# 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. Before predictive modeling to identify colitis, the data need to be curated from electronic health records (EHRs) since colitis does not have clear diagnosis codes and 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 before they are imported for more extensive expert curation of colitis episodes. In this work, we built a model to accurately identify colitis positive notes.

## 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. We used this dataset for model selection, training, and validation. The training set had 14,920 notes, the validation set had 3,730 notes, and the test set had 4,663 notes. We used Bidirectional Encoder Representations from Transformer (BERT)[1], a state-of-the-art natural language processing (NLP) model, as our base architecture. A constraint of BERT is that it can only accept texts of up to 512 words/token[1]. 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. Empirically, selecting relevant sections of the note had performance benefits. To select relevant segments, we trained a logistic regression using a bag-of-words (BOW)[2] model to predict colitis. We extract the top 10 words that were predictive for colitis and use those to filter segments. Before applying the model to a new dataset, it was filtered by curator keywords to improve precision. A secondary model to filter false positives was trained.

## Results

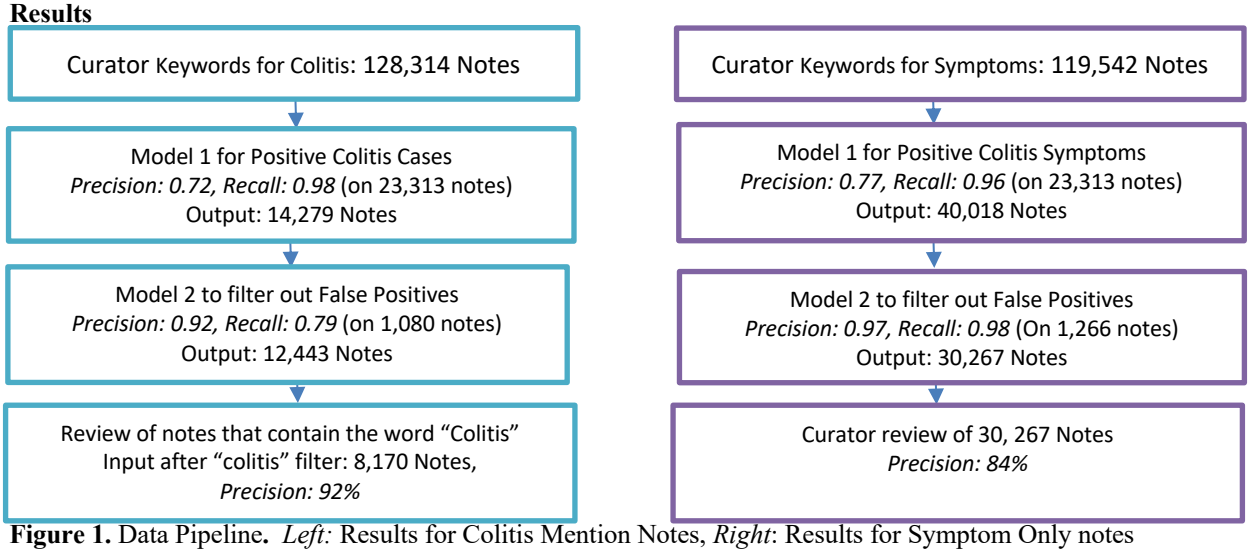


**Figure 1.** Data Pipeline**.** *Left:* Results for Colitis Mention Notes, *Right*: Results for Symptom Only notes

Figure 1 shows the data reduction and precision/recall for the models at each step of the curation pipeline on an unseen dataset. For the final step, the curators reviewed a random 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.

## Conclusion

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.

## 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. 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).