Category: Cross-Cutting Innovation and Collaboration – Work in Progress

NYU Prostate Cancer Automated Database (NYU Pro AD Pilot Project): Leverage AI in Operational Workflow

A. Bhatt, V. Lam, J. Sommer, N. Chowdhury, R. Belenkaya, D.R. Wise

Laura and Isaac Perlmutter Cancer Center at NYU Langone Health

1. Background

Well-built oncology registries are essential for understanding patient characterization, cohort identification, and developing optimal clinical research design. However, their creation is frequently constrained by labor-intensive manual data extraction, vulnerability to staffing shortages, and persistent backlogs that delay research output. To address these challenges, we piloted an AI-driven approach within the NYU Pro-AD (NYU Prostate Cancer Automated Database) project in our genitourinary oncology group, in collaboration with the Perlmutter Cancer Center data hub team. Leveraging large language models (LLMs) and natural language processing (NLP), we automated extraction of clinically relevant data from unstructured pathology reports and physician notes to create a structurally annotated, real-time clinicopathologic biospecimen registry that enhances analytical access.

2. Goals

- To enhance operational efficiency by reducing manual data entry and retrospective chart review and to optimize resource utilization and streamline research workflow
- To improve data processing timelines from months to days, enabling access to real-time updated data for faster data-driven clinical research

3. Solutions and Methods

A python-based pipeline, integrated with LLM and NLP tools, was used to process clinical notes and pathology reports, removing irrelevant metadata and standardizing data formats. The model extracted data of 1,200 prostate cancer patients enrolled in the NYU genitourinary tumor registry, a non-therapeutic investigator-initiated trial (IIT), from EPIC. The model extracted data for key variables, such as medical record numbers, histopathological findings, tissue collection site, date, procedure type and disease state. The model identified and labeled clinically relevant oncology terms based on predefined search criteria and mapped them into structured data fields for standardized export into an Excel spreadsheet.

4. Outcomes

The model processed 1,619 pathology reports from 1,200 patients, demonstrating robust and scalable performance. Validation against 169 manually curated records of specimen IDs, which showed 92.7 percent accuracy (95 percent CI: 88.88 percent –96.92 percent). The model generated a structured database of biospecimens with pathological variables (histopathological pattern, tumor grade, Gleason score, metastatic stage) and annotated according to prostate cancer working group three (PCWG3) diagnosis terms identified in physician notes. Implementation increased sub-study initiation capacity

from an average of one per year to three new projects in the last six months and reduced data retrieval timelines from months to automated recurring extraction. A key limitation was the variability in clinical terminology, resulting in incomplete or inconsistently captured data elements.

5. Lessons Learned and Future Directions

Future initiatives will focus on enhancing data integration and standardization of terminologies within oncology workflows. We aim to incorporate health disparity metrics, using population-level data to drive equitable care for underserved populations. Additionally, expanding our framework to integrate genomic data and longitudinal treatment outcomes will enable the development of predictive models for personalized therapies. By creating institutional oncology registries and leveraging advanced machine learning techniques, we strive to create scalable, reliable models that accelerate data-driven decision-making, support scientific research, and improve patient outcomes in cancer care using institutional real world patient data.