SEINE: There and Back Again
Dan Connolly
Biomedical informatics software engineer
Division of Medical Informatics
KUMC Medical Informatics

- Russ Waitman, Director of Medical Informatics
- Software Engineers: Dan Connolly, Nathan Graham, Bhargav Adagarla, Matt Hoag, Mike Prittie, Lav Patel, Nazma Kotcherla
- Analysts, Honest Brokers: Tamara McMahon, Sravani Chandaka, Li Huang, Rachel Gyore, Maren Wennberg
- Project Management: Steve Fennel, Brittany Zschoche, Hillary Sandoval
Your researchers are fisherman: wanting to land data to answer their research hypothesis.
The Fish: Diagnoses, Demographics, Observations, Treatments
The Fish: Diagnoses, Demographics, Observations, Treatments

Healthcare Enterprise Repository for Ontological Narration
SEINE = EDC <-> IDR
Pilot project:
Triple Negative Breast Cancer (TNBC) Registry
Pilot project: Triple Negative Breast Cancer (TNBC) Registry

Pilot experience:

- Instrument Validation
  - Age as drop-down vs. text with number validation
  - Repetition vs. longitudinal
- Data Integration (overlap)
  - Biospecimen Repository: 61%
  - NAACCR tumor registry: 56%
  - Social Security Death Index: 2%
REDCap Forms, Fields + Data -> i2b2 Folders, Concepts + Facts

- ETL code developed as part of HERON ETL
- For access, see MultiSiteDev in informatics.gpcnetwork.org
REDCap Forms, Fields + Data -> i2b2 Folders, Concepts + Facts

Road Not Taken: ODM

Poor fit for KUMC’s automated ETL process.
Data Privacy

- HERON is fully de-identified.
- ETL strips:
  - Free text
  - REDCap identifier fields
Access Control

• Norm: i2b2 UI respects REDCap access control
  • i2b2 projects configured real-time at i2b2 log-in
• Special case: REDCap project open to all i2b2 users
Access Control

- Norm: i2b2 UI respects REDCap access control
  - i2b2 projects configured real-time at i2b2 log-in
- Special case: REDCap project open to all i2b2 users

Monthly transition not smooth. Design improvements pending
SEINE DataBuilder: i2b2 -> REDCap

- Data Delivery for HERON i2b2 user community at KUMC
  - Pilot studies in 2012
- Federated Data collection in GPC
  - Breast Cancer
  - ALS
Data Delivery via REDCap

- REDCap output project has
  - Demographics CRF (age, sex, race, etc.)
    - From i2b2 patient dimension, plus
      - Diagnoses CRF if anything from i2b2 Diagnoses selected
      - Medications CRF
      - Lab Results CRF
      - ...
  - The project can be organized either
    - by-patient or
    - by-encounter.
SEINE DataBuilder: Diabetes + Vertigo Study

Pilot Experience:
- REDCap Graphical View & Stats sufficient for preliminary analysis
- Export to stats program

PMCID: PMC4791946
Identified, de-identified data fulfillment

- ~40 deliveries/month
- A common case: an identified cohort
  - Variables: just MRN
  - the customer then adds CRFs
Data Builder

1. Governance Committee OKs data request
2. Honest Broker uses Data Builder plug-in to specify data
i2b2 Timeline: Specify Data

- **Patient Set**
  - Self-service cohort query
- **Concepts**
### i2b2 Timeline: View Results

**LDS access role**

<table>
<thead>
<tr>
<th>Person</th>
<th>Date of Birth</th>
<th>Include in Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person #44206</td>
<td>6/28/1970</td>
<td>Include</td>
</tr>
<tr>
<td>Person #526833</td>
<td>1/9/2017</td>
<td>Include</td>
</tr>
</tbody>
</table>

- **Person #44206**
  - m 67 yr old White
  - 003-15324 Verbal Fluency [2,303 facts; 760 patients]
  - 008-15319 MOCA Score (out of 30) [3,479 facts; 2,419 patients]
  - 332 Parkinson's disease [165,266 facts; 8,114 patients]
  - Age [2,222,790 facts; 2,222,787 patients]
  - Body Mass Index [4,211,058 facts; 518,384 patients]
  - Ethnicity [2,246,187 facts; 2,246,184 patients]
  - G20 Parkinson's disease [86,381 facts; 4,624 patients]
  - Gender [2,246,187 facts; 2,246,184 patients]
  - Marital Status [2,246,187 facts; 2,246,184 patients]
  - Race [2,246,560 facts; 2,246,184 patients]

- **Person #526833**
  - f 82 yr old White
  - 003-9501244 Mini-Mental Status Examination [5,381 facts; 2,096 patients]
  - 005-12654 Verbal Fluency [2,034 facts; 1,610 patients]
  - 006-12655 Trailmaking A Time (secs) [1,680 facts; 1,442 patients]
  - 007-12658 Trailmaking A Errors [1,665 facts; 1,429 patients]
  - 008-12657 Trailmaking B Time (secs) [1,610 facts; 1,377 patients]
  - 009-12659 Trailmaking B Errors [1,592 facts; 1,361 patients]
  - 331.0 Alzheimer's disease [50,936 facts; 5,680 patients]
<table>
<thead>
<tr>
<th>Subject</th>
<th>Gender</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>4206</td>
<td>m</td>
<td>67</td>
</tr>
<tr>
<td>833</td>
<td>f</td>
<td>82</td>
</tr>
</tbody>
</table>
Many i2b2 facts -> 1 spreadsheet cell

<table>
<thead>
<tr>
<th>Subject</th>
<th>Gender</th>
<th>Age</th>
<th>Verbal Fluency</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>4206</td>
<td>m</td>
<td>67</td>
<td>???</td>
<td>???</td>
</tr>
<tr>
<td>833</td>
<td>f</td>
<td>82</td>
<td>???</td>
<td>???</td>
</tr>
</tbody>
</table>
Technique 1: Aggregation

<table>
<thead>
<tr>
<th>patient_num</th>
<th>BMI: count</th>
<th>BMI: first</th>
<th>BMI: last</th>
<th>BMI: median</th>
</tr>
</thead>
<tbody>
<tr>
<td>4206</td>
<td>19</td>
<td>23.5</td>
<td>24.2</td>
<td></td>
</tr>
<tr>
<td>833</td>
<td>4</td>
<td>25.1</td>
<td>24.8</td>
<td></td>
</tr>
</tbody>
</table>
## Technique 2: Record-per-encounter

<table>
<thead>
<tr>
<th>encounter_num</th>
<th>patient_num</th>
<th>Age</th>
<th>start_date</th>
</tr>
</thead>
<tbody>
<tr>
<td>934875</td>
<td>4206</td>
<td>67</td>
<td>2017-01-02</td>
</tr>
<tr>
<td>374830</td>
<td>4206</td>
<td>67</td>
<td>2017-01-09</td>
</tr>
<tr>
<td>302975</td>
<td>833</td>
<td>82</td>
<td>2017-01-08</td>
</tr>
</tbody>
</table>
GPC Federated Query: Breast Cancer Cohort

- **GPC common condition = Breast Cancer**
- **2015: Federated query to define survey cohort**
  - Putting RC11 from the [GPC Proposal](#) into practice
- **SEINE DataBuilder method used at 8 sites**
  - ~6 sites used common python code
  - ~2 sites developed work-alikes based on specs in GPC Wiki
Federating raw EAV data

- Initial BC query:
  - raw files sent to KUMC for aggregation into a spreadsheet
  - QC was labor intensive
    - Lots of round-trips
Like the CRC cell, **DataBuilder** calls the **PM cell** to validate authorization.

Raw EAV data was collected; aggregation for REDCap wasn’t deployed to sites.
REDCap as gatekeeper

- Initial BC query:
  - raw files sent to KUMC for aggregation into a spreadsheet
  - QC was labor intensive
    - Lots of round-trips

- Later queries (ALS, 2nd BC):
  - raw files were pivoted on-site before submission
  - REDCap data dictionary constraints provided much/most of the QC

Custom EAV->spreadsheet pivoting code; not generic SEINE edc_summary.py code
SEINE = REDCap<-i2b2 + REDCap->i2b2

- Registry
- Design
- Data Entry
- Bulk Import

- Cohort Discovery
- Data Visualization
- Data Integration
- Data Validation
- Instrument Validation

EDC System (REDCap)

- Pivoted Data
- Export to SAS, CSV, XLS
- Basic Statistics
- Annotation

Cohort Discovery & Data Delivery

EDC ETL

ETL from EMR + ...

SEINE manuscript
SEINE Full Cycle Case: Cancer Distress Registry

1. 3675 patients

- Registry Design
- Data Entry
- Bulk Import

EDC System (REDCap)

4. 150K facts on 4717 patients from EHR, tumor registry, vital statistics

5. data collection gap analysis

IDR (i2b2)

- Cohort Discovery
- Data Visualization
- Data Integration
- Data Validation
- Instrument Validation

3. cohort += 1042

- ETL from EMR + ...

- Pivoted Data
- Export to SAS, CSV, XLS
- Basic Statistics
- Annotation

SEINE manuscript

1. 3675 patients

3. cohort += 1042
DataBuilder scalability

- 10 or 20 minutes is typical at KUMC
  - i2b2 star schema on solid-state storage
- \geq 30K patients stresses the system
  - But does not break it
Support: open gpc-dev process

- DataBuilder in informatics.gpcnetwork.org wiki ->
  - SEINE Manuscript
  - heron_extract hg repository

- “if it breaks, you get to keep both pieces” ;)
- gpc-dev mailing list is open
- weekly teleconferences, annual meetings accommodate occasional guests
UTSW Success Story

- Used i2b2 to select a cohort of 2000 patients (out of 5M from i2b2 repository).
  - 20 additional observations along with Diagnoses, Procedures
  - Took ~3 hrs to run the DataBuilder
- Some post-processing for MRI etc. into REDCap.
- Just 4hrs total.

“Huge time saver” — T. Bosler