The Role of Federal Agencies to Advance Progress in EHRs: Cancer Surveillance as an Example

MARCH 1, 2022

Objectives

Understanding challenges in using EHR data for cancer surveillance

Examples:

- Automated extraction via DL/NLP across heterogeneous health care providers
- Source Data linkages – use cases of pharmacy data and genomic tests

Using EHR data for cancer surveillance

Central Cancer registries have legal authority to access to health records under state public health reporting regulations

Most relevant information from EHRs consists of unstructured text

The current process is manual review and abstraction

- EHRs do not consistently capture information (cancer patients seen at multiple different facilities and treatment centers)
- Manual process not sustainable for the Increasing volume of data required for surveillance (e.g. treatment detail, genomic characterization etc.)

The solution

- Focus on ***targeted*** EHR components for automated extraction
- Perform data linkages with source organizations (external partners)

Automated Extraction of Structured Data:

Focusing on targeted EHR components

Example: Automated extraction of structured data from unstructured text

Focus on most consistent and relevant EHR components for surveillance

- Pathology reports and radiology reports
- Contain key data on tumor characterization and status: site, histology, laterality, behavior, recurrence

Challenges

- Inconsistent content from unstructured text from thousands of pathologists
- >500 histology categories and >100 cancer site classifications
- Limited training data to enable robust algorithm development (rare tumors/histologies present class imbalance challenges)
- Need for high accuracy in registries (>97%) across all data elements being extracted

Automated Extraction using NLP/DL across heterogeneous data

NCI-DOE collaboration: develop algorithms to extract key structured data from ***unstructured pathology*** reports received from >400 pathology labs (4-5 Million path reports received in real time per year)

- Value of Automation:
  - Currently manually reviewed and annotated – automation to reduce labor (estimated >60,000 person hours per year on screening/extraction process)
  - SEER is moving towards near real time reporting of incidence- infeasible with manual extraction

- Solutions
  - Used existing "gold standard" manually extracted data from 7 registries to train the algorithm and develop an API for processing unstructured pathology reports
  - Iterative testing and development with review of mismatches for retraining
  - Development of ***uncertainty quantification*** process- identifies path reports with <97% accuracy and sends for human for review reports with lower accuracy

Using NLP/DL across heterogeneous data sources: Current Status

- Implemented in production work flow in 11 SEER registries
- Allow automated processing (no human review) of 23-25% of all path reports (>4 million received annually) with >98% accuracy across ALL FOUR data elements – site, histology, laterality and behavior
- API **18,000X faster than human** – translates to saving 11,500 labor hours for this process
- Leveraging work flow to:
  - iteratively improve algorithm (through feedback from manual process and 10% random sample for manual review of auto-coded)
  - Train personnel for increasing consistency and accuracy for non-auto-coded data
- Leveraging API for ***recurrent metastatic disease*** developed for pathology reports using transfer learning to apply to radiology reports

Data Linkages: leveraging source data for surveillance

Linkages with SEER

Linkages with source data (external partner organizations) offers an opportunity to

- Maximize all matches for data generated by source rather than through EHR
- Example Oncotype DX- manual reporting from hospital EHRs missed 40% of all tests

Done under surveillance reporting regulations – registries allowed to have PII for validation of matches

Challenges and opportunities for 2 use cases

- Pharmacy data
- Genomic test panels

Linkages- Use Case 1. Pharmacy Data

SEER registries receive all antineoplastic drugs provided by Walgreens/CVS/RiteAid as pharmacists are covered under state public health reporting regulations for cancer

Provides important information on treatment ***not routinely available to hospital registries (in EHRs)***

- Treatment for both initial course of therapy as well as subsequent treatment of recurrent disease
- Longitudinal coverage potentially identifies patients with recurrence

Examples of SEER-Linked* Pharmacy Data (2013-2020)

Sample of Tyrosine Kinase Inhibitor Use - 2013-2020

Drug Name	Patients	Filled Prescriptions
TARCEVA	2,129	17,423
SPRYCEL	1,934	27,729
IMATINIB MESYLATE	1,929	31,326
		8,035
		20,208
		4,580
		8,718
TAGRISSO	1,247	13,936
SUTENT	1,235	8,189
TASIGNA	1,020	15,830
CABOMETYX	791	4,939
INLYTA	744	4,884
LENVIMA	544	2,569
TYKERB	490	2,496
XALKORI	488	4,020

TKIs- 34 agents
188,000 Fills for >20,000 Patients

PARP Inhibitor** Use in 1,095 Patinets by Cancer Site from SEER-Linked Pharmacy Data (2017-2020)

Cancer Site	N Patients
Ovary	504
Breast	229
Other Female Genital Organs	132
Prostate	58
Peritoneum, Omentum and Mesentery	39
	31
	23
	13
Melanoma of the Skin	13
Thyroid	12
Brain	8
Urinary Bladder	7
Lung and Bronchus	6
Acute Myeloid Leukemia	4
Cervix Uteri	4
Other Biliary	4
Other cancer sites*	50

PARP Inhibitors 3 agents
7,000 Fills for 1,095 Pts.

* Sites with < 4 patients receiving agents

** Olaparib (approved 2014), rucaparib (approved Oct 2018), talazaporib (approved Oct 2018)

Use of a CDK 4/6 inhibitor (Palbociclib) in 4,302 Patients by Cancer Site (2013-2020)

Cancer Site	N Patients
Breast	4100
Corpus Uteri	28
Melanoma of the Skin	25
Thyroid	23
Retroperitoneum	22
	18
	17
	16
	13
Other Site	10
Kidney and Renal Pelvis	8
Non-Hodgkin Lymphoma	8
Ovary	7
Prostate	7

CDK 4/6 Inhibitors
Palbociclib
45,000 Fills for 4,302 Pts.

- Includes 11 of 20 SEER Registries- Pharmacy data from CVS/Walgreens/

Linkages- Use Case 1. Pharmacy Data

Challenges:

- Prescription coverage by insurers in the US is heterogeneous and inconsistent over time
  - Certain anti-neoplastic drugs are only provided by designated providers (i.e. CVS, Cardinal or PBMs) which can change annually
  - One and done- often the initial prescription can be filled anywhere – but subsequent prescriptions required to be filled by designated provider
- How do we reduce risk of bias in releasing the data via SEER?

Solutions:

- Release data under carefully controlled circumstances with review of proposed analysis
- Perform sandbox testing of linked data to understand where biases might be and comparison with more comprehensive data sources (SEER-Medicare Part D) to understand where gaps might exist
- Continue to increase coverage with additional linkage partners
  - Mail order
  - PBM – such as Optum-UHC linked data

Linkages- Use Case 2. Genomic Data

Unlike pharmacy data linkages with selected genomic pathology labs provide comprehensive information on all patients receiving the test

Examples:

- Oncotype DX 21 gene assay for breast,
- Oncotype and Decipher MGP for prostate,
- Castle MGP for melanoma

Linkages- Use Case 2. genomic data panels

Challenges:

- Assuring appropriate matches- patients often move over time and so matching based on name, dob and address (often only PII available from external orgs) may not match address at time of diagnosis in registry

Solution

- SEER Registries now link with Lexis Nexis to obtain longitudinal residential history that can permit adjudication of questionable matches
- Increases match rate significantly
- Permits manual review for uncertain matches

Conclusions

Using currently available EHR data for cancer surveillance remains challenging but-

- Specifically targeting EHR components can provide significant value
- Understanding what components of a patients health care EHRs actually contain is critical to optimizing the value (e.g. community oncology practice vs. free standing radiology center)
- It is not a fast or easy process!

Thank you for your attention.