Electronic medical records (EMR) are increasingly utilized in clinical practice and research, allowing for more efficient availability of rich patient records. However, most use of EMR is limited to coded, structured, administrative data, while the vast majority of patient information (e.g. disease subtype, severity, medical device usage, etc.), is tied up in narrative clinical notes. This makes rare clinical events difficult to identify in EMR, as they are seldom defined by specific diagnosis code.

Traditionally, this data has been extracted utilizing costly and time-consuming full manual chart review; however, rare events may require review of such a large array of records that manual chart review becomes impractical. Natural language processing (NLP) can be a solution; however, the number of relevant clinical notes and highlighted the finding of keywords/phrases, which required a identification of ≤1,100 (0.5%) patients with the specific subtype. Results for other subtypes will be presented.

This NLP-assisted semi-automated chart review approach maintains a large sample size for clinical research while managing the need for costly and time-consuming full manual chart review.

This demonstrates the ability to identify rare clinical events from EMR clinical notes when full NLP is not sufficient and without the need for costly and time-consuming full manual chart review.

This NLP-assisted semi-automated chart review approach maintains a large sample size for clinical research while managing workload necessary to extract meaningful information from EMR.

ABSTRACT

Background: Rare clinical events are difficult to identify in electronic medical records (EMR) as they are seldom defined by specific diagnosis code. Traditionally, they are extracted utilizing costly and time-consuming manual chart review, however, rare events may require review of such a large array of records that manual chart review becomes impractical. Natural language processing (NLP) can be a solution; however, the number of relevant clinical notes and highlighted the finding of keywords/phrases, which required a identification of ≤1,100 (0.5%) patients with the specific subtype. Results for other subtypes will be presented.

Methods: Patient case clinical event studies were identified using structured data, and relevant clinical notes were mined for keywords related to the subtype. Clinical annotators manually mined the resulting snippets of text and identified actual instances of the subtype. A full chart review of the documents associated with these events was completed for validation and reproducibility.

Results: 360,000 patients were included with a diagnosis of the general type of cancer. From those patients, 5,848,293 EMR records were obtained.

NLP-assistance using keywords/phrases significantly narrowed down the number of relevant clinical notes and highlighted the resulting keywords/phrases. This allowed for a much faster manual chart review process, which resulted in identification of ~1,100 patients (0.5%) with the specific subtype.

Conclusions: This NLP-assisted semi-automated chart review approach maintains a large sample size for clinical research while managing workload necessary to extract meaningful information from EMR.

REFERENCES


CONSIDERATIONS

NLP processes are often dynamic and iterative, and can be difficult to standardize. In addition, unstructured notes are always held to a systematic structure and can be highly variable across and between health systems. NLP must be carefully refined and validated.

Data available for NLP-driven extraction is limited to what the clinician documents in the clinical notes.

METHODS & RESULTS

Several rare clinical event studies were identified over the course of research. Here we present one example: identification of patients diagnosed with a specific subtype of cancer. A full chart review of the documents associated with these events was completed for validation and reproducibility.

Figure 1. Structured vs. Unstructured EHR Data

Figure 2. NLP-assisted Chart Review Process & Results

Figure 3. NLP-assisted Chart Review Process & Results

360,000 patients were included with a diagnosis of the general type of cancer. From those patients, 5,848,293 EMR records were obtained.

NLP-assistance using keywords/phrases significantly narrowed down the number of relevant clinical notes and highlighted the resulting keywords/phrases. This allowed for a much faster manual chart review process, which resulted in identification of ~1,100 patients (0.5%) with the specific subtype.

CONCLUSIONS

Figure 2. NLP-assisted Chart Review Process & Results

EHR data, including demographics, treatments, and diagnoses (NLP is a tool of the future, accuracy and completeness)

EHR data, including demographics, treatments, and diagnoses (NLP is a tool of the future, accuracy and completeness)

Figure 1. Structured vs. Unstructured EHR Data

Figure 2. NLP-assisted Chart Review Process & Results

Figure 3. NLP-assisted Chart Review Process & Results

360,000 patients with a diagnosis of the general type of cancer. From those patients, 5,848,293 EMR records were obtained.

NLP-assistance using keywords/phrases significantly narrowed down the number of relevant clinical notes and highlighted the resulting keywords/phrases. This allowed for a much faster manual chart review process, which resulted in identification of ~1,100 patients (0.5%) with the specific subtype.

CONCLUSIONS

This demonstrates the ability to identify rare clinical events from EMR clinical notes when full NLP is not sufficient and without the need for costly and time-consuming full manual chart review.

This NLP-assisted semi-automated chart review approach maintains a large sample size for clinical research while managing workload necessary to extract meaningful information from EMR.

REFERENCES


CONSIDERATIONS

NLP processes are often dynamic and iterative, and can be difficult to standardize. In addition, unstructured notes are always held to a systematic structure and can be highly variable across and between health systems. NLP must be carefully refined and validated.

Data available for NLP-driven extraction is limited to what the clinician documents in the clinical notes.

Limitations of this study include the variability of patient notes across health systems, the potential for keyword/phrases to be misspelled, and the inability to capture information from external sources.