Secondary Use of Clinical Data

- Electronic health records contain a large quantity of data drawn from multiple sources (e.g., administrative, laboratory, radiology, pharmacy) about each patient (Sarkar, 2010)

- Potential applications
  - Quantitative biomarkers derived from radiological images
  - Genotype-to-phenotype correlation analysis
  - Creation of teaching files
  - Clinical decision support by building predictive models
The Need for Deep Phenotyping
The Need for Deep Phenotyping

Our capacity to undertake [genotyping projects] will probably not be the limiting step in elucidating the genetic basis of common diseases. Instead, we may be held back by our inability to specify precisely the phenotypes (the observed manifestation of the genotypes) in those individuals whose genomes we investigate. Nat. Gen. 2003, 34: 15-21.
How Can We Get There?

- Leveraging the electronic health record for deep phenotyping
- Efforts are underway to develop the software infrastructure and collaborative networks
  - i2b2 (Harvard)
  - eMERGE network
  - Kaiser/VA
- Recent publications have demonstrated promising results

...recent studies demonstrate that they can be effectively employed for genetic studies using the information and biological ‘by products’ of health-care delivery.

Kohane IS. Nature Reviews Genetics, 2011.

Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium

We have identified five disease phenotypes with PPV of 73-98% and NPV of 98-100%.

Kho AN. Science Translational Medicine, 2011.
## Record-driven vs. Cohort-driven

<table>
<thead>
<tr>
<th></th>
<th>Health-record driven</th>
<th>Cohort-driven</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Timeliness</strong></td>
<td>Contemporary – reflective of current exposures</td>
<td>Varies with how often the cohort is refreshed</td>
</tr>
<tr>
<td><strong>Cost</strong></td>
<td>Marginal cost in addition to standard of care</td>
<td>Significant cost of parallel research infrastructure</td>
</tr>
<tr>
<td><strong>Populations studied</strong></td>
<td>Representative of a clinical care population</td>
<td>Determined by study/research protocol</td>
</tr>
<tr>
<td><strong>Assess comorbidity of genomic-phenotype</strong></td>
<td>Fragmentary</td>
<td><strong>Systematic</strong></td>
</tr>
<tr>
<td><strong>Family history</strong></td>
<td>Variable quality</td>
<td><strong>Consistent</strong></td>
</tr>
<tr>
<td><strong>Environmental exposure</strong></td>
<td><strong>Up-to-date and comprehensive</strong></td>
<td>Up-to-date to the extent that there is synchronization with health data</td>
</tr>
<tr>
<td><strong>Data accuracy</strong></td>
<td>Reliability will vary with only a subset being most accurate</td>
<td><strong>Systematic quality control of data collected</strong></td>
</tr>
<tr>
<td><strong>Range of data</strong></td>
<td><strong>Broad reflection of clinical states and exposures</strong></td>
<td>Determined by the protocol</td>
</tr>
<tr>
<td><strong>Consent</strong></td>
<td>Ambiguity exists in current IRB protocols (opt-out/opt-in models)</td>
<td><strong>Established protocol that has withstood scrutiny and debate</strong></td>
</tr>
</tbody>
</table>

Kohane IS. Nature Reviews Genetics, 2011.
But There Are Many Caveats

Limitations

- Inaccurate
- Incomplete
- Transformed and coded in ways that undermine meaning
- Unrecoverable for secondary uses
- Unknown provenance
- Insufficient granularity

Hersh W et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. Medical Care, 2013.
But There Are Many Caveats

Limitations

- Inaccurate
- Incomplete
- Transformed and coded in ways that undermine meaning
- Unrecoverable for secondary uses
- Unknown provenance
- Insufficient granularity

Potential Solutions

- Approaches to assess availability, completeness, and quality of the data
- Tools to manage data and their attributes
- Methods for comparative validation
- Standardized reporting
- Appropriate use of informatics expertise

Hersh W et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. Medical Care, 2013.
Understanding the evolution and risk of aneurysm rupture

MODELING INTRACRANIAL ANEURYSMS
Intracranial Aneurysms

- Outpouching that occurs at the point at which a cerebral artery departs from the Circle of Willis
  - Significant complications: Rupture (SAH), mass effect, stroke
- Can we accurately predict an individual’s risk of aneurysm rupture?
Study Characteristics

• Collection of **retrospective** and **prospective** observational clinical data
• Data will be collected **longitudinally** at varying time points during patient’s standard of care
• Data will need to be **structured** from clinical reports and medical images
  – Not all desired information is reported in the patient record and must be generated
• Multiple investigators involved
  – Need a centralized database to store collected information
Workflow

Electronic Health Record

Clinical data capture & workflow

Clinical Encounters -> Blood

Imaging studies

MR/CT angiography

Documentation

Reports / Lab Values

Research data & workflow

RNAseq/DNA methylation -> 3D morphometric analysis

Hemodynamic simulation

Extraction & characterization

Aneurysm Data Model

Integrated Observational Database
Data Model

- Important to enumerate what data elements are being collected, what are the permissible values
- Utilize common data elements when possible
- Flat file or relational database?
### Aneurysm Modeling Project
**University of California, Los Angeles**

**R01 EB000362**

<table>
<thead>
<tr>
<th>Case Number</th>
<th>Presentation</th>
<th># of Aneurysms</th>
<th>Medical History</th>
<th>Follow-up</th>
<th>Image analysis</th>
<th>Hemodynamics</th>
<th>Blood sample</th>
<th>Sequence data</th>
</tr>
</thead>
<tbody>
<tr>
<td>9917050</td>
<td>Incidental</td>
<td>5</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>9794297</td>
<td>Incidental</td>
<td>1</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>9715676</td>
<td>Incidental</td>
<td>1</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>9590221</td>
<td>Incidental</td>
<td>1</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>9548652</td>
<td>Incidental</td>
<td>2</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>9518604</td>
<td>SAH</td>
<td>2</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>9165729</td>
<td>Incidental</td>
<td>2</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>8798524</td>
<td>Incidental</td>
<td>1</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>8694295</td>
<td>SAH</td>
<td>1</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>8567437</td>
<td>Incidental</td>
<td>1</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
</tbody>
</table>

Showing 1 to 10 of 78 entries

#### Available Reports
- Aneurysm Characteristics
- Procedures
- Complications
- Ruptures

#### Summary
- Patient Summary
- Longitudinal Summary

#### Additional Reports
- Click for additional reports

#### Basic Filters
- **Select Form**: Patient, Aneurysm, Imaging Followup, Treatment, Clinical Followup, Hospital Course
- **Select Field**: Date of scan, Modality, Institution where imaging, Source of measurements, Aneurysm status, Treatment status

#### Selected Filters
- Gender: Female
- Procedure: Endovascular Coiling
- Aneurysm Status: Existing

---

**Export Data**  **Custom Export**
There is a hyperdense large intraparenchymal hematoma centered in the anterior aspect of the right frontal lobe...

**HEMATOMA**
- **Existence** = Present (thereIs)
- **Quantity** = 1
- **Size** = Large
- **Image characteristic** = Hyperdense
- **Hemorrhage Subtype** = Intraparenchymal
- **Location** in Right frontal lobe
- **Directionality** = Anterior aspect
Image Analysis

- Simple flow pattern
- Stable
- Divided
- Impact on dome
- Small impingement
- Concentrated jet
- Circular jet flow
- Wall shear stress

Hemodynamics

Morphology Analysis
## Data Challenges

<table>
<thead>
<tr>
<th>ID</th>
<th>Imaging Exam</th>
<th>OSH</th>
<th>Date</th>
<th>#</th>
<th>Side</th>
<th>Shape</th>
<th>Location</th>
<th>Status</th>
<th>AP</th>
<th>TR</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MRA Brain</td>
<td>Y</td>
<td>2012-10-04</td>
<td>1</td>
<td>LT</td>
<td>fusiform</td>
<td>MCA</td>
<td>untreated</td>
<td>19.1</td>
<td>19.6</td>
<td>24.9</td>
</tr>
<tr>
<td>1</td>
<td>MRA Brain</td>
<td>Y</td>
<td>2012-10-04</td>
<td>2</td>
<td>RT</td>
<td>saccular</td>
<td>MCA</td>
<td>untreated</td>
<td>1.5</td>
<td>2.1</td>
<td>2.2</td>
</tr>
<tr>
<td>1</td>
<td>MRA Brain</td>
<td>Y</td>
<td>2012-10-04</td>
<td>3</td>
<td>RT</td>
<td>saccular</td>
<td>ACOM</td>
<td>untreated</td>
<td>2.2</td>
<td>2.3</td>
<td>1.4</td>
</tr>
<tr>
<td>1</td>
<td>MRA Brain</td>
<td>Y</td>
<td>2012-10-04</td>
<td>4</td>
<td>RT</td>
<td>saccular</td>
<td>PCOM</td>
<td>untreated</td>
<td>3.6</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>IR Angiogram</td>
<td>N</td>
<td>2012-11-09</td>
<td>1</td>
<td>LT</td>
<td>fusiform</td>
<td>MCA</td>
<td>untreated</td>
<td>18</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>IR Angiogram</td>
<td>N</td>
<td>2012-11-09</td>
<td>2</td>
<td>RT</td>
<td>saccular</td>
<td>MCA</td>
<td>untreated</td>
<td>3.1</td>
<td>2.8</td>
<td>3.5</td>
</tr>
<tr>
<td>1</td>
<td>IR Angiogram</td>
<td>N</td>
<td>2012-11-09</td>
<td>3</td>
<td>RT</td>
<td>saccular</td>
<td>ACOM</td>
<td>untreated</td>
<td>3.4</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>IR Angiogram</td>
<td>N</td>
<td>2012-11-09</td>
<td>4</td>
<td>RT</td>
<td>saccular</td>
<td>PCOM</td>
<td>untreated</td>
<td>2.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>MRA Brain W CE</td>
<td>N</td>
<td>2013-02-04</td>
<td>1</td>
<td>LT</td>
<td>fusiform</td>
<td>ACOM</td>
<td>coiling</td>
<td>6.8</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>1</td>
<td>CTA Brain</td>
<td>N</td>
<td>2013-02-05</td>
<td>1</td>
<td>LT</td>
<td>fusiform</td>
<td>ACOM</td>
<td>untreated</td>
<td>4.4</td>
<td>5.4</td>
<td>11.4</td>
</tr>
<tr>
<td>1</td>
<td>IR Angiogram</td>
<td>N</td>
<td>2013-02-05</td>
<td>2</td>
<td>RT</td>
<td>coiling</td>
<td>ACOM</td>
<td>coiling</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>CTA Brain</td>
<td>N</td>
<td>2013-02-05</td>
<td>3</td>
<td>RC</td>
<td>coiling</td>
<td>ACOM</td>
<td>coiling</td>
<td>2.2</td>
<td>2.7</td>
<td>2.4</td>
</tr>
</tbody>
</table>

**Measurement variability**
- 2012-10-04 19.1 x 19.6 x 24.9
- 2012-11-09 18 x ?? x 23

**Data quality**
- 2012-10-04 MRA 3.6 x 2.4
- 2012-11-09 Angio 2.1

**Missing information**
- 2013-02-05 ?? x ?? x ??
Application: Predictive Modeling

**Model** | **Precision** | **AUROC**
--- | --- | ---
Logistic Regression | 69% | --
Naïve Bayes | 70% | 0.61
Bayesian Belief Network | 75% | 0.70
Using Informatics to Understand Practice Patterns/Quality

UTILIZATION OF IMAGING IN THE ELDERLY
Deficit Reduction Act of 2005 (DRA) significantly reduced the utilization of imaging in Medicare-aged individuals.

The effect of DRA on downstream patient care is unclear.

Current studies have examined aggregate administrative data; details are needed to complete the picture.

Study Objectives

• Characterize the utilization of medical imaging for elderly patients based on data from a large academic medical center

1. **Exploratory analysis**: Understand who undergo imaging; what types of exams do they receive; why studies are performed

2. **Practice-related analysis**: Examine the practice patterns and downstream value of imaging for specific cohorts
• Collection of **retrospective** observational clinical data

• Primarily focused on radiology data
  – **Structured fields** with some variations: E.g., anatomy field has 116 values: BREAST, RTBREAST, LTBREAST
  – **Structured fields** with a large number of variations
  – **Unstructured fields** such as free-text narratives

• Studying a broad population
  – Need automated tools to clean dataset
Methodological Challenges

Electronic health record

• Large amount of detail captured about individual patients
• Mostly unstructured
• Limited scope of institutions

Administrative dataset

• Information summarized as diagnosis/billing codes (CPT, ICD-9, NDC)
• Structured
• Any institution requesting reimbursement

General challenges with **observational data:** (1) inaccurate; (2) incomplete; (3) unknown provenance; and (4) insufficient granularity
Cohort Identification: UCRex

UCRex: 1.05M
UCSF: 0.4M
UCD: 0.8M
UCSD: 0.4M
UCI: 0.2M
Methods

• **Inclusion criteria**
  - Individuals 65 and older
  - Outpatients seen at UCLA Ronald Reagan Medical Center & Santa Monica Hospital
  - 2007-2012

• **Data sources**
  - Radiology information system (imaging exam details)
  - Electronic health record (pathology results)

• **Data access**
  - Service request through CTSI BIP
  - IRB Protocol Submission
  - Compliance Request to Interface Form
Methods

Westwood
n = 213,831
records

Santa Monica
n = 116,637
records

Other
n = 68,753 records

RIS Database
(Stores all records)

Data Collection Dates:
January 2007 – December 2012

Patient
Age

Patient
Gender

Type of
Study

Indications

Diagnostic
Codes

Date of Exam

Anatomical
Body Region

CT
MR
MG
CR
US
RF
PET

BIRADS
score

Pathology
result

Sample size:
399,221 records
77,010 unique individuals
### Data Wrangling

#### 41462 rows

<table>
<thead>
<tr>
<th>ExamDesc</th>
<th>patientId</th>
<th>PatientType</th>
<th>Gender</th>
<th>AgeAtExam</th>
<th>ExamDesc</th>
<th>OrgCode</th>
<th>icd9</th>
<th>DepartmentDesc</th>
<th>assessment_desc</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM BILAT DIAG MAM</td>
<td>5460098</td>
<td>4443787</td>
<td>F</td>
<td>87</td>
<td>MM BILAT SCREENING</td>
<td>SMO</td>
<td>V76.12</td>
<td>SMO BREAST IMAGING</td>
<td>Benign finding</td>
</tr>
<tr>
<td>MM BILAT DIAG MAM</td>
<td>5650762</td>
<td>5740025</td>
<td>F</td>
<td>88</td>
<td>MM BILAT SCREENING</td>
<td>SMO</td>
<td>V76.12</td>
<td>SMO BREAST IMAGING</td>
<td>Benign finding</td>
</tr>
<tr>
<td>MM BILAT DIAG MAM</td>
<td>56652918</td>
<td>4680786</td>
<td>F</td>
<td>84</td>
<td>MM BILAT SCREENING</td>
<td>SMO</td>
<td>V76.12</td>
<td>SMO BREAST IMAGING</td>
<td>Need additional Imaging evaluation</td>
</tr>
<tr>
<td>MM BILAT SCREENING</td>
<td>5782466</td>
<td>4585100</td>
<td>F</td>
<td>87</td>
<td>MM BILAT SCREENING</td>
<td>SMO</td>
<td>V76.12</td>
<td>SMO BREAST IMAGING</td>
<td>Benign finding</td>
</tr>
<tr>
<td>MM BILAT SCREENING</td>
<td>5849705</td>
<td>7397025</td>
<td>F</td>
<td>77</td>
<td>MM BILAT SCREENING</td>
<td>MP</td>
<td>V76.12</td>
<td>MP2 BREAST IMAGING</td>
<td>Negative</td>
</tr>
<tr>
<td>MM BILAT SCREENING</td>
<td>5865346</td>
<td>4994183</td>
<td>F</td>
<td>68</td>
<td>MM BILAT SCREENING</td>
<td>MP</td>
<td>V76.12</td>
<td>MP2 BREAST IMAGING</td>
<td>Negative</td>
</tr>
<tr>
<td>MM BILAT SCREENING</td>
<td>5938005</td>
<td>4444527</td>
<td>F</td>
<td>81</td>
<td>MM BILAT SCREENING</td>
<td>MP</td>
<td>V76.12</td>
<td>MP2 BREAST IMAGING</td>
<td>Benign finding</td>
</tr>
<tr>
<td>MM BILAT SCREENING</td>
<td>5947892</td>
<td>4735097</td>
<td>F</td>
<td>75</td>
<td>MM BILAT SCREENING</td>
<td>MP</td>
<td>V76.12</td>
<td>MP2 BREAST IMAGING</td>
<td>Negative</td>
</tr>
<tr>
<td>MM BILAT SCREENING</td>
<td>6041371</td>
<td>5238389</td>
<td>F</td>
<td>66</td>
<td>MM BILAT SCREENING</td>
<td>MP</td>
<td>V76.12</td>
<td>MP2 BREAST IMAGING</td>
<td>Benign finding</td>
</tr>
<tr>
<td>MM BILAT SCREENING</td>
<td>6056115</td>
<td>5374580</td>
<td>F</td>
<td>86</td>
<td>MM BILAT SCREENING</td>
<td>MP</td>
<td>V76.12</td>
<td>MP2 BREAST IMAGING</td>
<td>Benign finding</td>
</tr>
<tr>
<td>MM BILAT SCREENING</td>
<td>6099907</td>
<td>5298052</td>
<td>F</td>
<td>67</td>
<td>MM BILAT SCREENING</td>
<td>MP</td>
<td>V76.12</td>
<td>MP2 BREAST IMAGING</td>
<td>Benign finding</td>
</tr>
<tr>
<td>MM BILAT SCREENING</td>
<td>6075004</td>
<td>4941083</td>
<td>F</td>
<td>67</td>
<td>MM BILAT SCREENING</td>
<td>MP</td>
<td>V76.12</td>
<td>MP2 BREAST IMAGING</td>
<td>Benign finding</td>
</tr>
<tr>
<td>MM BILAT SCREENING</td>
<td>6121148</td>
<td>6120121</td>
<td>F</td>
<td>82</td>
<td>MM BILAT SCREENING</td>
<td>MP</td>
<td>V76.12</td>
<td>MP2 BREAST IMAGING</td>
<td>Benign finding</td>
</tr>
<tr>
<td>MM BILAT SCREENING</td>
<td>6122935</td>
<td>5139557</td>
<td>F</td>
<td>74</td>
<td>MM BILAT SCREENING</td>
<td>MP</td>
<td>V76.12</td>
<td>MP2 BREAST IMAGING</td>
<td>Benign finding</td>
</tr>
<tr>
<td>MM BILAT SCREENING</td>
<td>6122906</td>
<td>5285079</td>
<td>F</td>
<td>68</td>
<td>MM BILAT SCREENING</td>
<td>MP</td>
<td>V76.12</td>
<td>MP2 BREAST IMAGING</td>
<td>Benign finding</td>
</tr>
<tr>
<td>MM BILAT SCREENING</td>
<td>6141503</td>
<td>5233562</td>
<td>F</td>
<td>82</td>
<td>MM BILAT SCREENING</td>
<td>MP</td>
<td>V76.12</td>
<td>MP2 BREAST IMAGING</td>
<td>Benign finding</td>
</tr>
</tbody>
</table>

**Cluster**

- 55 choices
- Sort by name / count
- Change

**Cluster**

- 12 choices
- Sort by name / count
- Change
## Data Wrangling

### Table: Dataset Preview

<table>
<thead>
<tr>
<th>ExamDesc</th>
<th>examld_deld</th>
<th>patientid_deld</th>
<th>PatientTypeCod</th>
<th>Gender</th>
<th>AgeAtExam</th>
<th>ExamDesc</th>
<th>OrgCode</th>
<th>idc8</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM RT DIAG MAM</td>
<td>1125</td>
<td>6540929</td>
<td>6442052</td>
<td>O</td>
<td>82</td>
<td>MM RT SCREENING MAM</td>
<td>MP</td>
<td>611.72</td>
</tr>
<tr>
<td>MM RT DIAG MAM</td>
<td>1230</td>
<td>6554852</td>
<td>7025796</td>
<td>O</td>
<td>80</td>
<td>MM RT SCREENING MAM</td>
<td>MP</td>
<td>793.61</td>
</tr>
<tr>
<td>MM RT DIAG MAM</td>
<td>1422</td>
<td>6570960</td>
<td>5320319</td>
<td>O</td>
<td>87</td>
<td>MM RT DIAG MAM</td>
<td>MP</td>
<td>174.98</td>
</tr>
<tr>
<td>MM US RT BREAST</td>
<td>1635</td>
<td>6568931</td>
<td>5308549</td>
<td>F</td>
<td>88</td>
<td>MM US RT BREAST</td>
<td>MP</td>
<td>611.72</td>
</tr>
<tr>
<td>MM US RT BREAST</td>
<td>1705</td>
<td>6595254</td>
<td>5223560</td>
<td>F</td>
<td>66</td>
<td>MM US RT BREAST</td>
<td>MP</td>
<td>793.61</td>
</tr>
<tr>
<td>MM US BLAT BREAST DIAGNOSTIC</td>
<td>1710</td>
<td>6599446</td>
<td>6578314</td>
<td>F</td>
<td>85</td>
<td>MM US BLAT BREAST DIAGNOSTIC</td>
<td>SUO</td>
<td>611.72</td>
</tr>
<tr>
<td>MM US LT BREAST</td>
<td>1843</td>
<td>6604263</td>
<td>4951373</td>
<td>O</td>
<td>67</td>
<td>MM US LT BREAST</td>
<td>MP</td>
<td>611.72</td>
</tr>
<tr>
<td>MM US RT BREAST</td>
<td>1870</td>
<td>6606203</td>
<td>4615540</td>
<td>F</td>
<td>82</td>
<td>MM RT SCREENING MAM</td>
<td>MP</td>
<td>611.72(793.81)</td>
</tr>
<tr>
<td>MM US RT BREAST</td>
<td>1879</td>
<td>6606919</td>
<td>6346295</td>
<td>O</td>
<td>69</td>
<td>MM US RT BREAST</td>
<td>MP</td>
<td>174.06(611.72)</td>
</tr>
<tr>
<td>MM US RT BREAST</td>
<td>2049</td>
<td>6623123</td>
<td>6348096</td>
<td>O</td>
<td>89</td>
<td>MM US RT BREAST</td>
<td>MP</td>
<td>611.72</td>
</tr>
<tr>
<td>MM US BLAT BREAST</td>
<td>2613</td>
<td>6660957</td>
<td>5284517</td>
<td>F</td>
<td>65</td>
<td>MM US BLAT BREAST</td>
<td>MP</td>
<td>611.72</td>
</tr>
<tr>
<td>MM US BLAT BREAST</td>
<td>2774</td>
<td>6680558</td>
<td>6516965</td>
<td>O</td>
<td>65</td>
<td>MM US RT BREAST</td>
<td>MP</td>
<td>174.6</td>
</tr>
</tbody>
</table>

### Additional Notes:

- **MM RT DIAG MAM**: Mammography Right Diagnosis Mammogram
- **MM RT DIAG MAM**: Mammography Left Diagnosis Mammogram
- **MM US RT BREAST**: Mammography Ultrasound Right Breast
- **MM US LT BREAST**: Mammography Ultrasound Left Breast
- **MM US BLAT BREAST**: Mammography Ultrasound Bilateral Breast
- **ExamDesc**: Description of exam type
- **OrgCode**: Code for organization
- **idc8**: Additional data code

### Related Functions:

- **Cluster**: Group similar data points together
- **Filter**: Apply filters to view specific data
- **Refresh**: Update the displayed data

### Data Source:

- **acr_breast_mammo_test.csv**
# Data Wrangling

## Google Refine

<table>
<thead>
<tr>
<th>ExamDesc</th>
<th>PatientId</th>
<th>PatientType</th>
<th>Gender</th>
<th>AgeAtExam</th>
<th>ExamDesc</th>
<th>OrgCode</th>
<th>icd9</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM LAT DIAG MAM</td>
<td>5303851</td>
<td>CALL BACK</td>
<td>F</td>
<td>69</td>
<td>MM LAT DIAG MAM</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
</tr>
<tr>
<td>MM BL LAT DIAG MAM</td>
<td>6284163</td>
<td>O</td>
<td>67</td>
<td>MM BL LAT DIAG MAM</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
<td></td>
</tr>
<tr>
<td>MM BL LAT DIAG MAM</td>
<td>7284113</td>
<td>O</td>
<td>77</td>
<td>MM BL LAT DIAG MAM</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
<td></td>
</tr>
<tr>
<td>MM US BLRT</td>
<td>4402610</td>
<td>F</td>
<td>72</td>
<td>MM US BLRT</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
<td></td>
</tr>
<tr>
<td>MM US BLRT</td>
<td>4543087</td>
<td>F</td>
<td>70</td>
<td>MM US BLRT</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
<td></td>
</tr>
<tr>
<td>MM US BLRT</td>
<td>6227814</td>
<td>F</td>
<td>81</td>
<td>MM US BLRT</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
<td></td>
</tr>
<tr>
<td>MM US BLRT</td>
<td>5318083</td>
<td>F</td>
<td>75</td>
<td>MM US BLRT</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
<td></td>
</tr>
<tr>
<td>MM US BLRT</td>
<td>7284113</td>
<td>F</td>
<td>77</td>
<td>MM US BLRT</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
<td></td>
</tr>
<tr>
<td>MM US BLRT</td>
<td>4695592</td>
<td>F</td>
<td>79</td>
<td>MM US BLRT</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
<td></td>
</tr>
<tr>
<td>MM US BLRT</td>
<td>5020096</td>
<td>O</td>
<td>71</td>
<td>MM US BLRT</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
<td></td>
</tr>
<tr>
<td>MM US BLRT</td>
<td>4742285</td>
<td>O</td>
<td>67</td>
<td>MM US BLRT</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
<td></td>
</tr>
<tr>
<td>MM US BLRT</td>
<td>5148712</td>
<td>O</td>
<td>71</td>
<td>MM US BLRT</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
<td></td>
</tr>
<tr>
<td>MM US BLRT</td>
<td>6002402</td>
<td>O</td>
<td>66</td>
<td>MM US BLRT</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
<td></td>
</tr>
<tr>
<td>MM US BLRT</td>
<td>6170752</td>
<td>O</td>
<td>68</td>
<td>MM US BLRT</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
<td></td>
</tr>
<tr>
<td>MM US BLRT</td>
<td>6102856</td>
<td>CALL BACK</td>
<td>F</td>
<td>66</td>
<td>MM US BLRT</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
</tr>
<tr>
<td>MM US BLRT</td>
<td>6904517</td>
<td>F</td>
<td>65</td>
<td>MM US BLRT</td>
<td>MP</td>
<td>V16.3 [939.81]</td>
<td></td>
</tr>
</tbody>
</table>
Imaging Modalities by Age Group
Indications

- Abdominal/GI
- Cancer
- Brain and Neuro
- Endocrine
- Cardiopulmonary
- Genitourinary
- Muskuloskeletal
- Screening
- Trauma
Indications

- Abdominal/GI
- Cancer
- Brain and Neuro
- Endocrine
- Cardiopulmonary
- Genitourinary
- Muskuloskeletal
- Screening
- Trauma
Breast Screening: Analysis

Cohort Characteristics

- 5,501 individuals ≥ 75 years received at least one screening exam
  - 3,190 (57.9%) underwent multiple screening exams
  - 456 days on average between screening exams
- 17.5% of individuals had a follow-up diagnostic exam

# Screening Exams by Age
Breast Screening: Analysis

BIRADS Breakdown

- BIRADS 1: 14%
- BIRADS 2: 52%
- BIRADS 3: 19%
- BIRADS 4: 10%
- BIRADS 5: 5%

Rad-Path Concordance

- Benign: 31%
- Malignant: 69%

Individuals who received a diagnostic screening exam
- 964 cases total

Examining cases with BIRADS 4 or higher
- Of 137 cases, 128 were biopsied (93.4%)
IMPRESSION:

**RIGHT BREAST:** CATEGORY 2. Benign findings. No evidence of malignancy. Normal annual screening mammography is recommended in 12 months.

**LEFT BREAST:** CATEGORY 5. Findings are suspicious. Ultrasound-guided core needle biopsy is recommended at this time.

**OVERALL ASSESSMENT - CATEGORY 5 - HIGHLY SUGGESTIVE OF MALIGNANCY - APPROPRIATE ACTION SHOULD BE TAKEN.**

**FINAL DIAGNOSIS:**

A. **BREAST, LEFT, 11 MM, 12 O’CLOCK, 4 CM FROM NIPPLE (NEEDLE CORE BIOPSY):**
   - **Invasive ductal carcinoma** with lobular features, Grade 3, (90% of biopsy) Modified Bloom and Richardson score 8 of 9; Tubule formation: 3; Nuclear pleomorphism: 3; Mitosis: 2
   - Ductal carcinoma in situ (DCIS), solid type with focal central necrosis, high nuclear grade (< 5% of biopsy)
   - Lymph/vascular invasion: not identified
   - Microcalcifications: not seen

---

**Detect headings using regular expressions**

**Extract keywords such as anatomy (breast), laterality (right/left) and assessment (Category 5/invasive ductal carcinoma)**

**Match pathology and radiology findings (based on temporal ordering, anatomy, and side)**

**Identify relevant sentences from section (e.g., Left breast: Category 5)**

**Classify diagnosis (e.g., invasive ductal carcinoma → malignant) using conditional random field (CRF)**

**Generate metrics such as positive predictive value based on matched radiology-pathology findings**
• Electronic health records provide a rich resource for conducting retrospective studies on population health
  – **Structured** fields (e.g., diagnostic codes) are readily available, **unstructured fields** (e.g., narrative text) much more challenging
  – Awareness of inherent issues in observational data: inconsistencies, systematic biases, incompleteness

• Considerations
  – Defining a standard set of data elements from the beginning
  – REDCap is a good start but has its limitations
  – Data collection workflow needs to be coordinated
  – Many informatics tools are available to assist in this process
Tools Discussed

• UC Rex
  – http://ucrex.org/

• Open Refine
  – http://openrefine.org/

• NLM MetaMap

• Apache cTAKES (UIMA)
  – https://ctakes.apache.org/

• Slicer
  – http://www.slicer.org/