Enabling cancer research with end-to-end infrastructure for the request, exploration, and receipt of cancer registry data

Ian Foster – foster@uchicago.edu
University of Chicago

Jonathan Silverstein – j.c.s@pitt.edu
University of Pittsburgh

Collaboration Update
NCI ITCR Monthly PI Meeting, March 1 2019

NCI U24 CA209996, PI: Ian T. Foster
Collaborators

• University of Pittsburgh
  – Biomedical Informatics (Jonathan Silverstein)
  – Institute for Precision Medicine (Adrian Lee)

• UPMC
  – Hillman Cancer Center (Robert Ferris)
  – Network Cancer Registry (Sharon Winters)

• University of Chicago
  – Globus (Ian Foster)
CURRENT STATE: Step 1: Define desired data set

Researcher talks to registrar about what kind of patient cohort they want.

Registrar queries the registry for the researchers.

Lengthy, iterative, labor-intensive process
No mechanism for ad hoc exploration of the data by the researcher
Registrar delivers data via email, OneDrive, Sharepoint, mapped drive
No tracking/auditing, error prone, insecure
FUTURE STATE: Step 1: Define desired data set

Goal 1: Provide mechanism for self service exploration of registry data
FUTURE STATE: Step 2: Deliver data to researcher

Registrar delivers data to researcher

Registrar extracts data from Registry

Goal 2: Provide secure, auditable, access controlled mechanism to share data
Cancer Registry Records for Research (CR3) User Requirements

- De-identified data only may be in Search index.
- Researchers may not have record level access to Search index.
- Access to portal must be controlled.
- Researchers must use the specific identity that has been granted access permissions.
- UPMC identity must be an option.
- Access to portal by researcher must be logged.
- Render 2D graphs based on search results.
- Faceted search in real time with the desired logic.
Progress

• **Customized existing Globus portal code**
  – Aggregate charts for age range, gender, histology
  – Collapsible facet panels
• **Completed security assessments**
• **Extracted, curated and ingested registry data**
• **Demonstrated first version to potential users**
  – Basic prototype was developed within 1-2 month
  – Two diseases (i.e. breast, lung) were ingested into Globus Search
<table>
<thead>
<tr>
<th>Data Stats</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>7 Diseases</strong></td>
<td>&quot;laterality&quot;: &quot;Left - origin of primary&quot;, &quot;site&quot;: &quot;Lung, lower lobe&quot;, &quot;dx_year&quot;: &quot;2004&quot;, &quot;cs_lymph_node&quot;: &quot;0&quot;, &quot;recurrence&quot;: &quot;Yes&quot;, &quot;cancer_status&quot;: &quot;Evidence of this tumor&quot;, &quot;clinical_m&quot;: &quot;M0&quot;, &quot;site_code&quot;: &quot;C343&quot;, &quot;cs_stage&quot;: &quot;1B&quot;, &quot;scope_reg_ln_summ&quot;: &quot;4 or more reg LN removed&quot;, &quot;histology_code&quot;: &quot;80703&quot;, &quot;mets_at_dx&quot;: &quot;No distant metastasis&quot;, &quot;grade&quot;: &quot;Grade II: Mod diff, mod well diff,&quot; &quot;year_last_contact&quot;: &quot;2005&quot;, &quot;histology&quot;: &quot;Squamous cell carcinoma, NOS&quot;, &quot;dx_age_range&quot;: &quot;65-74&quot;, &quot;dx_age&quot;: &quot;71&quot;, &quot;race&quot;: &quot;White&quot;, &quot;disease_category&quot;: &quot;Lung&quot;, &quot;clinical_stage&quot;: &quot;1B&quot;, &quot;clinical_t&quot;: &quot;T2&quot;, &quot;gender&quot;: &quot;Female&quot;, &quot;spanish_origin&quot;: &quot;Non-Spanish; non-Hispanic&quot;, &quot;cause_of_death&quot;: &quot;Cancer related&quot;, &quot;path_n&quot;: &quot;N0&quot;, &quot;facility&quot;: &quot;Shadyside&quot;, &quot;path_t&quot;: &quot;T2&quot;, &quot;clinical_n&quot;: &quot;N0&quot;, &quot;path_stage&quot;: &quot;1B&quot;, &quot;path_m&quot;: &quot;MX&quot;, &quot;vitals&quot;: &quot;Dead&quot;</td>
</tr>
<tr>
<td><strong>Years 2004 - 2017</strong></td>
<td></td>
</tr>
<tr>
<td><strong>65,000 patients</strong></td>
<td></td>
</tr>
</tbody>
</table>

7 Diseases: Breast, Colon, Lung, Melanoma, Head and Neck, Ovary, Prostate
Cancer Registry Discovery Portal (CR3)

This data was curated by the Cancer Registry from UPMC hospitals and UPMC Hillman Cancer Center hospital-based clinics

About the CR3 data repository...
The UPMC Network Cancer Registry (Registry) is comprised of full time staff responsible for maintaining a standardized data system designed for the collection, management and analysis of demographic, diagnosis, staging, grading, treatment and progression data on patients having a diagnosis of cancer and other reportable conditions seen at UPMC hospitals and UPMC Hillman Cancer Center hospital-based clinics. More than 15-20 thousand new patients are added to the registry each year with decades of data already in place among hundreds of thousands of patients across essentially all cancer body-sites. The curation of this data is highly complete and maintained according to North American Association of Central Cancer Registries (NAACCR) standards across essentially all UPMC cancer treatment sites.

**The data contained within this repository has been de-identified and contains no PHI information it’s intended usage is for research cohort discovery.**

You have successfully logged in and can now begin searching the CR3 repository.

Get Started
Cancer Registry Discovery Portal (CR3)

This data was curated by the Cancer Registry from UPMC hospitals and UPMC Hillman Cancer Center hospital-based clinics.

Home

Overview of Cancer Types

- Breast: 36.5%
- Lung: 19.5%
- Melanoma, Cutaneous (Skin): 15.4%
- Head and Neck: 11%
- Colon: 10.9%
- RUS and Rectum: 9.5%
- Ovary: 9.5%
- Other Gynecological Cancers: 5.9%
- Other: 5.9%

Search

You can search the Cancer Registry repository by entering any search term (e.g., melanoma).

Also, for advanced searching see the advanced query documentation.

Search for a disease or condition (e.g., melanoma)
Based upon the criteria you have selected your cohort contains an estimated 12,322 cases.

You entered the following term(s) for your initial disease or condition search:

disease_category: "Lung"

Additional Comments for Request:

Review Email Request

Continue Exploring Data
Cancer Registry Records for Research (CR3) Advanced Features

- Modern technical approach (SaaS deployment)
- Federated Identity (each institution does authentication)
- Honest Brokerage with clinical data (local control)
- NAACCR Standard data (each institution can generate)
- Search via Globus (one search traverses institutions)
- Easy GUI to develop and use (includes faceted search)
- Easy search transmission (search “sentence” to registrar)
- Secure Data Liquidity (easy to move data to authorized)
CR3 Architecture v1.0

Researchers log in with UPMC/Pitt credentials to the CR3 Discovery Portal. Authentication is initiated to the Globus Auth (GA) service. Cohort aggregate counts are returned. Email requests are sent to CR3 Registry Staff.

The Elasticsearch Cancer Registry De-identified Data Index (minimal criteria data: e.g., staging) is connected to the Globus Search (GS) service. Cohort searches are initiated to GS.
Next Steps

• Prevent record level access by CR3 users
• Require specific identity (no linked identities) for CR3 access
• User testing
  – Registrars
  – Researchers
• Production release of CR3 v1
• Integrate UPMC identity provider
• Extend CR3 to identified data sharing
Long Term Goals

- Create a network of federated cancer registries
  - Deploy similar infrastructure at other cancer registries
  - Enable queries across multiple registries
  - Model proven in ITCR TIES/TCRN project with independent installation of TIES or connect a TIES installation to TCRN for multi-institutional queries

- Federation via Globus allows network to scale while control remains with each institution
  - Data owners input and export data sets, apply quality control
  - Data owners control all access policies
  - No overhead from ingestion of disparate datasets
  - Registry data remains at the location where it was generated
  - Identities are provided and authenticated by the institution, not Globus
  - System can scale almost infinitely if data owners provide their own storage resources
Questions?

j.c.s@pitt.edu
foster@uchicago.edu