

De-identification of unstructured (narrative text) clinical documents: Importance and Challenges

*SEER*DMS Face to Face meeting*

Rockville, MD, September 26-28, 2018

Valentina Petkov, MD, MPH

NCI/DCCPS/Surveillance Research Program

Presentation Outline

- Background – need for de-identification
- SEER evaluation of de-identification tools
- Summary of findings
- Proposed next steps

Background



Resources



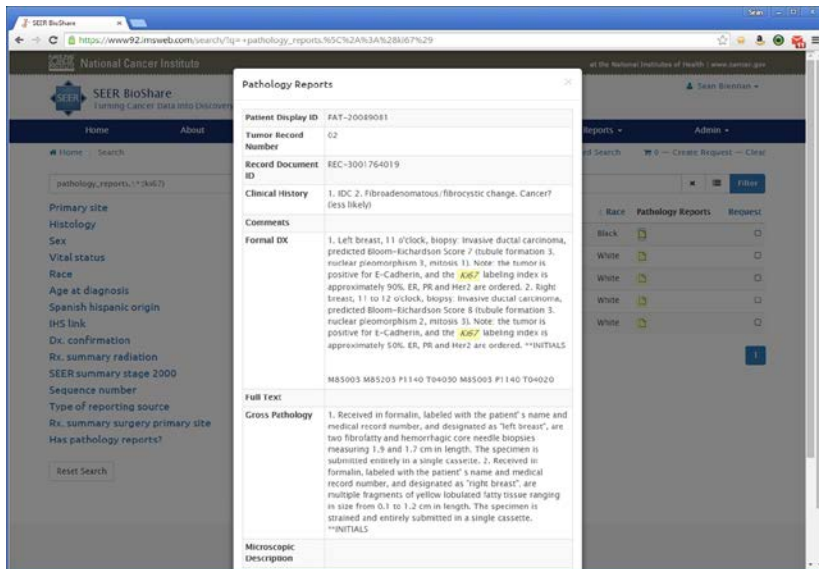
- NISTIR 8053: De-Identification of Personal Information (Oct. 2015)
 - <http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf>
- NIST Special Publications 800-188: De-Identifying Government Datasets (second draft, Dec. 2016)
 - http://csrc.nist.gov/publications/drafts/800-188/sp800_188_draft2.pdf

Importance to research

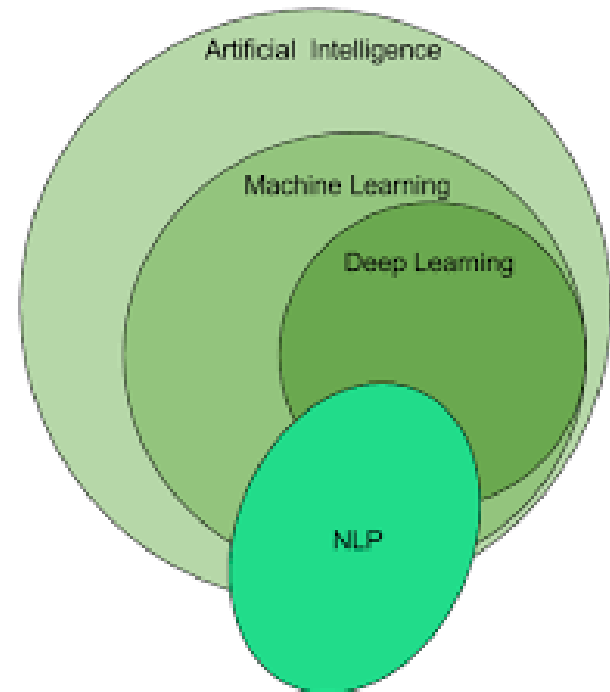
- Majority of relevant data from EMRs is in unstructured text format (estimated >65%)
- SEER registries collect increasing amount of clinical text document (Epath, radiology reports)
- There is need for researchers to access these unstructured text documents for
 - Capturing structured data
 - Access to large volumes of unstructured text documents to develop deep learning algorithms to enhance the ability to capture structured data without manual effort

Importance to SEER

- SEER linked Virtual Tissue Repository (VTR)
 - SEER VTR BioShare



- Automated abstraction of data from narrative clinical documents
 - NCI-SEER-DOE collaboration
 - DeepPhe



SEER Evaluation of De-identification tools

Two studies



De-identification evaluation protocol

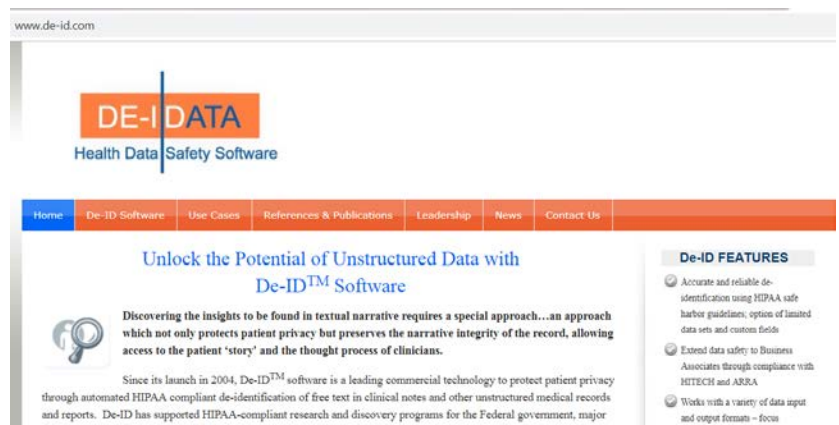
- 5 SEER Registries: CT, HI, KY, NM, and Seattle
 - IRB approvals
- Pathology report selection
 - 4000 randomly selected from reports received in 2011
 - 800/registry
 - Stratified by cancer site
 - 160 each: breast, lung, crc, prostate and other
- IMS provided technical instructions
- Each registry 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 and 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/>



The screenshot shows the homepage of the DE-ID website. At the top left, the URL "www.de-id.com" is displayed. The main header features the "DE-ID DATA" logo in orange and blue, with the tagline "Health Data Safety Software" below it. A navigation menu in orange includes links for Home, De-ID Software, Use Cases, References & Publications, Leadership, News, and Contact Us. The main content area has a heading "Unlock the Potential of Unstructured Data with De-ID™ Software" and a sub-heading "Discovering the insights to be found in textual narrative requires a special approach...an approach which not only protects patient privacy but preserves the narrative integrity of the record, allowing access to the patient 'story' and the thought process of clinicians." Below this is a paragraph about the software's history and use. On the right, a "De-ID FEATURES" section lists three key capabilities: accurate and reliable de-identification, extended data safety, and compatibility with various data formats.

www.de-id.com

DE-ID DATA
Health Data Safety Software

Home De-ID Software Use Cases References & Publications Leadership News Contact Us

Unlock the Potential of Unstructured Data with De-ID™ Software

Discovering the insights to be found in textual narrative requires a special approach...an approach which not only protects patient privacy but preserves the narrative integrity of the record, allowing access to the patient 'story' and the thought process of clinicians.

Since its launch in 2004, De-ID™ software is a leading commercial technology to protect patient privacy through automated HIPAA compliant de-identification of free text in clinical notes and other unstructured medical records and reports. De-ID has supported HIPAA-compliant research and discovery programs for the Federal government, major

De-ID FEATURES

- Accurate and reliable de-identification using HIPAA safe harbor guidelines; option of limited data sets and custom fields
- Extend data safety to Business Associates through compliance with HITTECH and ARRA
- Works with a variety of data input and output formats – focus

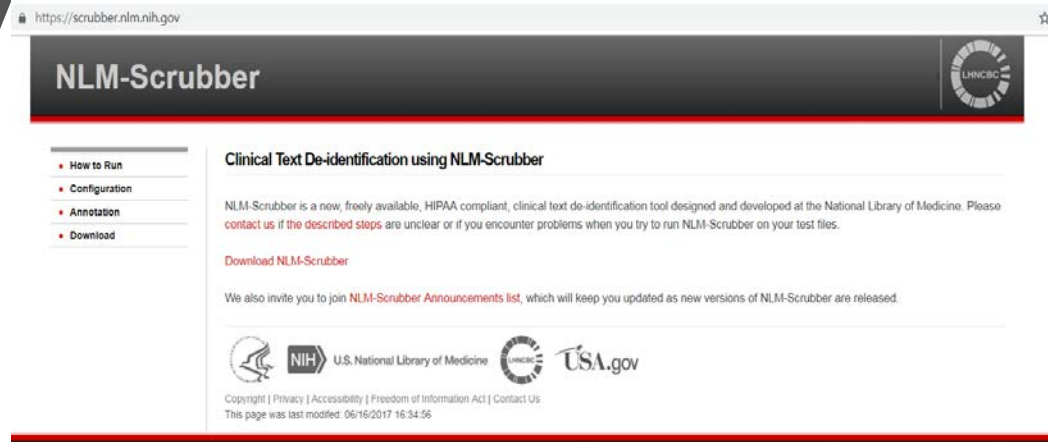
Performance of De-ID™ in five SEER registry

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/>



The screenshot shows the NLM-Scrubber website. At the top, the browser address bar displays <https://scrubber.nlm.nih.gov>. The page header features the title "NLM-Scrubber" in white text on a dark background, with the NLM logo on the right. A left-hand navigation menu lists: [How to Run](#), [Configuration](#), [Annotation](#), and [Download](#). The main content area is titled "Clinical Text De-identification using NLM-Scrubber" and contains the following text: "NLM-Scrubber is a new, freely available, HIPAA compliant, clinical text de-identification tool designed and developed at the National Library of Medicine. Please [contact us](#) if the [described steps](#) are unclear or if you encounter problems when you try to run NLM-Scrubber on your test files." Below this is a link for "Download NLM-Scrubber" and an invitation to join the "NLM-Scrubber Announcements list". The footer includes logos for NIH, U.S. National Library of Medicine, and USA.gov, along with copyright and privacy information: "Copyright | Privacy | Accessibility | Freedom of Information Act | Contact Us" and "This page was last modified: 06/16/2017 16:34:56".

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

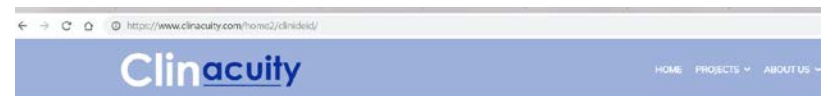
Summary

- 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

Other tools

- MIST, MITRE (<http://mist-deid.sourceforge.net/>)
 - Open source and free
 - Option for replacing with synthetic PII
 - Customized by a Harvard NLP group for clinical documents

- CliniDeID (former BoB, Best of Bread), Clinacuity (<https://www.clinacuity.com/home2/clinideid/>)
 - Option for replacing with synthetic PII



CliniDeID – Automatic clinical text de-identification

A diagram illustrating the CliniDeID de-identification process. It shows two columns of clinical text. The left column contains a medical record snippet with patient identifiers and a name. A red arrow points from this text to a central CliniDeID logo, which consists of a blue shield with a white question mark and a lock icon. Below the logo is the text "CliniDeID". A second red arrow points from the logo to the right column, which shows the same medical record snippet but with the patient identifiers and name replaced by synthetic text.

928701	7/13/2004 10:00:00 AM	Admission Date : 07/03/2004 Discharge Date : 07/12/2004 DISCHARGE DIAGNOSIS : RIGHT BICONDYLAR TIBIAL PLATEAU FRACTURE TIBIAL PLATEAU FRACTURE HISTORY OF PRESENT ILLNESS :Mr. Jones is an otherwise healthy 32 year old male attorney who was vacationing at Schesson Valley when he fell off his moped at a speed of approximately 25 miles per hour . He remembers the accident with no loss of consciousness. He landed on his right knee and noted immediate pain and swelling . He was taken by ambulance to Justice Healthcare where he had plain films that revealed a comminuted bicondylar tibial plateau fracture on the right. He was transferred to the McValley Medical Center for further evaluation and treatment. PAST MEDICAL/SURGICAL HISTORY : 1 10/20/2004/2010	327488	6/17/1994 12:00:00 AM	Admission Date : 06/07/1994 Discharge Date : 06/16/1994 DISCHARGE DIAGNOSIS : RIGHT BICONDYLAR TIBIAL PLATEAU FRACTURE TIBIAL PLATEAU FRACTURE HISTORY OF PRESENT ILLNESS :Mr. First is an otherwise healthy 32 year old male attorney who was vacationing at Robertson Falls when he fell off his moped at a speed of approximately 25 miles per hour . He remembers the accident with no loss of consciousness. He landed on his right knee and noted immediate pain and swelling . He was taken by ambulance to Hasring Healthcare where he had plain films that revealed a comminuted bicondylar tibial plateau fracture on the right. He was transferred to the Mercy Medical Center for further evaluation and treatment. PAST MEDICAL/SURGICAL HISTORY : 1 10/20/2004/2010
--------	-----------------------	--	--------	-----------------------	---

Other tools (cont.)

- Lexicon, Privacy Analytics
(<https://privacy-analytics.com/software/privacy-analytics-lexicon/>)
 - Option for replacing PII with synthetic PII
 - Same performance
 - Used by NIH Clinical Center and ASCO CancerLinQ



- Incognito, AIM
- 2014 i2b2/UTHealth shared task Track 1 challenge
 - Ten systems participated
 - Overall precision varies from 52 to 96%
- MIT researchers recently proposed a system based on deep learning

Journal of the American Medical Association, 2016, 311, 300-300
doi: 10.1093/jamia/ocw156
Advance Access Publication Date: 31 December 2016
Research and Applications



Research and Applications

De-identification of patient notes with recurrent neural networks

Franck Dernoncourt,^{1,*} Ji Young Lee,^{1,*} Ozlem Uzuner,² and Peter Szolovits¹

¹Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA,
²Computer Science Department, University at Albany, SUNY, Albany, NY, USA

Proposed Next Steps



Next steps

- Solicit interest of Cancer Data Ecosystem and Cancer Moonshot Initiative to
 - Identify and make available to researchers reliable and scalable de-id system(s)
 - Testing high-potential de-identification tools
 - Challenge/Hackathon
 - Market research to determine:
 - Current status
 - Interest to participate individually or in collaboration
 - Develop a set of “gold standard” clinical documents (pathology and radiology reports) with annotated and replaced PII to be used in competition/challenge for software customization, testing and

Next Steps: SEER registry role and participation

- Develop a set of “gold standard” clinical documents (pathology and radiology reports) with annotated and replaced PII to be used in competition/challenge for software customization, testing and validation
- 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

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%.

Technical aspects of PII annotation

- Use Clinical Data Annotation and Processing (CDAP) Pipeline
 - Currently available to the 4 registries participating in the NCI-DOE project
 - Same architecture will be replicated for the rest of registries
 - After training, registry staff or NCI contractor will access the system and annotate the reports
 - Annotation schema
 - All 18 HIPAA Safe Harbor identifiers
 - Institution/Medical practice/Laboratory name and address
 - Pathology report/specimen/slide number

Clinical document selection

- Random stratified sample
 - Include all labs and other entities feeding electronic clinical text documents to registries
 - Stratified by time period
 - Stratified report type
- The demographic section (header) will not be included for annotation
 - Could be used by each registry as ontology/reference DB in de-id tool

Metrics/analysis

- Recall (sensitivity) = $[TP/(TP+FN)]$
 - How many identifiers are we capturing (and how many are we missing)?
 - Important to registries and patients

- Specificity = $[TN/(TN+FN)]$
 - How much non-identifying info are we retaining?
 - Important to researchers

- Precision (positive predictive value) = $[TP/(TP+FP)]$
- F-measure = $2*[(Precision*Recall)/(Precision+Recall)]$

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.): > 99%
- 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%

Acknowledgments:

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, Jennifer Stevenson

A graphic featuring the words "Thank You!" written in a bold, red, cursive font. The text is set against a white background and has a slight drop shadow, giving it a three-dimensional appearance.

Capturing Tumor Genome and Germline Alterations in SEER

*SEER*DMS Face to Face meeting*

Rockville, MD, September 26-28, 2018

Valentina Petkov, MD, MPH

NCI/DCCPS/Surveillance Research Program



NATIONAL CANCER INSTITUTE



February 14, 2019

Outline

- *Current status*
- *The big picture – establishing an infrastructure*
- *Next steps – projects in the pipeline*

Current Status and Experience Collecting Tumor Genomics and Germline Alterations



Overview

- Importance to SEER
- Both tumor genome and germline alterations are determinants of response to therapy (predictive) and outcomes (prognostic)
- Issues with current collection of BMs as standard data elements
 - Limited to few BMs
 - Limitations to collections of new BMs
 - Rapidly changing landscape
 - Training
 - Quality: completeness and accuracy
 - Time lag

Oncotype DX linkage

- SEER performed 3 linkages
 - 2004-2012 dx year breast cancer cases
 - 2013 dx year
 - 2014-2015 dx year
- Data provided by Genomic Health (21- and 16- gene assays) are incorporated in SEER*DMS (8 variables)
- Linked Data are included in each November submission to SEER
- Data are released as specialized database upon request
- Approximately 40% of provided data were not captured in SSF22/23
- MOU re-linkage with SEER-Medicare in final stage

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 mutation tests
- 4 labs (Myriad, Invitae, 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 being analyzed
- Will be available to researchers through central registries

Establishing an infrastructure for collecting tumor genome and germline alterations



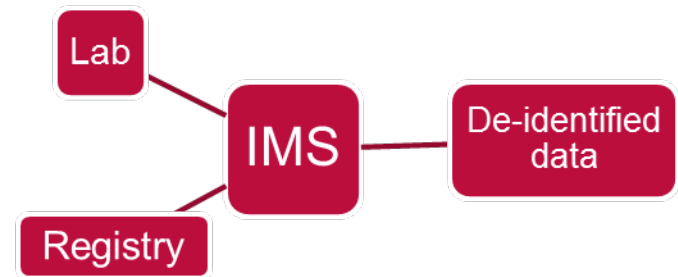
Regulatory aspects

- Do cancer registries have the authority to collect tumor genomic and genetic data?
- Communication sent to SEER PIs on 8/29
- Feedback received from 8 registries
 - All 8 registries are supportive and can collect genomic and genetic data
 - Two need state law change
 - One needs to investigate applicable privacy rules



Modes of data collection

- Traditional (manual) collection through a standard NAACCR abstract
- **Linkages with commercial companies/clinical laboratories or a third party data aggregators**
- Automated machine learning and deep learning algorithms



Genomic Data Evaluation

- What to collect
 - Manual data abstraction
 - Clinical guidelines
 - Complexity of data
 - Linkage source
 - Overlap with registry data
- Quality of collected genomic data
- Data Integration
 - Integrated in SEER*DMS
 - Stand alone data sets
- Data release plans and policies



Linkage projects in the pipeline



Linkage of OncotypeDX for IBC and DCIS

- Timeline: start in November
- Inclusion criteria
 - IBC 2004-2016 dx years - Tests 2004-2018
 - DCIS 2011-2016 dx year - Tests 2011-2018
- Rationale for re-linking
 - New registries
 - New software to be used (LinkPro)
 - Capturing tests on multiple primary tumors/ multifocal tumors/ multiple tests on the same tumor
- Strategies to eliminate duplicative work
 - Include flag for prior linkage in the PII file

Prostate cancer Multigene assays

- Used in the clinical practice but not supported by guidelines due to lack of evidence
- Prognostic, risk stratification, predictive
- Available tests
 - Oncotype DX for prostate (Genomic Health)
 - Decipher (GenomeDx)
 - Ploralis (Myriad)
- Commitment by Genomic Health and GenomeDx
 - Genomic Health prefer to conduct the linkage at the same time as BC linkage
- Linkage goals: test results for CTC; case finding

Other linkages

- Linkage of genetic tests (multigene panels)
 - Performed by 4-5 laboratories
 - Including all solid tumors dx 2013-2016
- Linkage of FoundationOne (Foundation Medicine)
 - GA registry
- Linkage with Caris
 - 590-gene panel; PD-L1, MSI
 - Registries to be determined

Other projects

- Project w/ Syapse: an 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 Seattle registry
 - Radiology reports
 - BMs, serological markers and multigene panels
- Project with Tempus: clinical lab and IT company for data integration
 - LA, IA, KY registries
 - Annotation of clinical documents
 - BMs, multigene panels, recurrence, therapy, outcomes





**NATIONAL
CANCER
INSTITUTE**

www.cancer.gov

www.cancer.gov/espanol