Background

Center for Childhood Cancer Research (CCCR) Biobank: The CCCR Biobank is a specimen biobioery at the Children’s Hospital of Philadelphia (CHOP), responsible for collecting multiple specimens over time from cancer cases affecting the pediatric population including solid tumors and hematologic malignancies for translational biomedical research.

Biorepository Informatics Infrastructure: Allows for scalable, modular, encrypted data integration, includes a custom extract, transform, and load (ETL) process to integrate data from systems used to manage biospecimens, de-identified clinical data, and other associated research data, as well as an open-source, web-based data discovery tool, developed by CHOP, used to display and query the filtered dataset. (See Systems Architecture)

Pathology Report Integration Workflow

Methods

ETL (Extract, Transform, Load) Framework – research protocol-based series of scripts, which import data from different sources, link and restructure data, remove protected health information (PHI), and ETI (Extract, Transform, Load) Framework

Benefits of automation:

• Reduction of time and effort spent on manual data entry tasks, standardization of data quality, creation of scalable workflows and processes for increasing specimen and data collection volumes

• 100% of sample procurement events have been linked with a pathology report, with the exception of ones containing erroneous data or formatting errors

• There is an ongoing effort to clean data and perform quality analysis to ensure the percentage of linked reports remains at 100%, that each sample procurement event is associated with only one pathology report, and to correct erroneous data and any import formatting errors

• The automated report linking process has substantially reduced the amount of time and effort spent on the linking process compared to manual linking:
  - Manual linking was estimated at approximately 10 minutes to complete a single sample procurement event link. For 500 events, manual effort would be approximately 5000 minutes (83.3 hours).
  - Automated linking for 500 events requires approximately 1.57 minutes (94 seconds), which is 3191 times faster than manual linking

• Preliminary work has been performed to test the integration of additional clinical report types using the same linking and integration workflow, with successful integration of Cytopathology reports from the EMR, demonstrating the potential application of this process to other projects.

Results

• Selection of minimum data set or common data elements

  • CCCR Biobank is a general purpose cancer biobioery, and as such there are a wide range of research interests which need to be accounted for

  • Inconsistent data quality

  • Electronic data capture forms contain variable data quality and/or quantity

  • LIMS data quality decreases for early records

  • Data import errors in EHR reports from other clinical laboratory management systems

  • Reliance on dates in unconnected systems can be a tenuous method for linking data

  • No direct links (keys/identifiers) between clinical EMR data and LIMS or REDCap data

  • EHR may not contain enough structured data to eliminate the need for electronic data capture forms

  • Unstructured data will require additional processing for extraction of structured data elements

Next Steps

• Continue refining the minimum data set

• Revise REDCap eDC forms based on automated clinical data extractors

• Application testing with biobioer staff and researchers/investigators

• Launch a production instance of the integrated database and Harvest data discovery tool for CCCR researchers

• Integration of other report types:
  - Cytopathology Reports
  - Solid Tumor Next Generation Sequencing Panels
  - Minimum Residual Disease (MRD) Reports

• Automate the selection of CCCR consented patients for EHR data extraction

• Natural Language Processing (NLP) of pathology reports to extract structured data elements

References


Download a PDF of this poster: