Labeling Chest X-Ray Reports with Markers of Longitudinal Change

Stanford CS224N Custom Project
Collaborator: Pranav Rajpurkar
Mentor: Gaurab Banerjee

Ryan Han
Department of Computer Science
Stanford University
ryanhan@stanford.edu

Christina Kwak
Department of Computer Science
Stanford University
kwakc@stanford.edu

Evan Saracay
Department of Computer Science
Stanford University
esaracay@stanford.edu

Abstract

Automatic label extraction from radiology text reports has enabled the construction of large, labeled datasets for training medical imaging models. However, existing approaches to report labeling focus on detecting the presence of disease at a single time point and do not capture longitudinal changes in disease progression. In this work, we introduce a BERT-based method for incorporating rule-based labels and a small set of manual annotations to label free text radiology reports as containing indications of disease progression, disease stability, or uncertain. We find that our best method, which includes iterative distillation, outperforms both the only existing rule-based labeler as well as simple BERT fine tuned only on manual annotations (AUROC: 0.861 vs 0.826 for simple BERT and 0.548 for sentence matching). Additionally, we demonstrate that our method produces the highest performance gains when manual annotations are scarce (AUROC at 50 samples: 0.815 vs 0.701 for simple BERT). Our method can be used to train computer vision models to monitor disease progression and thus has the potential to reduce patient harm and physician burnout. Additionally, our distillation approach is not specific to our labeling task, and can be built upon in future works to increase performance in other medical labeling tasks where reliable annotations are scarce.

1 Introduction

Chest X-rays (CXR) are the most common imaging examination and are critical to the screening, diagnosis, and management of a wide variety of medical conditions. Recently, the use of NLP to extract labels from radiology text reports has enabled the large-scale training of deep learning models for clinical applications such as detecting the presence of pneumonia or lung cancer [1, 2]. However, in contrast to these tasks which focus on the presence of disease at a single time point, clinical care represents a dynamic scenario where multiple time points are often compared in order to make a diagnostic or prognostic assessment. Extracting labels relating to longitudinal change from radiology text reports would thus enable the training of AI systems that facilitate tedious and time-consuming comparisons performed by radiologists.

Despite this clinical significance, relatively little effort has been made towards characterizing change in imaging datasets. Publicly available datasets such as MIMIC-CXR and CheXpert which were labeled using NLP do not contain labels pertaining to longitudinal change and, to the best of our

knowledge, the only existing work that focuses on longitudinal change in CXRs uses a rigid text matching approach to match frequent sentences pertaining to disease progression [3, 4, 5]. Previous works have demonstrated that pretrained BERT models are able to achieve greater performance and generalizability than rule-based radiology reports labelers [6, 7]. In this project, we explore how rule-based labelers and a small set of manual annotations can be combined with pretrained BERT models to extract labels for longitudinal change from free text radiology reports.

Specifically, we formulate the report labeling task as a multi-class classification problem where the input is a free-text radiology report and the output classes are disease progression (indication of deterioration or improvement), disease stability (indication of no clinically relevant change), and uncertain (no indication). Our approach begins with training a pretrained BERT model [8] on a small corpus of manual annotations followed by a randomly sampled corpus of weak (i.e. rule-based) labels, and then fine-tuning the model again on the manual annotations. We then use our trained model as a weak labeler and iteratively distill our model by using the same process to train another pretrained BERT model on the improved weak labels.

On a large publicly available database, MIMIC-CXR, we find that our method improves upon the only existing rule-based labeler as well as a simple BERT trained on manual annotations (AUROC: 0.861 vs 0.826 for simple BERT and 0.548 for sentence matching). Additionally, we demonstrate that our method produces the highest performance gains when manual annotations are scarce (AUROC at 50 samples: 0.815 vs 0.701 for simple BERT).

To the best of our knowledge, our method represents the first deep learning based natural language processing system for extracting labels for longitudinal change from free text radiology reports. The labels generated by our method can be used to train medical imaging models for monitoring disease progression and thus have the potential to reduce image read turnaround time, patient harm, and physician burnout. Additionally, our distillation approach is not specific to our labeling task, and can be built upon in future works to increase performance in other medical labeling tasks where reliable annotations are scarce.

2 Related Work

Label Extraction from Radiology Reports. There is a rich history of automated label extraction systems for free-text radiology reports ranging from intricate rule-based systems to systems that incorporate deep learning [9, 10, 11]. CheXpert [4] is one of the most common rule-based labelers for CXR and has been incorporated into deep learning based labelers such as CheXbert [6] and CheXpert++ [7]. Both of these papers find that training a pretrained BERT model on the output of CheXpert improves performance. However, while CheXbert explores the effect of including biomedical text during model pretraining, to the best of our knowledge no work has attempted to investigate whether this performance benefit is from pretrained language understanding or simply from training a neural network on a rule-based system.

Medical Report Labeling Our work is similar in motivation to approaches to reduce the number of annotations required for training medical report labelers [12, 13]. Weak labels generated through data programming have seen notable success in this field, such as [14] who used data programming in consultation with a clinician to label CT reports for intracranial hemorrhage in a limited time frame [14, 15]. In contrast to these methods which focus on the incorporation of expert knowledge into rule-based labelers, we present a generalized approach to take advantage of existing rule-based labelers and available manual annotations and demonstrate its value on the task of labeling longitudinal changes in CXR reports. Our method can thus be built upon by applying our distillation framework to other medical report labelers and tasks.

Distillation Knowledge distillation refers to the process of transferring knowledge from a teacher model to a student model [16]. Self-distillation, a special case of distillation where the student and teacher model are the same or vary in number of parameters, has been shown to be an effective method for improving performance of supervised computer vision and natural language processing models [17, 18]. Distillation has also been applied in the semi-supervised learning literature to leverage both strong and weak labels for natural images [19]. However, there is limited work applying semi-supervised distillation in the medical domain where the amount of reliable labels is often the scarcest [20, 21].

3 Approach

Baselines Our work builds off the rule-based labeler proposed in [5]. As the code and data for this approach are not publicly available, we implemented this approach from scratch based on the description released by the authors. This involved decomposing all training reports into sentences, counting sentence frequency (ignoring minor variations in spacing and punctuation), and manually labeling sentences that appeared more than 200 times. Out of these 523 frequent sentences, we found that 7 referenced disease progression and 39 referenced a lack of clinically significant change. These 46 labeled sentences appeared in 9.85% of the training data. Any report that contained a sentence referencing disease progression or a lack of clinically significant change was labeled as such, and reports that did not contain either of these types of sentences were labeled as uncertain.

Our second baseline is original yet builds off the first. From the 46 frequent sentences referencing longitudinal change, we manually extracted phrases responsible for the labeling of the sentences. This led to 14 phrases associated with disease progression and 4 phrases associated with lack of clinically significant change. Examples of such phrases are "improved", "no change", and "remain", and these phrases appeared in 57.04% of the training data reports. Reports were labeled using the same method as the first baseline.

Main Approach. Given a radiology report, we prepend a [CLS] token and encode the report with a pretrained BERT encoder before feeding the [CLS] embedding into a linear classification head with three outputs. These outputs are then passed through a softmax layer before applying categorical cross-entropy loss.

Following [6], we investigate various supervision strategies for training our BERT models. In *BERT-phm*, we train our model with weak labels generated by running our phrase matching baseline over the entire training set. In *BERT-man*, we train our model using strong supervision on a subset of manually labeled samples. We additionally investigate how we can combine these strategies in various orderings (*BERT-phm-man*, *BERT-man-phm*, *BERT-man-phm-man*) to leverage both our large weakly labeled dataset and our small subset of manual strong labels.

Another supervision technique we explored involved using a trained BERT model (e.g. *BERT-man*) to generate weak labels on the entire training dataset. We then train a default pretrained BERT model on the weakly labeled training dataset before fine tuning it on the manually labeled subset. This distillation process can then be repeated iteratively. This was motivated by the expectation that our best BERT-based method would be a better weak labeler than our phrase matching baseline.

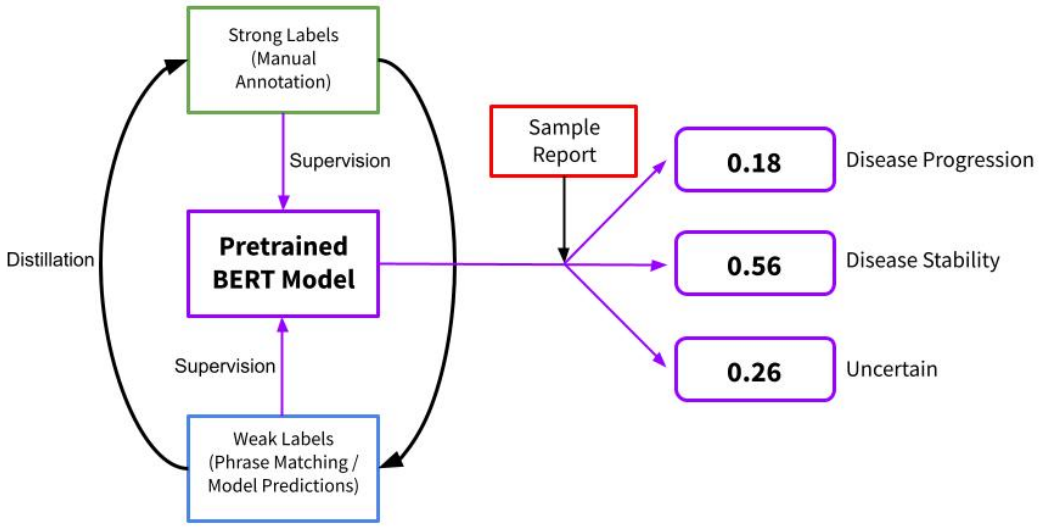

Figure 1: Distillation pipeline. BERT is trained on a combination of rule-based and manual labels and then iteratively distilled by using model predictions as weak labels to train another BERT.

4 Experiments

Data. We use the MIMIC-CXR dataset which consists of 337,110 chest x-ray images from 227,827 studies [3]. Each study has a single associated report, and we only use the 227,827 free text radiology reports (one per study) for our work. We preprocess the data by extracting and concatenating the "Findings" and "Impression" section from each report.

As with many medical imaging datasets, our ground truth was determined by manual annotation by human readers, with conflicts being determined by committee consensus after discussion [4, 6, 22, 23]. One notable difference is that our human readers were not clinicians but team members, which was due to both a lack of access to clinicians as well as due to the fact that our task formulation does not require clinical expertise (detecting indications of change vs. making a diagnosis).

A random subset of 250 patients (corresponding to 1000 reports) was selected for manual annotation by two team members. This subset was then split into a fine-tuning dataset of 399 reports (100 patients), a validation dataset of 90 reports (25 patients), and a test dataset of 510 reports (125 patients). All splits were created at the patient level in order to avoid information leakage, and the remaining 226,827 reports were used for training with weak labels. Our test dataset size is similar to that of CheXpert and CheXbert which use 500 and 687 manually annotated reports for evaluation respectively [4, 6]. We found that the class prevalence of our ground truth annotations were roughly comparable, with 194 reports indicating disease progression, 177 reports indicating disease stability, and 140 uncertain reports.

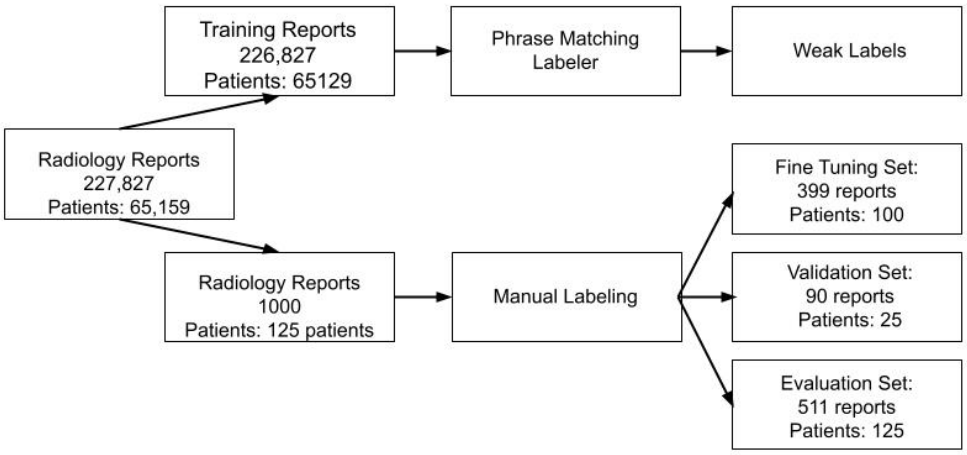

Figure 2: Data labeling pipeline. All splits were randomly generated at the patient level.

Evaluation method. We evaluate all methods using AUROC on the manually labeled test dataset. We determined this metric to be appropriate as our data are not heavily imbalanced and we are primarily interested in the ranking capabilities of our model (probabilities do not need to be calibrated). As our task is a multi-class classification problem, we calculate AUROC in a one-vs-rest manner and weight per-class AUROC by the support of each class when combining. We include 95% confidence intervals using the nonparametric bootstrap method with 1000 samples for our final results [24].

Experimental details. We fine-tune all layers of our model using Adam with a learning rate of $2 * 10^{-}5$ (as used in [8] for fine-tuning). Models were trained for 10 epochs on the manual fine-tuning set and for 1 epoch when training on weak labels for the entire training corpus (we found that performance did not significantly improve after 1 epoch, so we limited this due to computational constraints). Model selection was performed using validation loss, and training was done on a NVIDIA V100 GPU with a batch size of 18.

Results. We report the performance of the baselines and our models without distillation in Figure 3. We then investigate the effect of pretraining corpus in Table 1 and the effect of distillation in Figure 4. Figure 5 explores the performance of methods at varying numbers of manual fine-tuning samples.

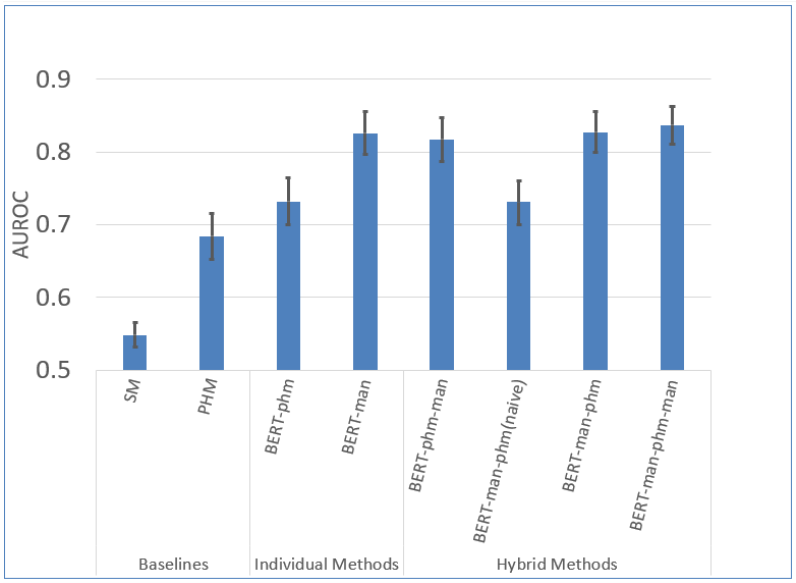

Figure 3: Effect of supervision strategy. All BERT models were fine-tuned from a standard pretrained BERT. Error bars indicate 95% confidence interval. Abbreviations: sentence matching (SM), phrase matching (PHM)

4.1 Supervision Strategy

As shown in Figure 3, we find that our proposed phrase matching baseline significantly outperforms the sentence matching baseline (0.684 vs 0.548). Indeed, the performance of the sentence matching baseline is near random, which makes sense in light of the fact that frequent sentences were only matched in 9.85% of all training reports. This is significantly different than the results reported by the authors of [5], who found that frequent sentence matching captured 77% of all reports in their private dataset. While it is diffcult to ascertain the cause of this difference since the code and data of the original paper are not made available, it seems likely that this is due to a difference in report style between datasets (frequent sentence examples in [5] are significantly shorter than frequent sentences in MIMIC-CXR). In contrast, we find that our phrase matching baseline matches 57.04% of all training reports, which is reflected in its higher performance. Predicted class distributions on the evaluation set for our baseline methods can be found in Appendix A.

We find that *BERT-man* achieves a statistically significant increase in AUROC over *BERT-phm* (0.826 vs 0.732). These results are notable as they indicate that strong supervision is preferable to weak supervision using the best rule-based labeler for our task, even with a several order of magnitude difference in training set size. Additionally, *BERT-phm* achieves a sizable yet not statistically significant increase in AUROC compared to the phrase matching baseline (0.732 vs 0.684). We hypothesize that this benefit is due to BERT's pretrained language understanding rather than simply due to training a neural network. The effect of pretraining is further explored in Table 1.

Among the methods in Figure 3 that incorporate both weak and strong supervision, we find that none of these hybrid methods lead to notable performance increases relative to *BERT-man*. *BERT-man-phm-man* was the only method that outperformed *BERT-man* (0.837 vs 0.829) but this difference was not statistically significant and did not lead to further performance increases when we continued to alternate training between strong and weak labels. Notably, *BERT-man-phm(naive)* had nearly the same performance as *BERT-phm* (0.731 vs 0.732). We rationalize as the model forgetting all knowledge learned from the manual labels due to the vast difference in training set size between manual and phrase-matching labels. Thus, for the rest of the *BERT-man-phm*-based methods, we matched the phrase-matching dataset size to the number of manual labels available.

4.2 Biomedical Language Representations

Based on the superior performance of *BERT-phm* to our phrase matching baseline, we hypothesized that BERT's pretrained language understanding enables it to label reports more accurately than rule-based methods. To investigate this, we fine-tuned BERT on manual annotations from random initialization as well as from checkpoints pretrained on different biomedical text datasets.

Pretraining method	**AUROC (95% CI)**
General pretraining	
Random Initialization	0.682 (0.650, 0.714)
BERT [8]	0.826 (0.797, 0.853)
Biomedical pretraining	
BioBERT [25]	0.831 (0.803, 0.858)
ClinicalBioBERT [26]	0.814 (0.785, 0.842)
BlueBERT [27]	0.807 (0.778, 0.835)

Table 1: Effect of pretraining strategy. All BERT models were trained using the manual fine-tuning set. All pretrained models had a statistically significant improvement over random initialization.

As shown in Table 1, all pretraining strategies for BERT perform comparably and significantly better than random initialization. Notably, BERT fine-tuned from random initialization performs comparably to the phrase-matching baseline, which confirms our hypothesis that the performance benefit of BERT derives from pretrained language understanding rather than simply due to training a neural network. The fact that BERT pretrained on various biomedical text databases did not perform significantly better than default pretraining is surprising as [6] find that BlueBERT performs significantly better than default pretraining for their task. One explanation of this could be that labeling indications of longitudinal change does not require as many domain-specific concepts as labeling specific diagnoses. However, we leave the exploration of this hypothesis to future work.

4.3 Distillation

We explore the effect of distillation on our best performing method which is *BERT-man-phm-man*. As shown in Figure 4A, each iteration of distillation leads to a non-inferior model with incremental increases in AUROC over the iteration space we explored (limited due to time and compute). After five iterations of distillation, our distilled model has a sizeable but not statistically significant increase in AUROC relative to *BERT-man* (0.861 vs 0.826).

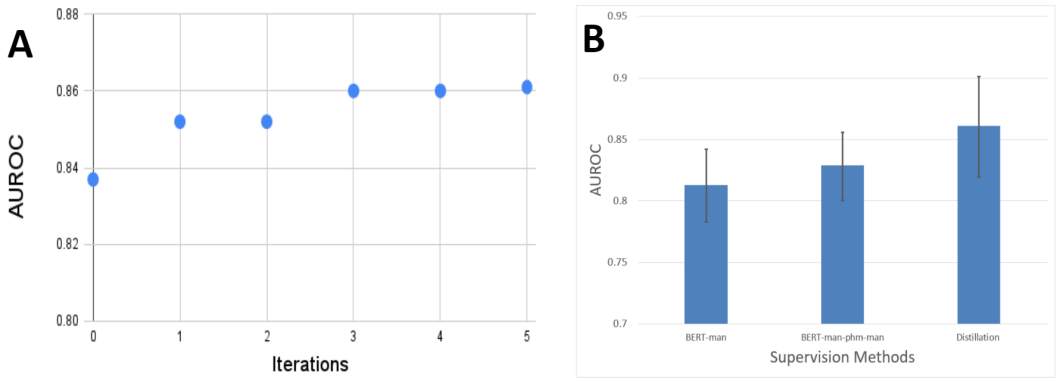

Figure 4: Effect of Distillation. **A:** AUROC of distilled *BERT-man-phm-man* across iterations.**B:** Comparison of model performance before and after distillation.

4.4 Fine Tuning Set Size Reduction

Our final analysis involved investigating the performance benefit of distillation at different numbers of fine tuning labels. In Figure 5, we see that while the performance of *BERT-man* is roughly linear

with respect to number of training samples, the performance of *BERT-man-phm-man (distilled)* is less dependent on number of training samples. This translates to distillation having a greater performance benefit in lower label settings and, indeed, the difference between *BERT-man* and *BERT-man-phm-man (distilled)* at 50 samples is statistically significant (0.701 vs 0.815). Additionally, we highlight the fact that the performance of *BERT-man-phm-man (distilled)* at 50 samples is comparable to that of *BERT-man* at 399 samples (0.815 vs 0.827), despite having 8 times less labels. This demonstrates the promise of distillation in settings when manual labeled samples are scarce (e.g. clinical settings). Continuing to observe the limits to which our approach can be applied presents an interesting avenue for future work.

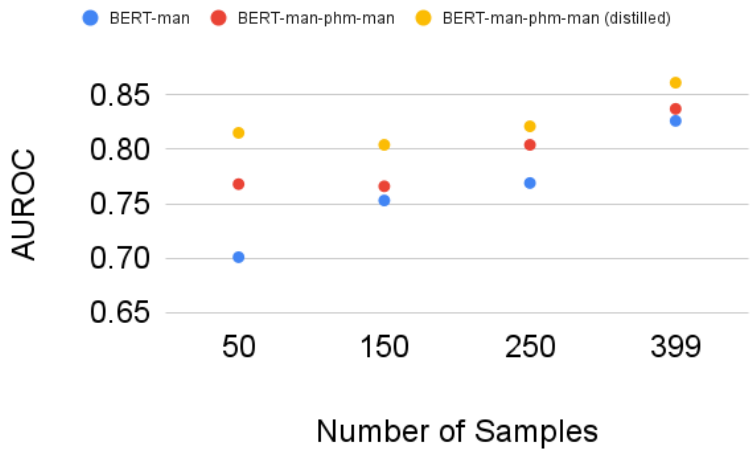

Figure 5: Effect of training set size. Difference between BERT-man-phm-man (distilled) and BERT-man at 50 samples is statistically significant.

5 Analysis

No.	Example	Label	Phrase matching	BERT man-phm-man (distilled)
1	there is a left-sided port-a-cath with the distal lead tip at the cavoatrial junction. there has been **decrease** in the right-sided pleural effusion. left-sided pleural effusion is again seen. no pneumothoraces are seen. there is mild prominence of the pulmonary interstitial markings stable	change	change	change
2	a **newly placed** feeding tube is coiled in the pharynx. the tube does not reach the esophagus. tube reposition is required	change	uncertain	change
3	as compared to the previous radiograph the pre-existing left pleural effusion **has been drained** with a pigtail catheter. the pigtail catheter is in place. the effusion has almost completely resolved. there is no safe evidence of left pneumothorax. **otherwise** the radiograph is **unchanged**. moderate cardiomegaly	change	no change	change
4	the dobbhoff tube is in the stomach. there **continues** to be retrocardiac opacity consistent volume loss infiltrate effusion. right ij cordis is **again visualized**. aeration in the right lung is **improved**	change	uncertain	uncertain

Figure 6: Qualitative analysis of selected failure cases.

In 6, we show a few select examples that highlight how our model works. We chose to compare only the strong label, weak label (using PHM), and the BERT man-phm-man-distilled, which was our best performer.

In Example 1 we see that all three results matched, which show that our models can recognize clear indications of change. PHM used the phrase "decrease" in order to label the report, and it seems like BERT man-phm-man-distilled was able to pick up on the meaning of this phrase. Given that this phrase appears in our most common phrase list, BERT has plenty of opportunities to learn that this

phrase indicates change from the PHM weak labels.

Example 2 and 3 show times when our weak labels were incorrect, meaning that phrase matching was insufficient in understanding and identifying change in the report. The phrase "newly placed" in example 2 points to change, however, this was not a common phrase found in our list, and it was thus unable to identify the report as having a sign of change. However, our BERT man-phm-man-distilled model was able to identify this change correctly. This indicates that our model is capable of learning some general concept of change without explicitly being told that a phrase corresponds to change. In example 3, we see that the phrase "has been drained", which is not on our common phrase list, is something that would indicate change, so this caused our PHM to fail. However, PHM is able to match the phrase "unchanged" which comes up later in the report which leads to a false no change. BERT man-phm-man-distilled is not constrained to the first common phrase that it sees, so unlike PHM, it is able to more holistically analyze the report and pick up on indications of change.

Example 4 shows another failure of PHM as it picks up on phrases like "continues" and "again visualized" and fails to get to the end of the report where change is indicated. However, for this example, BERT man-phm-man-distilled also fails, perhaps due to multiple conflicting phrases within the report. This means that the trained BERT model failed to learn that "improved" is a clear indication of change that trumps the other conflicting signs in the report.

6 Conclusion

In this study we leverage NLP methods for extracting longitudinal change from free-text radiology reports. We utilize the method proposed in [6] to combine both rule-based labelers and a small manually annotated set to train a BERT model that far outperforms its rule-based labeler, motivated by [5]. Our best performer (distillation) achieves an AUROC of 0.826, while our phrase matching baseline achieves an AUROC of 0.684. Our model can be used to generate large labeled sets for training medical imaging models. With the worldwide shortage of radiologists, improving these type of imaging models can help accelerate the analysis of CXR's and lead to life saving improvements in the medical system.

We also demonstrate that our distillation method, motivated by [16][17][18], provides a process for effectively training BERT models on tasks with scarce manual annotations. As seen in Section 4.4, with small sets of strong labels, our distillation model far outperforms BERT-man which is a BERT model that has been fine tuned on manual annotations. Our method for distillation presents an approach that can be used in conjunction with other contexts such as few-shot learning, which we have shown to be successful. This is especially useful in the medical domain where labeling large data sets is simply not an option given that clinician time is extremely valuable and such datasets may not exist.

Future work includes exploring the bounds at which distillation no longer provides incremental increases in AUROC. The results of Section 4.3 and 4.4, using distillation on our best performing model with smaller sample sets, also show promising avenues for future research that can possibly increase the performance of our best model. Motivated by our qualitative analysis, another possible avenue for exploration could be using sentence level embeddings and aggregating them to come up with report level predictions, as this would potentially facilitate the model learning to deal with conflicting phrases.

References

[1] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Yi Ding, Aarti Bagul, Curtis P. Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225, 2017.

[2] Fatemeh Homayounieh, Subba Digumarthy, Shadi Ebrahimian, Johannes Rueckel, Boj Friedrich Hoppe, Bastian Oliver Sabel, Sailesh Conjeti, Karsten Ridder, Markus Sistermanns, Lei Wang, Alexander Preuhs, Florin Ghesu, Awais Mansoor, Mateen Moghbel, Ariel Botwin, Ramandeep Singh, Samuel Cartmell, John Patti, Christian Huemmer, Andreas Fieselmann, Clemens Joerger,

Negar Mirshahzadeh, Victorine Muse, and Mannudeep Kalra. An Artificial Intelligence–Based Chest X-ray Model on Human Nodule Detection Accuracy From a Multicenter Study. *JAMA Network Open*, 4(12):e2141096–e2141096, 12 2021.

[3] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019.

[4] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019.

[5] Dong Yul Oh, Jihang Kim, and Kyong Joon Lee. Longitudinal change detection on chest x-rays using geometric correlation maps. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 748–756, Cham, 2019. Springer International Publishing.

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

[7] Matthew B. A. McDermott, Tzu Ming Harry Hsu, Wei-Hung Weng, Marzyeh Ghassemi, and Peter Szolovits. Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output, 2020.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[9] Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald M. Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *CoRR*, abs/1712.05898, 2017.

[10] Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, 2001.

[11] Choi HA Cartwright WB 4th Hinds PS Chamberlain JM Yadav K, Sarioglu E. Automated outcome classification of computed tomography imaging reports for pediatric traumatic brain injury. *Acad Emerg Med*, 2016.

[12] Christopher Re ´ James I. Huddleston Nicholas J. Giori Scott Delp Alison Callahan, Jason A. Fries and Nigam H. Shah. Medical device surveillance with electronic health records. *npj DigitalMedicine*, 2019.

[13] Matthew P. Lungren Imon Banerjee, Matthew C. Chen and journal = Journal of Biomedical Informatics year = 2018 Daniel L. Rubin, title = Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort.

[14] Khaled Saab, Jared Dunnmon, Roger Goldman, Alex Ratner, Hersh Sagreiya, Christopher Ré, and Daniel Rubin. Doubly weak supervision of deep learning models for head ct. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 811–819, Cham, 2019. Springer International Publishing.

[15] Jared Dunnmon, Alexander Ratner, Nishith Khandwala, Khaled Saab, Matthew Markert, Hersh Sagreiya, Roger E. Goldman, Christopher Lee-Messer, Matthew P. Lungren, Daniel L. Rubin, and Christopher Ré. Cross-modal data programming enables rapid medical machine learning. *CoRR*, abs/1903.11101, 2019.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv*, 2015.

[17] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. *CoRR*, abs/1905.08094, 2019.

[18] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[19] Qizhe Xie, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Self-training with noisy student improves imagenet classification. *CoRR*, abs/1911.04252, 2019.

[20] Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. 2022.

[21] Huisi Wu, Jiasheng Liu, Fangyan Xiao, Zhenkun Wen, Lan Cheng, and Jing Qin. Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion. *Medical Image Analysis*, 78:102397, 2022.

[22] Rajpurkar P. Haghpanahi M. et al. Hannun, A.Y. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, (25):65–69, 2019.

[23] M. et al. Gavrielides. Clinical Decision Support for Ovarian Carcinoma Subtype Classification: A Pilot Observer Study With Pathology Trainees. *Archives of Pathology Laboratory Medicine*, (144):869–877, 2020.

[24] B. Efron and R. Tibshirani. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, 1(1):54 – 75, 1986.

[25] 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*, 36(4):1234–1240, 2020.

[26] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical BERT embeddings. *CoRR*, abs/1904.03323, 2019.

[27] 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 preprint arXiv:1906.05474*, 2019.

A Appendix

Set	Change	No Change	No Indication
Validation	31	34	25
Fine Tuning	132	97	73

Table 2: **Class Distributions of Manually Labeled Sets**

Class	Change	No Change	No Indication
Sentence matching	5	53	452
Phrase matching	150	146	214

Table 3: **Class Distribution of Baseline Methods on Evaluation Set**