

Update on Virtual Tissue Repository Initiative, De-identification efforts and Genomics projects

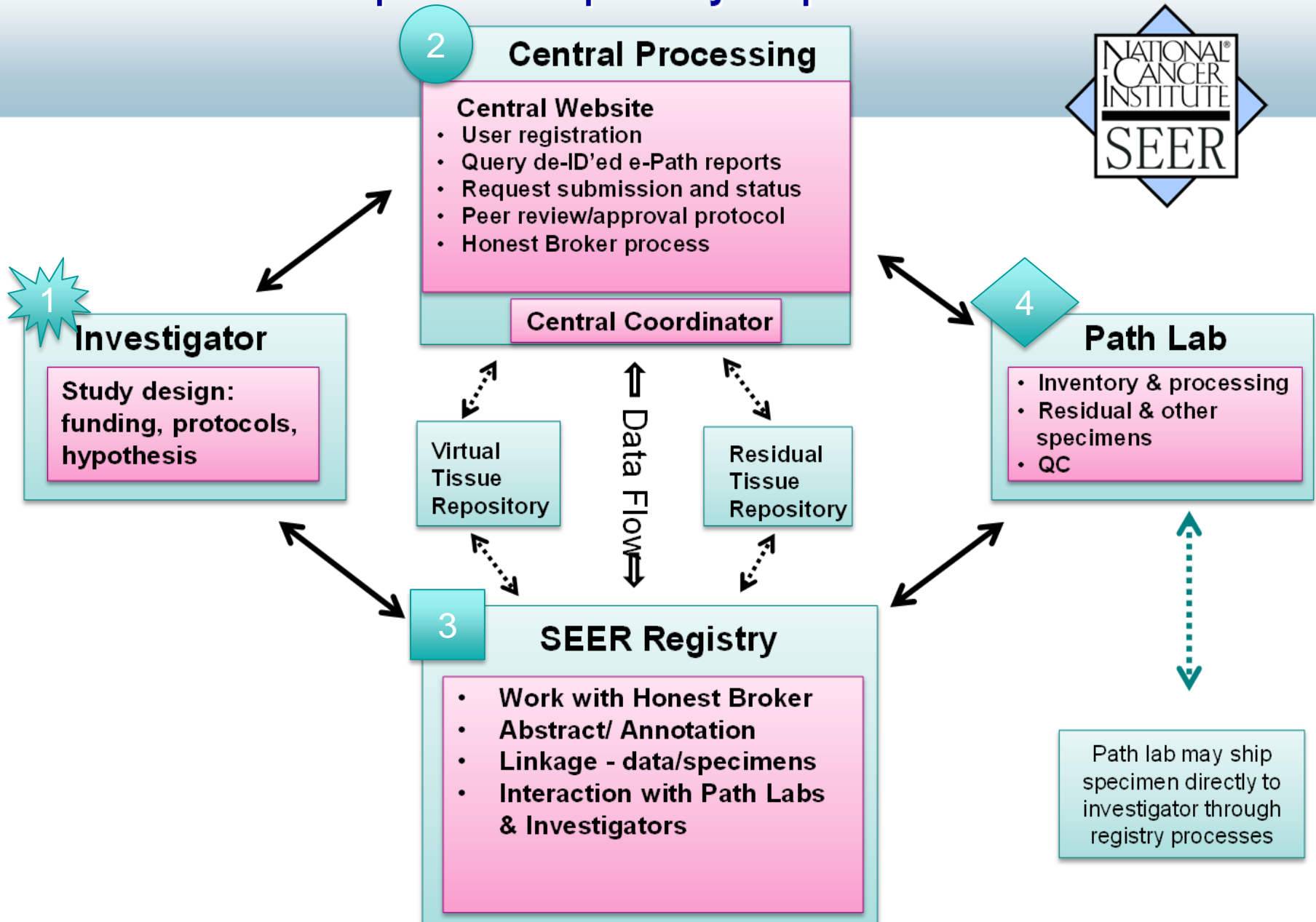
*SEER*DMS F2F meeting*

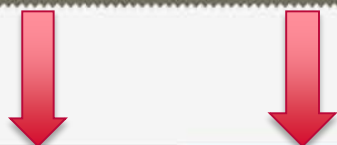
Valentina Petkov, MD, MPH

VTR: the big picture



SEER Biospecimen Repository Proposed Workflow





Histology

- 8140/3: Adenocarcinoma, NOS (# of cases = 3,498)
- 8500/3: Infiltrating duct carcinoma, NOS (C50._) (# of cases = 1,721)
- 8070/3: Squamous cell carcinoma, NOS (# of cases = 894)
- 8720/2: Melanoma in situ (C44._) (# of cases = 520)
- 8720/3: Malignant melanoma, NOS (C44.0_) (# of cases = 458)
- 8000/3: Neoplasm, malignant (# of cases = 370)
- 8742/2: Lentigo maligna (C44._) (# of cases = 253)
- 8130/2: Papillary transitional cell carcinoma, non-invasive (C65.9, C66.9, C67._) (# of cases = 248)
- 8010/3: Carcinoma, NOS (# of cases = 186)

Find... Cancel **Filter Results**

Enter search

- Primary site
- Histology**
- Sex
- Vital status
- Race
- Age at diagnosis
- Spanish hispanic origin
- IHS link
- Dx. confirmation
- Rx. summary radiation
- SEER summary stage 2000
- Sequence number
- Type of reporting source
- Rx. summary surgery primary site

Search 6 — Create Request — Clear

Filter

				Race	Request
				White	<input type="checkbox"/>
				White	<input checked="" type="checkbox"/>
				White	<input checked="" type="checkbox"/>
				White	<input type="checkbox"/>
Q	Female	83	8520/3: Lobular carcinoma, NOS (C50._)	White	<input type="checkbox"/>
Q	Female	79	8380/3: Endometrioid carcinoma (C56.9)	White	<input type="checkbox"/>
Q	Male	71	8140/3: Adenocarcinoma, NOS	Unknown	<input type="checkbox"/>
Q	Female	54	8430/3: Mucoepidermoid carcinoma	Unknown	<input type="checkbox"/>
Q	Male	79	8550/3: Acinar cell carcinoma	White	<input type="checkbox"/>
Q	Female	66	8201/2: Cribriform carcinoma in situ (C50._)	White	<input type="checkbox"/>

Found 16,577 results in 63 milliseconds.

Reset Search

SEER BioShare

https://www92.imsweb.com/search/?q=+pathology_reports.%5C%2A%3A%28ki67%29

National Cancer Institute

SEER BioShare
Turning Cancer Data Into Discovery

Home About

Home / Search

pathology_reports.*(ki67)

Primary site
Histology
Sex
Vital status
Race
Age at diagnosis
Spanish hispanic origin
IHS link
Dx. confirmation
Rx. summary radiation
SEER summary stage 2000
Sequence number
Type of reporting source
Rx. summary surgery primary site
Has pathology reports?

Reset Search

at the National Institutes of Health | www.cancer.gov

Sean Brennan

Reports Admin

ed Search 0 — Create Request — Clear

Filter

Race Pathology Reports Request

Black
White
White
White
White

Pathology Reports

Patient Display ID PAT-20089081

Tumor Record Number 02

Record Document ID REC-3001764019

Clinical History 1. IDC 2. Fibroadenomatous/fibrocystic change. Cancer? (less likely)

Comments

Formal DX 1. Left breast, 11 o'clock, biopsy: Invasive ductal carcinoma, predicted Bloom-Richardson Score 7 (tubule formation 3, nuclear pleomorphism 3, mitosis 1). Note: the tumor is positive for E-Cadherin, and the **Ki67** labeling index is approximately 90%. ER, PR and Her2 are ordered. 2. Right breast, 11 to 12 o'clock, biopsy: Invasive ductal carcinoma, predicted Bloom-Richardson Score 8 (tubule formation 3, nuclear pleomorphism 2, mitosis 3). Note: the tumor is positive for E-Cadherin, and the **Ki67** labeling index is approximately 50%. ER, PR and Her2 are ordered. **INITIALS

M85003 M85203 P1140 T04030 M85003 P1140 T04020

Full Text

Gross Pathology 1. Received in formalin, labeled with the patient's name and medical record number, and designated as "left breast", are two fibrofatty and hemorrhagic core needle biopsies measuring 1.9 and 1.7 cm in length. The specimen is submitted entirely in a single cassette. 2. Received in formalin, labeled with the patient's name and medical record number, and designated as "right breast", are multiple fragments of yellow lobulated fatty tissue ranging in size from 0.1 to 1.2 cm in length. The specimen is strained and entirely submitted in a single cassette. **INITIALS

Microscopic Description

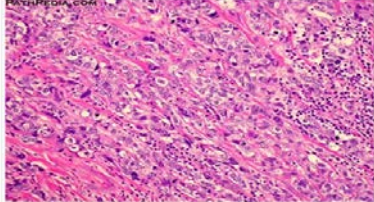
Pathology Reports

Patient Display ID PAT-20089081

Tumor Record Number 02

Record Document ID REC-3001764019

Clinical History 1. IDC 2. Fibroadenomatous/fibrocystic change. Cancer? (less likely)



M85003 M85203 P1140 T04030 M85003 P1140 T04020

To do

- Identify reliable de-identification software and incorporate it with SEER*DMS
- Finish the VTR pilot in 7 registries
- Obtain funding for the scaled program
- Establish VTR policies and procedures



VTR pilot in 7 SEER registries

Objectives

- To inform us in establishing best practices
 - Can the registries do it?
 - Registry regulatory requirements (IRB approvals, MTAs, DUA, etc)
 - Pathology labs regulatory issues
 - Retrieval and processing of specimen
 - Detailed clinical annotation
 - Effort and cost at each step

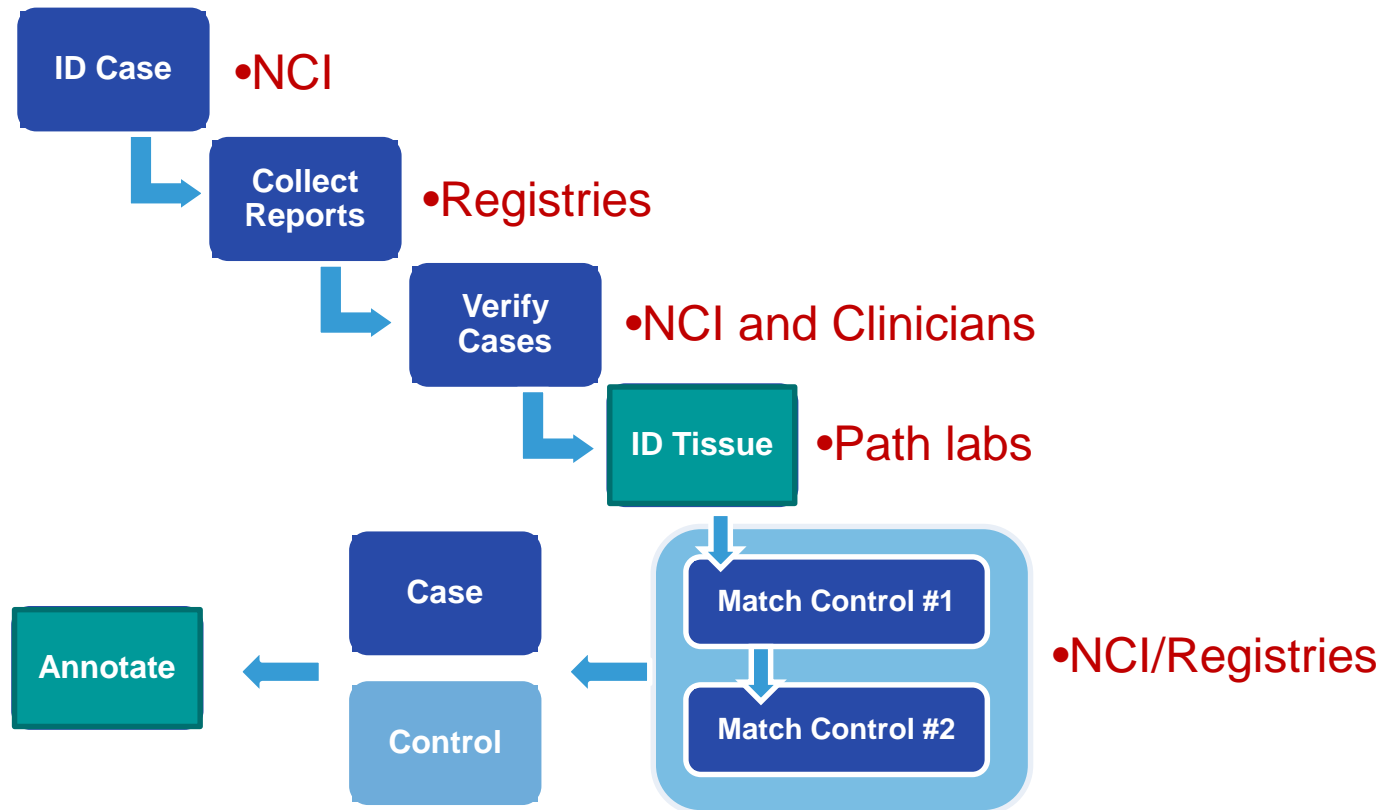
Methods

- RRSS in 7 SEER registries: GrCA, CT, HI, KY, IA LA, UT
- Pathology inventory: 42 item web-based questionnaire to local pathology labs – completed
 - Storing/sharing biospecimens
 - Sharing/providing histology slides
 - Digitization of images
 - Terms of release for research

Methods (cont)

- Two use cases: case-control matched study design
 - Study 1: Unusual outcome in early stage breast cancer (LN0)
 - Cases: < 30 mo survival w COD=BC
 - Controls > 60 months survival
 - Matched deterministically on HR status and probabilistically on age, race, year of dx, tumor size, histology, radiation, number of LN examined
 - Study 2: Unusual outcome in pancreatic adenocarcinoma
 - Cases: > 60 months survival
 - Controls < 24 month survival w COD=PC
 - Matched deterministically on mets and LN status and probabilistically on age, race, gender, anatomical location, radiation therapy

SEER-VTR Pilot Workflow



Custom annotation of biospecimen

- Detailed systemic therapy (agents, dose, frequency, duration)
- Radiation therapy
- Co-morbidities
- Biomarkers

PANCREAS

Abstractor Comments (confidential):

Complete the form for all drugs of interest, include single and total dose dispensed, dose units, and frequency of administration

Systemic Therapy Agent 1

Drug Other Dose Schedule Flag

Single Dose Dose Units Other

Route of Administration Other Dose Frequency Other

Total Dose Prescribed Start Date / / End Date / /

Code	Description
00	None given
01	5-Fluorouracil (5-FU)
02	Capecitabine (Xeloda)
03	Cisplatin (CDDP, Platino)
04	Oxaliplatin (Eloxatin)
05	Epirubicin (Ellence)
06	Mitomycin
07	Irinotecan (Camptosar)
08	Gemcitabine

Current status:

- Determination of tissue availability – 95% completed
- Custom annotation - 25% completed
- Need additional cases and controls
- Timeline: 9/2017 - 9/2018



Substudy: Digital imaging

- Collaboration with CBIIT, Emory and Stony Brook universities
- Objectives:
 - Can registries successfully collect and transfer images
 - Incorporation with image viewer/ image analysis software
 - Feature extraction – nuclear morphology and lymphocyte infiltration
- 5 participating registries
- 700 images
- Current status: 130 images collected and transferred to IMS and Emory

Substudy: Genomic sequencing

- Pancreatic cancer
- Sponsored by PanCAN
- WES on 100 case-control pairs performed by a commercial lab
- Clinical and sequencing data will be stored at IMS
- Ultimate goal is to make the data available to the larger research community (Genomic Data Commons/ dbGap)
- Current status: protocol developed; IRB submissions
- Timeline
 - Sequencing 7/17-7/18
 - Initial evaluation of data and analysis 7/18-7/19
 - Data available to research community: 2020

Acknowledgements

SEER Registries (GC, CT, HI, IA, LA, KY and UT)

IMS

NCI/SRP: L Penberthy, V Petkov, S. Hussey, M Matatova, S Friedman, A. Wang, M Yu, P Fearn, former: S. Altecruse, R Moravec, J Botten

NCI/other E. Gillander, D. Carrick, Ed Helton, Ulrike Wagner

PanCAN

Emory U: Ashish Sharma

Stoney Brook U: Joel Saltz

SEER Evaluation of De-identification tools

Two studies



De-identification evaluation protocol

- 5 SEER Registries: CT, HI, KY, NM, and Seattle
 - IRB approvals
- Pathology report selection
 - 4000 randomly selected from reports received in 2011
 - 800/registry
 - Stratified by cancer site
 - 160 each: breast, lung, crc, prostate and other
- IMS provided technical instructions
- Each registry performed the de-identification
- Reviewed and compared de-id tool output to original report
- Recorded number of occurrences PII was missed by PII category
- Automated count of de-id phrases by PII category

Performance measurement

- De-identification rate
 - PII phrase level
 - N de-identified phrases/All PII phrases
 - PII at patient level
 - N patients w/ missed PII/800
 - Calculated per each PII category, overall and per registry
- Limitations
 - N de-id phrases counted based on PII tag (includes over scrubbing)
 - De-id rates for names of patients and providers cannot be calculated separately

DE-ID™

<http://www.de-idata.com/>

Performance of De-ID™ in five SEER registry

PHI type	De-Id phrases N	Missed phrases N	All PHI phrases	PII phrase DeID rate	N pts w/ missed PII	Pt level DeID rate
Names	13030	88	13118	0.993	19	0.995
Dates	8717	31	8748	0.996	23	0.994
Phone Numbers	909	0	909	1.000	0	1.000
Places	1532	0	1532	1.000	0	1.000
Street Addresses	350	10	360	0.972	7	0.998
Zip Codes	844	0	844	1.000	0	1.000
ID Numbers	1358	77	1435	0.946	51	0.987
Total PHI	26740	206	26946	0.992	100	0.975
Path Numbers	1678	1310	2988	0.562	810	0.798
Institutions	1355	1673	3028	0.447	825	0.794
Total de-id info	29773	3189	32962	0.903	1735	0.566

NLM scrubber

Beta Version tested

<https://scrubber.nlm.nih.gov/>

Performance of NLM scrubber in CT SEER registry

NLM scrubber tags	N phrases de-id	N phrases missed	Total N phrases	N patients not de-id	De-id rate phrases	De-id patients
Personal name pt name+provider name	5130	0+8	5138	0	0.998	1.000
Address	466	1	467	1	0.998	0.999
Alphanumeric ssn+mrn+phone+path#	1420	0+0+0+179	1599	77	0.888	0.901
Date	1393	1	1394	1	0.999	0.999
Total	8409	189	8598	79	0.978	0.899

Performance of NLM scrubber in HI SEER registry

NLM scrubber tags	N phrases de-id	N phrases missed	Total N phrases	N patients not de-id	De-id rate phrases	De-id patients
Personal name pt name+provider name	6783	29+35	6847	13	0.991	0.984
Address	356	0	356	0	1.000	1.000
Alphanumeric ssn+mrn+phone#+path#	1057	0+0+0+5	1062	3	0.995	0.996
Date	883	1	884	1	0.999	0.999
Total	9079	69	9149	17	0.992	0.979

Other tools

- *PARAT, Privacy Analytics*
- *MIST, MITRE*

Summary

- Reasonable performance for PII (with the exception of Seattle and to a lesser degree HI)
- Suboptimal for Institution and pathology specimen IDs
- Inconsistency across reports and registries
 - De-ID within a report
- Registries opinion: **generally not satisfied**
 - KY and CT: NLM scrubber performed better and more user friendly
 - Seattle: both tools performed the same; NLM easier to use
 - HI and NM: performance the same

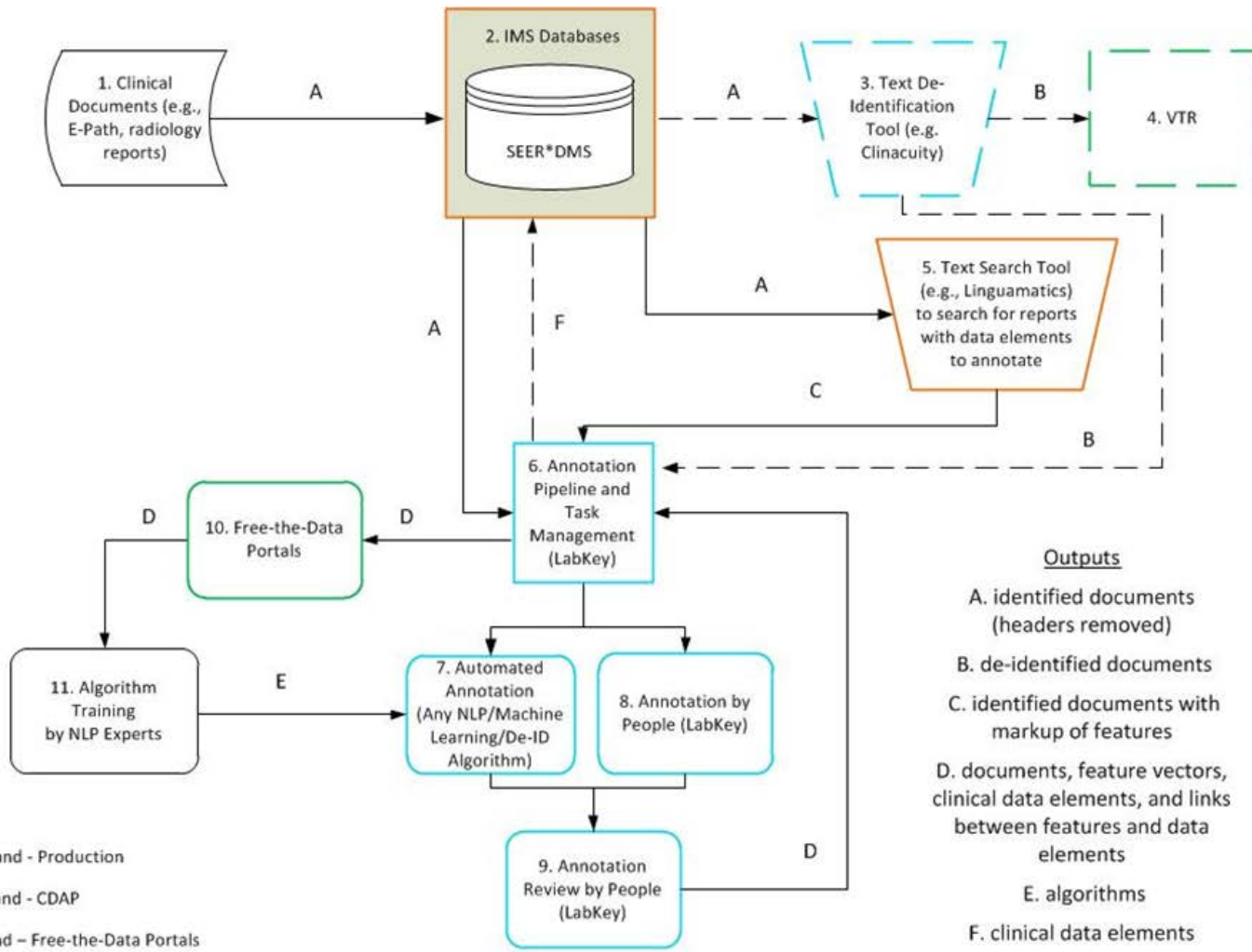
Next steps

- PII annotation on representative sample of ePath reports
- Testing high-potential de-identification tools
 - Latest version of NLM scrubber
 - BoB

PII Annotation Protocol for Narrative Clinical Text

- Annotation of PII - all PII is clearly marked and categorized in the text
- CDAP pipeline will be used for annotation
- Each registry will annotate a sample of reports
- PII annotated reports will be used for:
 - Customization and training of de-identification tools
 - Validation/testing of the tools prior to deployment
 - Validation/testing each time major revisions/versions of the tools are introduced

CDAP



Annotation Process

https://zseercdap-labkey.seerdm.com/labkey/NCI%20CDAP%20pipeline/nlp-view... Check Point Mobile - Main Pathology Report - 01-REC... Welcome to AgilQuest's OnBo...

LabKey Server Search LabKey Server

NCI CDAP pipeline NLP D

Pathology Report - 01-REC-3000639415

Field Results

Dates

Date: Apr 20, 2013

Names

First Name: Edward

Last Name: Simms

Other Name: R.

Addresses

Number:

Street:

City: Medford

State/Province/Country: OR

Postal Code:

Size: 0.9 x 0.7 x 0.5 cm (black) and 1.2 x 1 x 0.7 cm (blue)

Submitted/Blocks: Entirely/2./r/n</Item>

<Item naaccrId="textPathMicroscopicDesc">MicroscopicDescription:

A. Ductal carcinoma in situ is seen involving the entire specimen with multiple foci (4) of microinvasive (less than 1 mm) ductal carcinoma. The ductal carcinoma in situ is seen focally involving the superior and superficial surgical margins. In addition there is crush artifact at the surgical margin. The DCIS shows clinging/micropapillary and papillary patterns with intermediate nuclear grade (Grade II). Focal reaction compatible with previous biopsy site is seen.

B. Microscopic evaluation was performed. Final diagnosis was rendered based on gross and microscopic findings.

<Item naaccrId="textPathNatureOfSpecimens">A: Right breast tissue, excision (fresh)

B: Right breast tissue at 9 o'clock, excision (fresh)</Item>

<Item naaccrId="textPathSuppReportsAddenda">ResultsComments:

The following results were performed at Medford, OR and reported by Edward R. Simms, M.D. on Apr 20, 2013.

INTERPRETATION:

BREAST CANCER PROGNOSTIC PANEL:

*1, BLOCK # A7 (INVASIVE CARCINOMA)

ESTROGEN RECEPTOR:	90%
PROGESTERONE RECEPTOR:	83%
HER-2/neu (ACIS score):	0.8 (NO OVEREXPRESSION)

COMMENT:

ERPR

Analysis is performed using ChromaVision Automated Cellular Imaging System (ACIS) on formalin-fixed paraffin-embedded section stained by immunohistochemical methods on the Ventana Benchmark XT automated stainer using antibodies against ER (SP1 IVD), PR (clone IE2 IVD). Though the largest studies have used 10% as a threshold for positivity, others have recommended a cutoff as low as 1%.

Annotation schema

- All 18 HIPAA Safe Harbor identifiers
- Institution/Medical practice/Laboratory name and address
- Pathology report/specimen/slide number

Registry selection

- All registries are eligible to participate
 - Registry decision
- **Benefits**
 - Tool customization will take into account registry specific variability
 - The same set of reports can be used for assessment of multiple tools and later versions of tools
 - Annotation by preset rules will allow for comparability across registries and tools
- **Costs**
 - Will require some time investment at the registry
 - Training (1-2 hours)
 - Annotation of 100 documents is estimated at 17 hours but can vary

Proposed metrics/goals

- Patient name: > 99%
- Other names (relatives; providers, etc.): > 98%
- SSN: 100%
- Dates: > 98%
- Other identification numbers (MRN, account #, insurance plan #): > 99%
- Patient address (street, city, zip code): > 98%
- Patient phone, fax, email, URL: > 99%
- Specimen/slide/path report #: > 97%
- Institution/lab name: > 97%
- Institution address: > 97%

Resources



- NISTIR 8053: De-Identification of Personal Information (Oct. 2015)
 - <http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf>
- NIST Special Publications 800-188: De-Identifying Government Datasets (second draft, Dec. 2016)
 - http://csrc.nist.gov/publications/drafts/800-188/sp800_188_draft2.pdf

Acknowledgment:

SEER registries: CT, HI, KY, NM, and Seattle

NCI team: Morris Spencer, Paul Fern, Steve Friedman, Lynne Penberthy

IMS team: Rusty Shields, Dave Annette, Laurie Buck, Linda Coyle

NIH/NLM: Mehmet Kayaalp

USC: Stephane Meystre

Tumor genomics and germline mutations



Overview

- Importance to SEER
- Both tumor genome and germline mutations are determinants of response to therapy and outcomes
- Issues with current collection of BMs as standard data elements
 - Limited to few BMs
 - Quality: completeness and accuracy
 - Rapid change in landscape and time lag
- SEER plan: tumor genomics and germline mutations to be collected as part of regular cancer surveillance
 - Mostly in automated ways

Oncotype DX linkage

- Currently in third year - plan to finish by the end of August
- Data on 21-gene assay available to researchers as a specialized data set
- 16-gene assay data analyzed currently
- Assessment will determine data release policy
- Incorporation in SEER-Medicare: MOU
- Ongoing research collaboration with Genomic Health on research projects, presentations and articles

GA-CA genetic linkage (Genlink study)

- Primary objective: to determine the feasibility of collecting germline mutations for cancer surveillance
- IRB approved study in 4 registries
- Breast and ovarian cancer cases diagnosed 2013-2015 (>100,000)
- Linked to single or multipanel germline mutations tests
- 4 labs (Myriad, Invite, Ambry, and GeneDX)
- Labs provided 1.5 million records for 1.1 million persons
- 26% of SEER cases successfully linked
- De-identified data set is currently analyzed
- Will be available to researchers through central registries
- 2017 linkage to capture fully 2015 dx year
- Plans to scale to SEER program
 - 2018 linkage will be open to all SEER registries that can collect these data as part of regular cancer surveillance

Collaboration with Syapse

- IT company that harmonizes genomic data across labs, integrates them with clinical data, displays the data in chronological and structured way, link targetable genes/mutations to available drugs both for standard of care and clinical trials
- Pilot in GA
 - Conducted as cancer surveillance activity
 - Link data from 2 genomic labs, preferably multipanel tests
 - Gardient360 – 70 gene liquid bx test covering all actionable gene mutations

Other projects

- Foundation Medicine
 - FoundationOne (solid tumors): >300 genes in all 4 classes of alteration for solid tumors plus MSI and tumor mutational burden
 - FoundationOne Heme: > 400 genes interrogated and >250 RNA sequence genes
 - FoundationACT (>60 genes; liquid bx)
 - FoundationFocus CDxBRACA: first FDA approved companion dx for both germline and somatic BRCA mutation in Ovarian ca-response to PARP inhibitors
- Prostate Biomarkers - 3 major players
 - Prolaris test (Myriad)
 - Decipher (GenomeDX)
 - OncotypeDX (Genomic Health)



**NATIONAL
CANCER
INSTITUTE**

www.cancer.gov

www.cancer.gov/espanol