# Towards a Scalable Clinical Data Annotation and Processing Pipeline to Support Cancer Surveillance

Paul Fearn, PhD, MBA Surveillance Research Program Division of Cancer Control and Population Sciences paul.fearn@nih.gov



LabKey User Conference Oct 5<sup>th</sup>, 2017

# Context of Cancer Surveillance



### **Cancer Surveillance 101**



### The SEER Cancer Registries



## NCI SEER and Surveillance Research Program (SRP)

- SRP organizationally
- SRP goals for SEER registries
  - Natural language processing (NLP) and other automation to improve efficiency and reliability and accuracy of data collection
  - Linkages of new detailed diagnostic and treatment data sources to improve
  - Data quality initiatives
  - Grow and sustain national population-based research resource

### Surveillance Informatics Branch (SIB) Initiatives

- Data Acquisition and Linkages
- SEER\*DMS enhancements
- NLP: DOE collaboration, inter-agency collaboration with CDC, FDA, de-identification, data quality assessment and improvement



https://surveillance.cancer.gov/branches/sib/

# Data Acquisition and Linkage



From Donna Rivera



## **SEER\*DMS Enhancements**

Usability assessment and human computer interaction design framework



## NCI DOE Pilot 3: Overview

#### Population Information Integration, Analysis and Modeling

Improve the effectiveness of cancer treatment in the "real world" through computing



## **NCI-DOE Pilot 3 - Aims and Mission**

Aims	NCI	DOE
Aim 1: Information Capture Deep text comprehension of unstructured clinical text to improve the capacity of the cancer surveillance program	Enhance breadth, quality and timeliness of data collected in SEER with confidence-rated (UQ) automated data element identification from unstructured data sources	Advance <i>descriptive</i> analytics with scalable deep learning based NLP tools for CORAL and exascale architectures
Aim 2: Information Integration Scalable graph, visual, and in- memory heterogeneous data analytics and inference methods to understand drivers in patterns of cancer outcomes and predict clinical endpoints	Integrate a variety of new data sources containing relevant clinical information feeding cancer patient surveillance databases to understand patient outcomes and cancer drivers in the real world population - beyond clinical trials	Advance <i>predictive</i> analytics with scalable statistical inference tools, graph and visual analytics for CORAL and exascale architectures
<b>Aim 3: Population-level Modeling</b> Data-driven modeling of patient- specific and population level cancer health trajectories to guide precision cancer care	Develop capability for data- driven modeling of individual patient outcomes using an integrative cancer patient model based on comprehensive data from Aims 1 and 2	Advance <i>prescriptive</i> analytics that involve large-scale data-driven modeling and simulation for CORAL and exascale architectures

# Challenges



### **NCI-DOE Pilot 3 Challenges**

- Develop scalable NLP tools and processes
- Need for robust and scalable tool for annotation and review
  - Developing training and validation datasets through human annotation and review
  - External review and validation of human annotation
  - Review and validation of NLP algorithms with UQ
- Scale to multiple batches, schemas
- Tracking and sharing data through a secure data sharing portal

# Role of LabKey





#### Clinical Document Annotation and Processing (CDAP) Pipeline





#### **CDAP** Pipeline





#### Working List for Managers, Annotators, and Reviewers

straction Task List										
3 case	es for mo	e to assig	n							
GRID	VIEWS 🔻	REPORTS -	CHARTS - EXPOR	T - PRIN	T PAGING 🔻					
GRID	VIEWS 🔻	REPORTS -	CHARTS - EXPOR Report Number	MRN	T PAGING - Last Modified	Pulled Date	Status	Document Type	Abstractor	Reviewer
GRID	DETAILS >	REPORTS - Tasks ASSIGN >	CHARTS - EXPOR Report Number 01-REC-0000000441	T - PRIN MRN 484932	T PAGING - Last Modified 2016-10-08 21:07	Pulled Date	Status Ready for assignment	Document Type Pathology Reports	Abstractor	Reviewer
GRID	DETAILS DETAILS >	REPORTS - Tasks ASSIGN > ASSIGN >	CHARTS - EXPOR Report Number 01-REC-0000000441 01-REC-0000000442	T PRIN   MRN 484932   485263 485263	T PAGING - Last Modified 2016-10-08 21:07 2016-10-08 21:07	Pulled Date	Status Ready for assignment Ready for assignment	Document Type Pathology Reports Pathology Reports	Abstractor	Reviewer

#### 2 cases for me to abstract

GR	ID VIEWS 👻	REPORTS -	CHARTS - EXPORT	PRINT	PAGING -					
	r	Tasks	Report Number	MRN	Last Modified	Pulled Date	Status	Document Type	Abstractor	Reviewer
	DETAILS »	ABSTRACT )	02-REC-3000679345	01036995	2016-10-08 21:07		Ready for initial abstraction	Pathology Reports	kristinf	reviewer1
	DETAILS »	ABSTRACT )	02-REC-3000697853	00833487	2016-10-08 21:07		Ready for initial abstraction	Pathology Reports	kristinf	

#### 3 cases for me to review

	GRID	GRID VIEWS - REPORTS - CHARTS - EXPORT - PRINT PAGING -									
			Tasks	Report Number	MRN	Last Modified	Pulled Date	Status	Document Type	Abstractor	Reviewer
		DETAILS »	<b>REVIEW &gt;</b>	01-REC-0000000443	485401	2016-10-08 21:08		Ready for review	Pathology Reports		kristinf
		DETAILS »	<b>REVIEW</b> >	01-REC-0000000445	486414	2016-10-08 21:09		Ready for review	Pathology Reports		kristinf
(		DETAILS »	REVIEW >	01-REC-3000698369	10046190	2016-10-08 21:09		Ready for review	Pathology Reports		kristinf

#### Main Annotation UI

APPROVE

REPROCESS

#### Pathology Report - 02-REC-000000188

ClassifiedDiseaseGroup	lung	1
ALK Test		
ALK test performed:	Yes	1
ALK test method:	FISH-based	1
ALK test result:	Negative	1
EGFR Test		
EGFR test performed:	Yes	1
EGFR test method:	PCR-based	1
EFGR test result:	Negative	1

Tumor>

<Item naaccrId="tumorRecordNumber">02</Item>

<Item naaccrId="recordDocumentId">REC-0000000188</Item>

<Item naaccrId="textPathClinicalHistory">Clinical Diagnosis and History L LUNG MASS LEFT UPPER LOBE PULMONARY NODULE OPER: L VAT WITH L UPPER LOBECTOMY POSSIBLE LEFT THORACOTOMY L VAT L LOWER LOBECTOMY FROZEN SECTION Intraoperative Consult FSDX#1: CARCINOMA, NON-SMALL CELL. TUMOR IS 2 MM FROM STAPLE LINE MARGIN. \*\*INITIALS IHC/MER/ap FSDX#3: SHAVE BRONCHIAL AND VASCULAR MARGINS FREE OF TUMOR. \*\*INITIALS

<Item naaccrId="textPathComments">4000 \*\*PLACE, KY \*\*ZIP-CODE Copy To: \*\*NAME[ZZZ], M.D., \*\*NAME[YYY M]. Specimen(s) Received 1: LEFT UPPER LOBE PULMONARY NODULE-FS-ls 2: LEFT UPPER LOBE PULMONARY NODULE-FRESH FOR PRECISION-ls 3: LEFT UPPER LOBE-FS-ls 4: STATION 7-ls 5: L 10-ls 6: AP WINDOW-ls</litem>

<Item naaccrId="textPathFormalDx">Final Diagnosis 1: LEFT UPPER LOBE WEDGE RESECTION: ADENOCARCINOMA, MODERATELY WELL-DIFFERENTIATED (GRADE 2). MAXIMUM ROSS DIMENSION OF TUMOR IS 15 MM. VISCERAL PLEURAL INVASION TUMOR UNIFOCAL. OT IDENTIFIED. VASCULAR INVASION: NOT IDENTIFIED. MARGIN OF WEDGE FREE OF 'UMOR. COMMENT: Tumor appears to be arising in association with scar. I feel tumor on outing histology is consistent with origin from lung. However, immunoperoxidase stains ill be performed in an effort to support this impression. 2: LEFT UPPER LOBE WEDGE: SUBMITTED TO PRECISION THERAPEUTICS. 3: LEFT UPPER LOBE: RESIDUAL CARCINOMA NOT DENTIFIED. METASTATIC CARCINOMA PRESENT IN 1 OF 7 PARABRONCHIAL LYMPH NODES. HAVE BRONCHIAL AND VASCULAR MARGINS FREE OF TUMOR. 4: STATION 7: BENIGN LYMPH FRAGMENTS OF BENIGN LYMPH NODE. 6: AP WINDOW LYMPH NODE: ODE. 5: L10: RAGMENTS OF BENIGN LYMPH NODE. Electronically Signed Out By \*\*NAME[XXX], M.D., \*PLACE</Item>

<Item naaccrId="textPathGrossPathology">Gross Description 1: Received fresh for frozen diagnosis designated \"left upper lobe pulmonary nodule\" is a wedge resection of lung 10.5 m up to 5.7 m up to 2.0 m. Starle line of closure is present. One

In view mode, an annotator can see the abstracted values on the left compared with the original document text on the right.

#### Highlight supporting text in source document

ClassifiedDiseaseGrou	p: lung	1	**DATE[Jan 18 2012] Addendum Comment See separately scanned report regarding EGFR Mutation Analysis from Genzyme Laboratories in electronic medical file. Their
🚍 ALK Test			patients with non-small-cell lung cancer and without identifiable mutations are reported to be responsive to EGFR tyrosine kinase inhibitor therapies. Reviewing Pathologist:
ALK test performed:	Yes	1	**NAME[OOO NNN], Ph.D. **INITIALS This certifies that I have reviewed and electronically signed this report.
ALK test method:	FISH-based	1	**NAME[MMM], M.D., **NAME[LLL] SUPPLEMENTAL REPORT Date Ordered: **DATE[Jan 24 2012] Status: Signed Out Date Complete: **DATE[Jan 24 2012] By:
ALK test result:	Negative	1	**NAME[KKK, JJJ III] Date Reported: **DATE[Jan 24 2012] Addendum Comment See separately scanned FISH report from Genzyme Laboratories in the electronic
EGFR Test			medical records file. Their report in part: Negative for a rearrangement involving the ALK gene. Three and four copies of ALK were observed in 72.0% of cells, suggesting the presence of a neoplastic cell population with gains of chromosome 2 or 2p. Signed:

Clicking on a field result highlights all of the supporting strings of text in the document that provide evidence for the value (label) selected

#### 🚯 CDAP Free the Data Portal

Louisianna (LA)

Louisianna (LA)

Datasets

### "Free the Data" Portal

QSearc	ch CDAP Fr	h CDAP Free the Data Portal					
		Help 🕶	depuy 🕶				
Datasets	Tasks	Docume	nt Level				

Task	Schema	Workflow	Document Type	Dataset Link
ALK/EGFR Task1	D metadata.json	ALK and EGFR Path Reports	Pathology	LA_abstracts_abstracts_ALKEGFR_batch1.nlp.zip
ALK/EGFR Di	C metadata.json	ALK and EGFR Path Reports	Pathology (Encrypted)	LA_abstracts_abstracts_ALKEGFR_batch1.nlp.zip.gpg
			Manually coded path report fields	LA_package1_manually-coded-path-report-fields.06-07-2017 05.46.49PM.csv
			Manually coded path report fields (Encrypted)	LA_package1_manually-coded-path-report-fields.06-07-2017 05.46.49PM.csv.gpg
			Path Report XML	LA_package1_path-report-xml-ext13.20170607175443.zip
			Path Report XML (Encrypted)	LA_package1_path-report-xml-ext13.20170607175443.zip.gpg
			CTC Data	LA_package1_ctc-data_20170607175425.csv
			CTC Data (Encrypted)	LA_package1_ctc-data_20170607175425.csv.gpg
			LA DOE eRad	LA_DOE_e-Rad.zip.gpg

#### **Complex Annotation Workflow**



## Lessons Learned



#### Lessons Learned

- Have not scaled up volume yet; still scaling for complexity (variety), data quality (veracity), and velocity
- Looked at many annotation tools; none build for this scale
- Engaging LabKey more as a partner than a vendor

## Results





#### Where we are now

- Multiple instances of LabKey up and running in secure enclaves for annotating documents from three SEER cancer registries
- 3 batches of ALK/EGFR annotated
- Enhancements in-progress

## Enhancements and Support Completed for this project

- 1. Development of Batch Assignment of Cases
- 2. Development of Alternate Metadata Location Configuration
- 3. Development of NCI Folder Template
- 4. Development to Improve XML Document Display
- 5. Development to Specify Metadata Location, Document Type, Disease Group via Protocol
- 6. Development of Ability to Have Second and Third Review Step of Approved Results
- 7. Provision of Help Desk Support

#### Planned Phase 2 Enhancements <u>In-Progress</u>

#### **Structured Abstract Result Export**

Expert Services Provider shall implement a feature for authorized researchers to download the results associated with a particular annotation task in a structured format (LabKey's JSON format) so that is easy to import the data into external tools.

#### Process to Remove Non-printable Characters when Importing Text Documents

Expert Services Provider shall implement a process that removes any non-printable characters (e.g., anything less than ASCII 31) from imported text files.

#### **Custom View Providing Task & Dataset Details**

Expert Services Provider shall implement a feature in the annotation interface that gives authorized users a view of the Tasks and Datasets details in the Natural Language Processing (NLP) pipeline

#### Add Document Counts to Task & Dataset View

Expert Services Provider shall implement a feature in the annotation interface that allows authorized users to view document counts as a part of the Tasks and Datasets custom view in the Natural Language Processing (NLP) pipeline.

#### Select Multiple Values from a Dropdown

Expert Services Provider shall modify fields in the annotation interface to allow abstractors and reviewers to select one or more values from the list provided in the dropdown menu. The selected values must be delimited or maintained as a list of separate values.

# Future plan



#### Next steps

- Annotation of path reports for breast cancer recurrence
- Review and validation of NLP and UQ from DOE labs
  - Site, laterality, histology, grade, behavior
- List of anticipated enhancements to CDAP pipeline in 2018

### Potential Phase 3 Enhancements for 2018

- Integration with NLP algorithms from DOE labs
- Multiple independent annotators with adjudication and inter-annotator agreement reports
- Branching logic for annotation schemas
- Case-level annotation (multiple source documents)
- Refinements to annotation workflow and UI
- Refinements to the administration workflow and UI
- Annotation with hierarchical vocabulary in addition to by defined fields
- Annotation of relations
- Document lifecycle reporting
- Making workflows more automated and robust

## Acknowledgements

- NCI Staff
  - Jessica Boten
  - Donna Rivera
  - Marina Matatova
  - Spencer Morris
- IMS
  - Linda Coyle
  - Glenn Abastillas
- DOE labs

- LabKey
  - Kristin Dubrule
  - Adam Rauch
  - Michael Gersch
- SEER Registries
  - Seattle, LA, KY, GA
- Other contributors
  - Emily Silgard

#### Questions?

## Paul Fearn paul.fearn@nih.gov





www.cancer.gov/espanol

www.cancer.gov