Accelerated Data Curation of Colitis Cases

Protiva Rahman, Ph.D., Cheng Ye, Ph.D., Kate Mittendorf, Ph.D., Michele LeNoue-Newton, Ph.D., Christine Micheel, Ph.D., Daniel Fabbri, Ph.D.
Vanderbilt University Medical Center, Nashville, Tennessee

Introduction

While immune checkpoint inhibitors (CPI) have improved cancer care, one of their main adverse events is CPI-induced colitis. Predicting CPI-induced colitis in patients can inform therapy selection. However, before predictive modeling, the data need to be curated from electronic health records (EHRs) since colitis does not have clear diagnosis codes and it can be documented in a variety of ways (proctocolitis, CPI-associated diarrhea, etc.). Curating positive colitis cases is an onerous task -- keyword search identifies over 200,000 notes which need to be manually reviewed for accuracy before they are imported into the database for more extensive expert curation of colitis episodes. In this work, we build a model to accurately identify colitis positive notes for review. For notes that only mentioned colitis symptoms, our deep learning pipeline reduced the number of reviewed symptom notes by 75% and had a precision of 84%. For colitis mention notes, our algorithm had an overall precision of 92% and reduced the number of notes from 128,314 notes to 8,170, indicating a 93.4% reduction in note review burden.

Methods

The goal of the colitis curation task is to identify EHR notes which are positive for colitis or one of the symptoms of colitis, i.e., diarrhea or bloody stool. Prior to building extraction models, curators manually reviewed 23,313 notes for 703 patients using keyword search. Of these, 1,994 notes were positive for colitis within the diagnostic differential, 3,906 were positive for presence of diarrhea, and 548 were positive for presence of bloody stool. Curators also highlighted the part of the note that was relevant in their curation. We use this dataset for model selection, training, and validation. The training set has 14,920 notes, the validation set has 3,730 notes, and the test set has 4,663 notes.

Model Building and Selection

We use Bidirectional Encoder Representations from Transformer (BERT)[1], a state-of-the-art natural language processing (NLP) model, as our base architecture. A limitation of BERT is that it can only accept texts of up to 512 words. Since EHR notes are usually longer than that, they need to be split into multiple segments, with the prediction from each segment aggregated to get the final label for the note. We are interested in identifying positive colitis cases and symptoms which might only be mentioned in a specific segment. Therefore, we aggregate by selecting the maximum value instead of majority voting, i.e., if any segment is flagged as positive, the entire note is positive. Given this segmentation approach, choosing the right segments can be important. So, our first experiment compared three methods: (1) Use all segments from a note, (2) Randomly select segments from a note, (3) Use the highlighted sentences provided by the curator. The third model using the curator's sentences outperformed the other two.

Hence, selecting the relevant segments from notes instead of using all segments has performance benefits. However, for new datasets, we would not have access to highlighted segments a priori. To this end, we compare 4 different strategies for selecting the segments that are input into BERT. The baseline model uses all segments of the note. The second model uses curator-specified keywords to filter segments. For the third model, we train a logistic regression using a bag-of-words (BOW)[2] model. We use the term frequency-inverse document frequency (TF-IDF)[2] metric to rank the top 1000 words in our notes. We use these words as binary features to train a logistic regression predicting colitis. We extract the top 10 words that were predictive for colitis and use those to filter segments. The results for these models are shown in Table 1.

While logistic keywords model had the highest recall, precision[3] is quite low. The curator keywords have a higher precision, but slightly lower recall. Since colitis has very low incidence rates (1994 positive in 23,313 notes), we are not willing to sacrifice recall for precision. Our fourth model then is trained using the logistic keyword filters, but the test set is filtered by the curator keywords. This model has the best overall performance.

Application and Optimization

We applied the logistic keyword filtering model to a completely new and unseen cohort of 128,314 notes, which contained curator keywords. After reviewing the first 1,080 notes, the curators found an 18% false positive rate, i.e., 194 notes were not colitis cases, most of which mention colitis as a *potential* side effect at therapy initiation. We then trained a second model using the 1,080 notes to identify false positive colitis notes. After reviewing another small sample, curators found that notes that did not contain the word "colitis" were not relevant. Our final workflow (Figure

1) then was to filter data by curator keywords, apply model 1(logistic+curator in Table 1) for colitis identification, apply model 2 for false positive filtering, and then send the positive notes containing colitis for curator review. We built two pipelines, one for colitis mention notes and one for symptom mention notes.

Results

Table 1. Precision/Recall and Training time for Segment Selection Strategies in the colitis mention notes

	Entire Dataset	**Curator Keywords**	**Logistic Keywords**	**Logistic + Curator**
Precision (class = 1)	.76	.71	.57	**.72**
Recall (class = 1)	.96	.97	.98	**.98**
Training Time (hrs)	2.5	**1**	1.5	1.5

Table 1 shows the results of the different segment selection strategies. The fourth model performed the best. It also decreases training time to 1.5 hrs. from the 2.5 hrs. to train on the entire dataset.

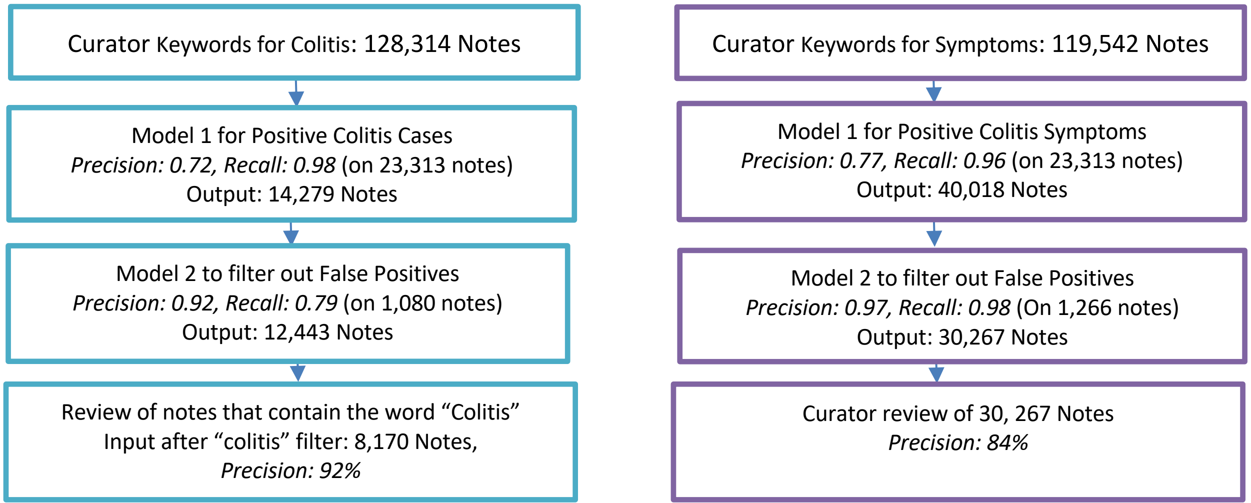
Curator Keywords for Colitis: 128,314 Notes
Model 1 for Positive Colitis Cases
Precision: 0.72, Recall: 0.98 (on 23,313 notes)
Output: 14,279 Notes
Model 2 to filter out False Positives
Precision: 0.92, Recall: 0.79 (on 1,080 notes)
Output: 12,443 Notes
Review of notes that contain the word "Colitis"
Input after "colitis" filter: 8,170 Notes,
Precision: 92%
Curator Keywords for Symptoms: 119,542 Notes
Model 1 for Positive Colitis Symptoms
Precision: 0.77, Recall: 0.96 (on 23,313 notes)
Output: 40,018 Notes
Model 2 to filter out False Positives
Precision: 0.97, Recall: 0.98 (On 1,266 notes)
Output: 30,267 Notes
Curator review of 30, 267 Notes
Precision: 84%

Figure 1. Data Pipeline. *Left:* Results for Colitis Mention Notes, *Right*: Results for Symptom Only notes

Figure 1 shows the data reduction and precision/recall for the models at each step of the curation pipeline. For the final step, the curators reviewed a sample of 20 negative notes but did not find any false negatives. So, we only report accuracy on the positive notes, i.e., precision, for the last step. The colitis mention pipeline had a precision of 92% and reduced the note review load by 93.4%. The symptoms pipeline had a precision of 84% and reduced the note review load by 75%. The slightly lower performance for the symptom model can be attributed to imperfect training data, since the initial set of 23,313 had false positive labels which were curated by non-expert reviewers. Even so, our models greatly accelerated data curation on an unseen cohort.

Conclusion

Data curation is a bottleneck for many informatics research pipelines. Building semi-automated curation tools can accelerate this process. While there are many NLP tools for prediction and sentence completion, customizing them for data extraction is non-trivial. Tuning a curation pipeline requires a tight working loop between data scientists and curators with domain expertise[4], as demonstrated by this work. We show that identifying relevant text to be fed into the model has an impact on performance. The next step is to use attention scores to identify sections that were relevant to the model. Curators can use these to further expedite their review. Applying similar methods to other extraction domains can be greatly beneficial for democratizing research data.

References

1. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.
2. Martineau J, Finin T. Delta tfidf: An improved feature space for sentiment analysis. InProceedings of the International AAAI Conference on Web and Social Media 2009 Mar 20 (Vol. 3, No. 1, pp. 258-261).
3. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. InProceedings of the 23rd international conference on Machine learning 2006 Jun 25 (pp. 233-240).
4. Rahman P, Nandi A, Hebert C. Amplifying domain expertise in clinical data pipelines. JMIR Medical Informatics. 2020 Nov 5;8(11):e19612.