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

1. Investigator
   - Study design: funding, protocols, hypothesis

2. Central Processing
   - Central Website
     - User registration
     - Query de-identified e-Path reports
     - Request submission and status
     - Peer review/approval protocol
     - Honest Broker process

3. SEER Registry
   - Work with Honest Broker
   - Abstract/Annotation
   - Linkage - data/specimens
   - Interaction with Path Labs & Investigators

4. Path Lab
   - Inventory & processing
   - Residual & other specimens
   - QC

Path lab may ship specimen directly to investigator through registry processes.

Flow Diagram:
- Investigator to Central Coordinator
- Central Coordinator to Central Processing
- Central Processing to SEER Registry
- SEER Registry to Path Lab
<table>
<thead>
<tr>
<th>Race</th>
<th>Request</th>
<th>Histology</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td></td>
<td>8140/3: Adenocarcinoma, NOS (# of cases = 3,498)</td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>8500/3: Infiltrating duct carcinoma, NOS (C50..) (# of cases = 1,721)</td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>8070/3: Squamous cell carcinoma, NOS (# of cases = 894)</td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>8720/2: Melanoma in situ (C44..) (# of cases = 520)</td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>8720/3: Malignant melanoma, NOS (C44.0..) (# of cases = 458)</td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>8742/2: Lentigo maligna (C44..) (# of cases = 253)</td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>8130/2: Papillary transitional cell carcinoma, non-invasive (C65.9, C66.9, C67..) (# of cases = 248)</td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>8010/3: Carcinoma, NOS (# of cases = 186)</td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>8520/3: Lobular carcinoma, NOS (C50..)</td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>8380/3: Endometrioid carcinoma (C56.9)</td>
</tr>
<tr>
<td>Unknown</td>
<td></td>
<td>8140/3: Adenocarcinoma, NOS</td>
</tr>
<tr>
<td>Unknown</td>
<td></td>
<td>8430/3: Mucoepidermoid carcinoma</td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>8550/3: Acinar cell carcinoma</td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>8201/2: Cribriform carcinoma in situ (C50..)</td>
</tr>
</tbody>
</table>

Found 16,577 results in 63 milliseconds.
### Pathology Reports

<table>
<thead>
<tr>
<th>Patient Display ID</th>
<th>PAT-20089081</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor Record Number</td>
<td>02</td>
</tr>
<tr>
<td>Record Document ID</td>
<td>REC-3001764019</td>
</tr>
<tr>
<td>Clinical History</td>
<td>1. IDC 2. Fibroadenomatous/fibrocystic change. Cancer? (less likely)</td>
</tr>
<tr>
<td>Comments</td>
<td></td>
</tr>
<tr>
<td>Formal DX</td>
<td>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 K677 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 K677 labeling index is approximately 50%. ER, PR and Her2 are ordered. <strong>INITIALS</strong></td>
</tr>
</tbody>
</table>

**Full Text**

- M85003 M85203 P1140 T04030 M85003 P1140 T04020

**Gross Pathology**

- 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 stained and entirely submitted in a single cassette. **INITIALS**
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

- **ID Case**
  - NCI
- **Collect Reports**
  - Registries
- **Verify Cases**
  - NCI and Clinicians
- **ID Tissue**
  - Path labs
- **Annotate**
- **Case**
- **Control**
  - Match Control #1
  - Match Control #2
  - NCI/Registries
Custom annotation of biospecimen

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

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

Beta Version tested

https://scrubber.nlm.nih.gov/
## Performance of NLM scrubber in CT SEER registry

<table>
<thead>
<tr>
<th>NLM scrubber tags</th>
<th>N</th>
<th>N</th>
<th>Total N</th>
<th>N</th>
<th>De-id</th>
<th>De-id</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>phrases</td>
<td>phrases</td>
<td>phrases</td>
<td>patients</td>
<td>rate phrases</td>
<td>patients</td>
</tr>
<tr>
<td>Personal name pt name+provider name</td>
<td>5130</td>
<td>0+8</td>
<td>5138</td>
<td>0</td>
<td>0.998</td>
<td>1.000</td>
</tr>
<tr>
<td>Address</td>
<td>466</td>
<td>1</td>
<td>467</td>
<td>1</td>
<td>0.998</td>
<td>0.999</td>
</tr>
<tr>
<td>Alphanumeric ssn+mrn+phone+path#</td>
<td>1420</td>
<td>0+0+0+179</td>
<td>1599</td>
<td>77</td>
<td>0.888</td>
<td>0.901</td>
</tr>
<tr>
<td>Date</td>
<td>1393</td>
<td>1</td>
<td>1394</td>
<td>1</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>Total</td>
<td>8409</td>
<td>189</td>
<td>8598</td>
<td>79</td>
<td>0.978</td>
<td>0.899</td>
</tr>
</tbody>
</table>
## Performance of NLM scrubber in HI SEER registry

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

Pathology Report - 01-REC-3000639415

Size: 0.5 x 0.7 x 0.5 cm (black) and 1.2 x 1 x 0.7 cm (blue)
Submitted/Blocks: Entirely/2.1/3

Microscopic Description:
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 surg is within a millimeter of all other surgical margins. In addition there is crush artifact at the surgical margin. Involve ment cannot be entirely excluded. 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.

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

Interpretation:
Breast Cancer Prognostic Panel:
+1, Block 7 A7 (Invasive Carcinoma)

Estrogen Receptor: 90%
Progesterone Receptor: 83%
HER-2/neu (ACIS score): 0.8 (No Overexpression)

Comment:
ER/PR
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)

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)