Enriching the value of real-world oncology data with important clinical information from unstructured data sources for better clinical insight generation

Verma V, Rastogi M, Paul A, Gaur A, Daral S, Kukreja I, Nayyar A, Roy A, **Khan S**, Arora R, Mantri G

RWD62

Introduction

- Real-world data (RWD) from sources like administrative claims or patient registries offer valuable insights into patients' health status and healthcare delivery outside of clinical experiments like randomized controlled trials (RCTs).
- However, the use of RWD has limitations in providing deeper insights into patients' clinical profiles beyond the primary diagnosis. Therefore, the use of RWD is limited in cases where treatment and care innovations are driven by patients' clinical classification.
- Oncology, for instance, requires knowledge of patients' clinical profiles, including cancer staging, metastatic status, genetic profile, and biomarkers, to design precision medicine and targeted treatment pathways.
- Unstructured electronic health record (EHR) data, such as patients' notes, contains valuable clinical information that is not captured in structured databases like claims or EHRs. Thus, extracting and integrating this information with structured databases offers a panoramic view of patient journeys.

Objective

The objective is to develop an NLP model capable of extracting relevant information from physician notes which can enrich the existing structured databases. We chose colorectal cancer (CRC) as a use case for developing the data enrichment model development.

Method

- Optum's de-identified Market Clarity Database was used to identify patients with primary colorectal cancer from January 2022 to December 2022. Patients with other concomitant cancer types were excluded.
- Following clinical data elements associated with colorectal cancer were considered for data extraction:
  - Cancer staging (Numeric & TNM)
  - Metastatic status (Yes/No)
  - Biomarkers: Microsatellite Instability (MSI) -Stable/Unstable; Mismatch repair protein deficiency (d-MMR) – proficient/deficient
- Physician notes (unstructured EHR data) from Optum's Physician Notes database were used after de-identifying them by removing all protected health information (PHI) and personally identifiable information (PII).
- A sampling strategy was developed based on the frequency distribution of data elements. Notes that are rich in data elements were selected for manual annotation.
- Annotated notes were divided into train, test, and validation sets. The training set was used for NLP model development. Validation and test sets were used to evaluate model performance.
- Machine learning-based classification models (NER classification) and transformer-based NLP models (BERT) were used to identify relevant clinical texts from patients' notes. The models were fine-tuned using the annotated data.
- The model's accuracy was assessed in terms of precision, recall, and F1 scores for each concept.
- In this analysis, four iterations were required to achieve the desired accuracy level (precision> 80%, Recall >70%).

Results

- Out of 2,541 colorectal cancer patients, only 659 (25.9%) had colorectal as their primary cancer type and had at least 1 physician note. The total number of physician notes associated with these patients were 104,360, out of which only 21,553 (20.6%) notes in total had >=1 data element in scope.
- Based on the frequency distribution, a total of 1,671 (~8%) notes were considered for manual annotation exercise in four batches which were used to create training, validation and testing samples [Table 1].
- Model performance was evaluated based on precision, recall and F1 score [Table 2]. Performance was calculated at instance level, where positive prediction of all the words in a phrase was considered as true positives.

Table 1: Frequency distribution of data elements for sampling approach

Data Element Details in Notes	Total Notes	Annotation 1	Annotation 2	Annotation 3	Annotation 4	Total
Details on CRC but no biomarkers	18,596	185 (1.0%)	220 (1.2%)	254 (1.3%)	0 (0.0%)	659 (3.5%)
Details on both CRC and biomarkers	1,958	117 (5.9%)	110 (5.6%)	103 (5.2%)	120 (6.1%)	450 (22.9%)
Details on biomarkers but no CRC	233	116 (49.7%)	28 (12.0%)	84 (36.0%)	0 (0.0%)	228 (97.8%)
Note of CRC patient but no details on disease	692	138 (19.9%)	110 (15.8%)	53 (7.6%)	0 (0.0%)	301 (43.4%)
Notes of CRC patient but just mention of disease	74	74 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	74 (100%)
Column Total	21,553	608 (2.8%)	464 (2.1%)	479 (2.2%)	120 (0.6%)	1,671 (7.7%)
Notes sample		2.8%	(+) 2.1%	(+) 2.22%	(+) 0.6%	7.7%

Fig 1: Frequency distribution in annotated data

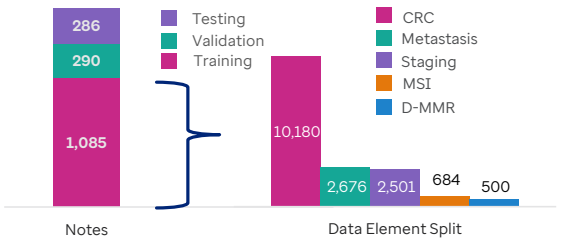

Table 2: NER classification model performance on combined batches

Data elements	Precision	Recall	F1-score	Support
CRC	0.96	0.96	0.96	945
Staging	0.88	0.93	0.91	257
Metastasis	0.96	0.93	0.94	662
D-MMR	0.78	0.91	0.84	54
MSI	0.91	0.97	0.94	128

Conclusion and Limitation

- Higher precision, recall, and F1 score (>80%) for all concept terms showed the NLP model could extract specific data elements from unstructured text data, which can be used to supplement structured claim data.
- The natural language model developed for this analysis is specific to Colorectal cancer and may not perform as expected for other cancer types. Further model development is required to achieve similar performance statistics.

Optum