Use of NLP to Augment Real-World Data (RWD) and Identify Eligible Patients at Scale for Oncology Studies

Avi Raju, Jessica K. Paulus, Zhaohui Su, Genc Doko, Nicholas J. Robert
Ontada, Boston, MA

MSR30

Background and Objectives

- The substantial volume of unstructured data contained within electronic health records (EHRs), with estimates as high as 80% (Li et al., 2022), represents a rich and untapped source of insights for real-world evidence studies.
- Practical examples of the application of natural language processing (NLP) approaches are needed to assess their performance in generating fit-for-purpose real-world studies from unstructured EHR data.
- We evaluated the application of NLP technologies to support feasibility assessments of real-world studies and the implementation of complex inclusion/exclusion criteria to identify eligible patients.

Methods

- The primary data sources are large sets of text-based clinical notes from iKnowMed[SM], an oncology-specific EHR system.
- Pretrained healthcare NLP models from John Snow Labs, Inc. as well as internally-developed models were utilized to identify patients meeting complex clinical criteria across 20 feasibility assessments for varying cancer types.
- Prior to utilization, the models were evaluated using internal validation datasets to measure precision, recall, and F1 scores to ensure model accuracy and reliability.
- Once the NLP models generated results, a randomly selected sample (5-10%) of the results were validated by clinical annotators for each assessment to further validate the model.
- A workflow illustrating the end-to-end production NLP process for the feasibility assessments is shown in Figure 1.

Reference:
Li, I., Pan, J., Goldwasser, J., Verma, N., Wong, W. P., Nuzumlalı, M. Y., Rosand, B., Li, Y., Zhang, M., Chang, D., Taylor, R. A., Krumholz, H. M., & Radev, D. (2022). Neural Natural Language Processing for unstructured data in electronic health records: A review. *Computer Science Review*, *46*, 100511. https://doi.org/10.1016/j.cosrev.2022.100511

Methods

Figure 1. Feasibility assessment NLP workflow

Step 1 – Clinical notes: Clinical notes are extracted for the patient cohort and timeframe of interest.

Step 2 – Clinician guidance: A clinician provides guidance on the keywords to use to find the desired variables.

Step 3 – Note excerpt selection: Clinical notes are filtered for the keywords provided by the clinician.

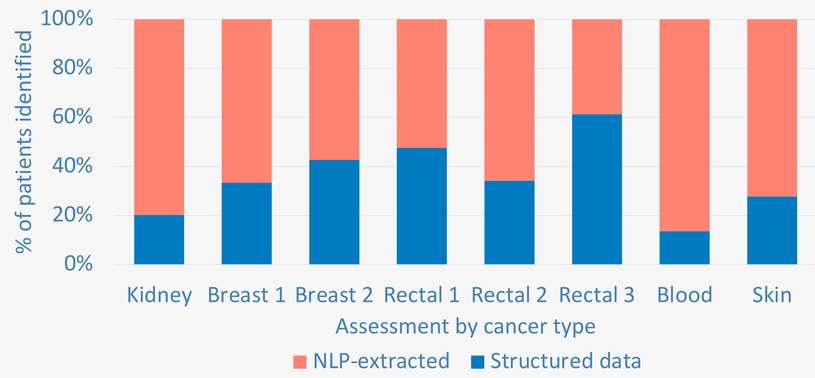
Step 4 – Model execution: A validated NLP model is executed to extract the values of the clinical variables of interest.

Step 5 – Subject Matter Expert validation: Clinical annotators validate a sample of the model results.

Step 6 – Result presentation: Aggregated counts are shared with the client to determine if the sample size is sufficient to propose a real-world evidence study.

Results

Figure 2. Structured vs. NLP-extracted data source distribution for feasibility assessments



- In 8 of the 20 assessments conducted, comparable structured data was available. NLP increased the desired patient population by an average of a 3.5-fold (range 1.1-7.4). The median sample size pre-NLP was 110 patients, and post-NLP, it was 315 patients. Figure 2 illustrates the percentage of patients identified by data source (structured vs. NLP).
- In another 9 of the same 20 assessments, NLP identified a patient cohort when comparable structured data was not available. The median number of patients identified in such assessments was 244 (range 40 – 1482 patients).
- In the remaining 3 assessments, NLP significantly narrowed down the population for manual chart review, resulting in a mean reduction of 80% (range 69-89%).

Conclusions

Augmentation of feasibility exercises with NLP can boost sample sizes by significant margins, increasing the statistical power and clinical generalizability of proposed studies.

This approach offers a scalable solution for conducting real-world evidence studies and supplements high-effort and costly activities such as manual chart abstraction.