The Presence of Highly Similar Notes Within the MIMIC-III Dataset

Rodney A. Gabriel, MD, Sanjeev Shenoy, MS, Tsung-Ting Kuo, PhD, Julian McAuley, PhD, Chun-Nan Hsu, PhD.
University of California, San Diego

Abstract

Highly similar or even identical notes are problematic when one is compiling statistics or training predictive algorithms that model the language or attributes in notes. We developed an algorithm to identify and characterize highly similar notes within the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III) dataset. We found that there were multiple instances of exact copies, common outputs, and template notes that form the public domain MIMIC-III dataset.

Background/Introduction

- With the advent of the electronic health record (EHR), there became a widespread presence of copy-and-pasting when composing clinical notes.
- Highly similar or even identical notes are problematic when one is compiling statistics or training predictive algorithms that model the language or attributes in notes.
- For example, duplicates may cause an outlier detection algorithm to erroneously identify a rare condition as being common or a predictive model may erroneously identify correlations between symptoms.
- Correcting duplication in clinical note data is challenging, as the sources of duplication are widely observed but poorly understood.
- De-duplication and data cleansing approaches may improve the quality of clinical note corpora as a vital information source for meaningful use of EHR.

Materials and Methods

- We sought to explore the presence of highly similar notes within the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III) dataset (2,065,096 notes).
- We used Jaccard Similarity (JS) and an approximation algorithm that minimizes pairwise comparisons to find all similar notes in the dataset exhaustively (Figure 1).
- Figure 2 shows the results. We defined three classes of similar notes: 1) Exact copy: JS=1, same patient/date 2) Common output: JS=1, diff patient or date 3) Similar notes: JS > 0.7 and count their proportions (Figure 3).
- We then created a validation list consisted of pairs of documents with JS >= 0.3 by generating 2 million random pairs and calculating the Jaccard Similarity (JS) for each pair. We used this list to validate if notes were clustered appropriately (Table 1).

Results/Evaluation

Figure 2. Line graph illustrating the number of total notes and clusters generated from the algorithm based on Jaccard Similarity threshold.

Figure 3. Proportion of notes classified as exact copies, common output notes, or similar notes based on Jaccard Similarity threshold setting.

Conclusions

- We found that there were multiple instances of exact copies, common outputs, and similar notes from the public domain MIMIC-III dataset.
- It is unclear for the reasons of highly similar notes, but this could be related to pervasive practice of copy-and-pasting, note template utilization, common output notes from automated machines, and technical errors in note processing.