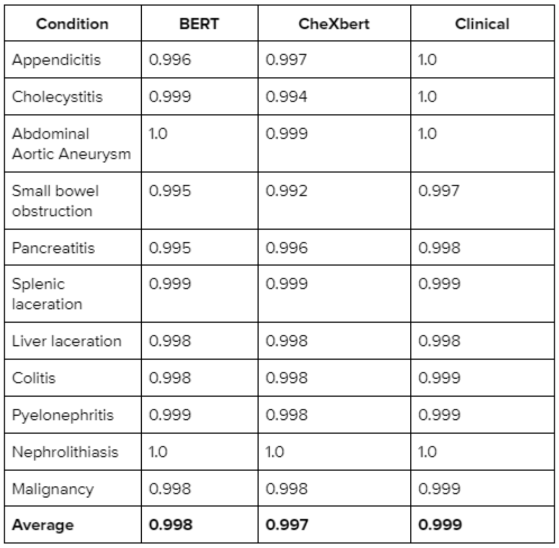
Maciej Mazurowski Joseph Lo Ricardo Henao Geoffrey D Rubin Lawrence Carin. Rachel Draelos, David Dov. 2021. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. In *Medical Image Analysis*.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. 2020. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *arXiv.org*.

David Wood and Thomas Lynch, Jeremy ... Booth. 2020. Automated labelling using an attention model for radiology reports of mri scans (alarm). In *arXiv.org*.

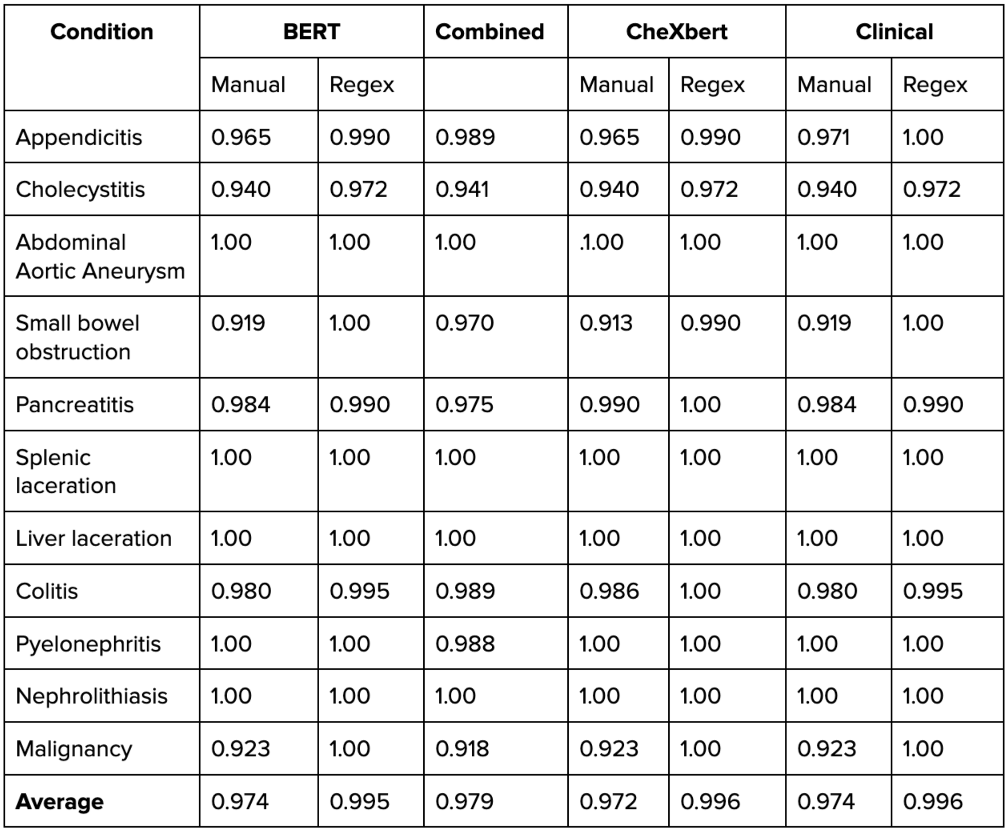
A Appendix

In this section, we include accuracy scores for each of the three models when evaluated with manual labels and with regex labels. We see that they perform very well when evaluated with regex labels and comparable to Regex when evaluated with manual labels. We also include an additional evaluation of the models with a large amount of Regex data for CT Abdomen, where ClinicalBERT demonstrates a strong performance.



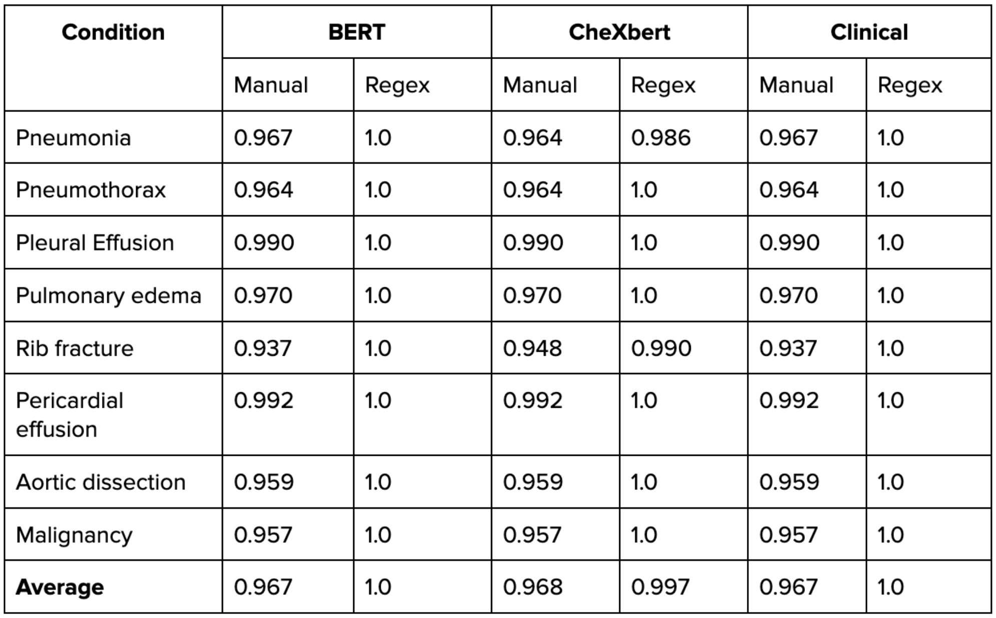
Condition	BERT	CheXbert	Clinical
Appendicitis	0.996	0.997	1.0
Cholecystitis	0.999	0.994	1.0
Abdominal Aortic Aneurysm	1.0	0.999	1.0
Small bowel obstruction	0.995	0.992	0.997
Pancreatitis	0.995	0.996	0.998
Splenic laceration	0.999	0.999	0.999
Liver laceration	0.998	0.998	0.998
Colitis	0.998	0.998	0.999
Pyelonephritis	0.999	0.998	0.999
Nephrolithiasis	1.0	1.0	1.0
Malignancy	0.998	0.998	0.999
Average	**0.998**	**0.997**	**0.999**

Figure 3: CT abdomen F1 scores for all BERT models against Regex labels



Condition	BERT		Combined	CheXbert		Clinical	
	Manual	Regex		Manual	Regex	Manual	Regex
Appendicitis	0.965	0.990	0.989	0.965	0.990	0.971	1.00
Cholecystitis	0.940	0.972	0.941	0.940	0.972	0.940	0.972
Abdominal Aortic Aneurysm	1.00	1.00	1.00	.1.00	1.00	1.00	1.00
Small bowel obstruction	0.919	1.00	0.970	0.913	0.990	0.919	1.00
Pancreatitis	0.984	0.990	0.975	0.990	1.00	0.984	0.990
Splenic laceration	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Liver laceration	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Colitis	0.980	0.995	0.989	0.986	1.00	0.980	0.995
Pyelonephritis	1.00	1.00	0.988	1.00	1.00	1.00	1.00
Nephrolithiasis	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Malignancy	0.923	1.00	0.918	0.923	1.00	0.923	1.00
Average	0.974	0.995	0.979	0.972	0.996	0.974	0.996

Figure 4: CT abdomen F1 scores for all BERT models against Manual and Regex labels



Condition	BERT		CheXbert		Clinical	
	Manual	Regex	Manual	Regex	Manual	Regex
Pneumonia	0.967	1.0	0.964	0.986	0.967	1.0
Pneumothorax	0.964	1.0	0.964	1.0	0.964	1.0
Pleural Effusion	0.990	1.0	0.990	1.0	0.990	1.0
Pulmonary edema	0.970	1.0	0.970	1.0	0.970	1.0
Rib fracture	0.937	1.0	0.948	0.990	0.937	1.0
Pericardial effusion	0.992	1.0	0.992	1.0	0.992	1.0
Aortic dissection	0.959	1.0	0.959	1.0	0.959	1.0
Malignancy	0.957	1.0	0.957	1.0	0.957	1.0
Average	0.967	1.0	0.968	0.997	0.967	1.0

Figure 5: CT chest F1 scores for all BERT models against Manual and Regex labels

Condition	BERT		CheXbert		Clinical	
	Manual	Regex	Manual	Regex	Manual	Regex
Intracranial hemorrhage	0.664	1.0	0.664	1.0	0.664	1.0
Subarachnoid hemorrhage	1.0	1.0	1.0	1.0	1.0	1.0
Subdural hemorrhage	1.0	1.0	1.0	1.0	1.0	1.0
Epidural hemorrhage	1.0	1.0	1.0	1.0	1.0	1.0
Intraparenchymal hemorrhage	1.0	1.0	1.0	1.0	1.0	1.0
Skull fracture	0.942	0.985	0.942	0.985	0.947	0.975
Stroke	1.0	1.0	1.0	1.0	1.0	1.0
Cerebral edema	0.970	0.985	0.970	0.985	0.975	1.0
Diffuse axonal injury	1.0	1.0	1.0	1.0	1.0	1.0
Average	0.953	0.997	0.953	0.997	0.954	0.997

Figure 6: CT head F1 scores for all BERT models against Manual and Regex labels

Condition	BERT		CheXbert		Clinical	
	Manual	Regex	Manual	Regex	Manual	Regex
Pneumonia	0.930	1.0	0.930	1.0	0.930	1.0
Pneumothorax	0.979	1.0	0.979	1.0	0.979	1.0
Pleural Effusion	0.969	1.0	0.969	1.0	0.969	1.0
pulmonary edema	0.936	1.0	0.936	1.0	0.936	1.0
rib fracture	0.959	1.0	0.959	1.0	0.959	1.0
infection	0.828	0.988	0.828	0.988	0.828	0.988
aspiration	0.923	1.0	0.923	1.0	0.923	1.0
cardiomegaly	0.989	1.0	0.989	1.0	0.989	1.0
opacity	0.938	1.0	0.938	1.0	0.938	1.0
atelectasis	0.845	1.0	0.845	1.0	0.845	1.0
Average	0.930	1.0	0.930	1.0	0.930	1.0

Figure 7: XR chest F1 scores for all BERT models against Manual and Regex labels