SEER*DMS 2017

Integrating new technologies and data streams

Linda Coyle

Information Management Services, Inc.
SEER*DMS Registries

2005  Metropolitan Detroit CSS
2006  Connecticut Tumor Registry
2007  Hawaii Tumor Registry
2007  New Mexico Tumor Registry
2007  Alaska Native Tumor Registry
2008  Cherokee Nation Cancer Program
2009  State Health Registry of Iowa
2009  Louisiana Tumor Registry
2010  Seattle Cancer Surveillance System
2011  Utah Cancer Registry
2012  Georgia Center for Cancer Statistics
2013  New Jersey State Cancer Registry
2016  New York State Cancer Registry
2017  Minnesota Cancer Surveillance System
2018  Kentucky Cancer Registry
Surveillance Research Program, NCI
surveillance.cancer.gov

Surveillance Informatics Branch
Supporting informatics tools for the next generation of cancer surveillance systems.

Mission Statement
Our mission is to advance precision cancer surveillance through the application of informatics tools, methods, and architectures. Research areas of focus for the branch include but are not limited to:

- innovative methods and approaches for acquiring, abstracting, analyzing, sharing, and visualizing data for cancer surveillance,
- automated and cost effective methods for data capture and integration, including abstraction facilitated with natural language processing (NLP) and machine learning,
- acquisition, linkage, and integration of traditional cancer registry data with new data sources to expand and enhance the scope of cancer surveillance research across the cancer continuum and fuel population-level precision oncology research, and
- enhancements of SEER Data Management System (SEER*DMS).

Highlights
- Natural Language Processing to Support Cancer Surveillance
- Blog on the Collaboration with Department of Energy
Data Sources – SEER*DMS Registries

Standard sources of medical data:

- Hospitals
- Pathology Laboratories
- Cancer Treatment Facilities
- Diagnostic Imaging Centers
- Physician Practices
- Clinics
- Nursing Homes/Hospices
- Other Cancer Registries

New data streams for SEER*DMS registries are expected to include:

- Insurance Claims
- Pharmacy Data
- Physician Reports
- Genetic Information
New data sources will increase the number of:

- Reports per patient
- Data items collected:
  - Biomarkers
  - Genetic markers
  - Detailed treatment information

And provide potential resources for

- Passive follow-up
- Case finding
Today’s Presentation

The SEER Program is:

- Increasing the amount and types of data collected
- Implementing technical solutions to process the data and mitigate impact on cancer registrars

Increases in data streams affect all aspects of a data management system: hardware, software, user experience.

This presentation will review the impact on:

- System Infrastructure
- SEER*DMS Software
- SEER*DMS Registries
System Infrastructure

Changes in the computing environment to support new data streams and technologies
2013-2017 – Converted all registries from a server housed at the registry to the computer center in order to increase security and reduce costs.

Added value – this model makes it easier to adapt to changing needs; much easier to accept and process high volume data streams.

Registry’s instance of SEER*DMS is hosted at an IMS Computer Center.

Each registry has a separate and independent enclave of servers.

Registry users access SEER*DMS via a Web browser using a secure connection.
Registry enclaves in the data center can be expanded – to support additional applications.
Data Transmissions to SEER*DMS

Standard method:

- Organization sends data to registry’s local environment
- Registry transmits data to SEER*DMS autoloader; or staff upload data via the SEER*DMS application

This is the recommended method for local and regional data sources (hospitals, path labs, etc).
National organizations can stream data to a single, dedicated island in an IMS computer center.

This island is stand alone; cordoned off from everything else by a firewall.

An IMS-maintained service transfers data from the data feed island to the appropriate registry island.
Making National Data Available to Registries

Unlimited Systems is now transmitting data for 5 registries (as of March 2017).

Requirements:
- A single agreement between IMS and Unlimited Systems. A similar agreement would need to be made with each national organization.
- An amendment to the Connection Agreement between IMS and the registry. The intent is for this to cover all national data sources.

Benefits:
- Saves time and effort for each registry in terms of infrastructure to receive and transmit data.
- National organizations are more likely to onboard – less hassle dealing with one entity (IMS) instead of 4 to 15 registries.
Changes to the application and database to support new data streams and technologies
Design once, use many. Each new data stream:

- Large amount of raw data
- Variety of data items and file formats
- Cannot be processed by the registry in their standard workflow --- it would be overwhelming.

Solution:

- One data store for high volume data types
- Separate from traditional source records
- Use database structure and algorithms that support all data types (claims, pharmacy, MU2, etc).
Workflow for High Volume Data

Import Data
- Load data into “PRE_RECORD” table. Data stored in JSON format.
- Sample file formats: ANSI 837 – 5010A for Claims; CDA for MU2 data
- JSON – self-describing; and well-suited for the multi-level data in claims and other data. One JSON definition for claims; another for MU2, etc.

Code Fields
- Perform some standardization of the data: address normalization; conversion to ICD-O-3 codes, etc.

Match to Database
- Match against Patient Sets
- Use SEER MP/H rules for tumor level matching

Evaluate
- Targeted review of data – identify cases for manual review. Example: Manually update treatment data for cases identified via data searches

Define Automation
- Based on the manual evaluation and data analysis, define business rules for auto-processing the data. For example, automatically code chemo fields based on claims data.

This workflow is currently being used for claims in GA.

The same principles will be used to evaluate other data sources.
Advantages:
- Adding a field is much easier.
- JSON object is one column in the "pre_record" table.
- Same table can be used for different data types; the JSON column differs by type.

Disadvantage:
- It is harder for typical users to query using SQL. IMS will provide viewers and data searches.

This claim indicates chemo was administered on 03-30-2015. J9310 = Rituximab injection.
SEER*DMS Claims Viewer

JSON viewer is available – primarily for IMS staff to resolve issues.

Registrars see the Claims Viewer in the right panel of the Patient Set editor.
SEER*DMS Road Map

To handle new data streams, we must:
- Develop automated solutions for claims and other new data. Implement solutions in phases – first phase will be more manual than later phases.
- Improve current processes. Increase automation of standard data types (abstracts, path reports, etc).

2017-2018 Goals
- Use Natural Language Processing to set site, histology, behavior in pathology reports
- Fully Automate Tumor Linkage
- Automate Patient and Tumor Consolidation
An API is a tool that makes it easy to re-use code. SEER MP/H rules can be used throughout the system.

APIs allow organizations to share code with each other. A broader community can work toward technical solutions.
Automated Tumor Linkage

Auto-link source records, using rules that align with SEER 2007 multiple primary and hematopoietic database rules (MP/H).

Auto-linking in SEER*DMS prior to 2016:
- Death certificates – linked per NAACCR Death Clearance Manual guidelines
- Pathology reports – linked to same CTCs as their matching abstracts
- Re-submitted abstracts for older cases – linked on matching rules defined by registries based on simplified version of 2007 MP/H
- Non-analytic abstracts – treated as case finding records in many registries and linked to matching CTCs
- Abstracts that are “re-submissions” to registry; processed for follow-up only

Priority of SEER*DMS development 2016-2017:
- Increase the percentage of abstracts and pathology reports that are auto-linked as part of effort to increase overall automation in SEER*DMS
Multiple Primary Rules in SEER*DMS

2007 MP/H Rules released in SEER*DMS v16 (January 30, 2015)
- Not used for auto-linking in first release; primarily a “test” release
- Rules used to set higher priority for tasks related to probable new cases
- Executed the rules across registry databases to test the implementation. Registry CTR volunteers evaluated the results.
- Contacted Florida Cancer Data System (FCDS) to discussed possibility of testing IMS implementation vs FCDS results

Rules implemented in the initial release
- Hematopoietic Rules (histology of 9590 – 9989) using hematopoietic database; year dx 2001 and later.
- 2007 Multiple Primary Rules – year dx 2007 and later

Limited release of 2007 MP/H Rules in SEER*DMS v16 (Jan 30, 2015)

June 2015 – Pete Ransdell (Kentucky Cancer Registry) described KY’s implementation at annual NAACCR conference. IMS approached KCR to compare results and join forces – KCR and IMS are both Java shops

April 2016 – IMS completed review of KCR and IMS implementations

2007 MP/H Rules – reviewed differences with KCR. Minimal differences. Some changes to KCR approach; some changes to IMS; and a few that need SEER review. Overall – the results aligned nicely.

KCR’s implementation included additional rule sets:

- Hematopoietic 1998 Rules (year dx 2000 or earlier)
- 2004 Solid Tumor and Benign Brain Rules (year dx 2006 or earlier)

KCR and IMS agreed that a single Java library should be used; KCR stated the preference for IMS to maintain the Java library and API.

IMS implemented all revisions based on the review and implemented rules for 1998 Hematopoietic and 2004 Benign Brain using KCR algorithms and SEER manuals.
Multiple Primary Rules in SEER*DMS

Using the MP/H rules in SEER*DMS 2016:

- Setting priority flags on tasks indicating new incident case for an existing patient
- Auto-linking higher % abstracts in some registries – registries are evaluating rules on test and dev servers; increasing use in production
- Linking claims to CTCs

Next steps (2 to 3 months)

- Auto-link “all” matching abstracts and path reports, but require manual consolidation per the registry’s current rules.
- May have some exceptions for certain sites, eg, hematopoietic diseases
- Auto-link and auto-consolidate all matching abstracts that provide new treatment data, but do not have changes to other data items (date of diagnosis, staging, etc). Trigger a manual review for limited number of cases (review of anomalies)
Tumor consolidation reviews all reports for each tumor/cancer and creates a “best” cancer, stage, and treatment record for each tumor associated with the patient.
Technical view:

SEER*DMS workflow supports auto-consolidation of patient and tumor level data items.

Reality:

A large number of patient-level data items are auto-consolidated, but treatment fields are the only tumor-level data items auto-consolidated in most registries.

To this point, registry rules have determined the level of automation. Each registry’s auto-consolidation rules were implemented, but no registry auto-consolidated tumor items prior to their SEER*DMS migration.

Pre 2016 workflow supported auto-consolidation rules for tumor level data items were supported, improvements were needed to do it more efficiently.

2016 - overhaul of workflow in v17 -- better support for auto-consolidation of all data items

There is still work to be done to define business rules and in the technical implementation, for example, need to improve validation of source records prior to consolidation.

CCB auto-consolidation workgroup will define rules based on KCR rules, FCDS rules, NAACCR guidelines, and subject matter experts from the SEER*DMS registries.
SEER*DMS Registries

Defining appropriate uses and workflows for high volume data.
<table>
<thead>
<tr>
<th>Year</th>
<th>Registry</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>Louisiana Tumor Registry</td>
</tr>
<tr>
<td>2010</td>
<td>Seattle Cancer Surveillance System</td>
</tr>
<tr>
<td>2012</td>
<td>Georgia Center for Cancer Statistics</td>
</tr>
<tr>
<td>2018</td>
<td>Kentucky Cancer Registry</td>
</tr>
<tr>
<td>Year</td>
<td>Registry Name</td>
</tr>
<tr>
<td>------</td>
<td>------------------------------------------------------</td>
</tr>
<tr>
<td>2006</td>
<td>Connecticut Tumor Registry</td>
</tr>
<tr>
<td>2007</td>
<td>Hawaii Tumor Registry</td>
</tr>
<tr>
<td></td>
<td>New Mexico Tumor Registry</td>
</tr>
<tr>
<td>2008</td>
<td>State Health Registry of Iowa</td>
</tr>
<tr>
<td>2009</td>
<td>Louisiana Tumor Registry</td>
</tr>
<tr>
<td>2011</td>
<td>Utah Cancer Registry</td>
</tr>
<tr>
<td>2012</td>
<td>Georgia Center for Cancer Statistics</td>
</tr>
<tr>
<td>2013</td>
<td>New Jersey State Cancer Registry</td>
</tr>
<tr>
<td>2016</td>
<td>New York State Cancer Registry</td>
</tr>
</tbody>
</table>
Auto-linking Workgroup

- 2007: Hawaii Tumor Registry
- 2008: Alaska Native Tumor Registry
- 2011: Utah Cancer Registry
- 2012: Georgia Center for Cancer Statistics
- 2016: New York State Cancer Registry
Claims Workgroup

2007  ●  New Mexico Tumor Registry

2011  ●  Utah Cancer Registry

2012  ●  Georgia Center for Cancer Statistics

2013  ●  New Jersey State Cancer Registry

2018  ●  Kentucky Cancer Registry
SEER*DMS Registry Workgroups

- Make data available for research & development of algorithms – NLP, Claims, MU2
- Collaborate with NCI, DOE, and IMS staff to establish the CDAP architecture and protocols
- Work with IMS to define and analyze workflows
  - IMS creates system tasks to compare new, automated workflow against data that were processed manually
  - Registries evaluate discrepancies between automated decisions and manual
  - Analyze claims and MU2 data to define workflows
Acknowledgments

- Suzanne Adams
- Dave Annett
- Scott Depuy
- Chuck May
- Nicki Schussler
- and the entire SEER*DMS team!

Thank you!!