Researcher Access to Clinical Data
Introduction to Clinical Research
July 25, 2013

Diane Alejo
Diana Gumas
Sam Meiselman
David Plaut
Chris Shafer
David Thiemann, MD

Where do researchers get data?

TODAY
- Electronic Patient Record (EPR) – 4M patients
- Allscripts / Sunrise Clinical Manager (JHH – inpatient)
- Meditech (Bayview)
- Many Departmental Systems (ED, OR, Anesthesia)
- Multiple result reporting systems (Lab, Rad, etc)
- IDX (professional fees)
- Casemix Datamart (diagnoses, procedures)

This is NOT an exhaustive list!

FUTURE – an “Epic” opportunity
Methods for Data Access

Historically: Researcher Negotiates Access
With Clinical System Technical Staff

-Logistically & technically challenging

Approaches for Data Access

Back end or informal access underestimates the complexity of the data and likely violates HIPAA and JHM PHI protection policies

Approved access points:

1. Clinical Research Management System (CRMS)
2. EPR2020 - researcher self-serve for patients on your studies
3. Center for Clinical Data Analysis (CCDA) Service
4. I2b2 cohort discovery tool
5. caTissue – Biospecimen data
1. Clinical Research Management System (CRMS)

CRMS Contains:
- Administrative data about your study & participants
- Can create research “Forms”

You can extract data from CRMS via
- Canned Reports (e.g. enrollment report, demographics)
- Ad-hoc querying using SQL database language

For more info:
CRMSHelp@jhmi.edu

2. EPR2020 Clinical Data Repository

EPR- JHH & JHBMC
4.4M Patients,
36.2M Visits
16.2M Documents
8.6M Rad Results
402M Lab Results
5.8M Inpatient Diagnoses
1.7M Problems

EPR2020
Read-Only
Data Discovery
For:
Clinical Care
Research
Quality Improvement

Clinical Research Management System
HDM/Casemix
JHCP
Other JHH/JHBMC
Clinical Systems
External Labs
Additional Hospitals

Patients on Study
Diagnoses/Procedures
FUTURE
### EPR2020 Research Patients on “My Studies”

**Research Support:**
- Research Patients
- Cohorts
- Clinical Detail
- Info

<table>
<thead>
<tr>
<th>Name</th>
<th>MNIN</th>
<th>Protocol ID #</th>
<th>CRMS #</th>
<th>Study Title</th>
<th>Study Status</th>
<th>Recent Encounter</th>
<th>Recent Location</th>
<th>Age</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA_000025</td>
<td>CRMS-0001</td>
<td>Carotid Artery</td>
<td>Active</td>
<td>08/09/2011</td>
<td>VAS</td>
<td>77</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA_000025</td>
<td>CRMS-0002</td>
<td>Carotid Artery</td>
<td>Active</td>
<td>08/09/2011</td>
<td>VAS</td>
<td>60</td>
<td>F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA_000025</td>
<td>CRMS-0003</td>
<td>Carotid Artery</td>
<td>Active</td>
<td>08/09/2011</td>
<td>VAS</td>
<td>70</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA_000025</td>
<td>CRMS-0004</td>
<td>Carotid Artery</td>
<td>Active</td>
<td>08/09/2011</td>
<td>VAS</td>
<td>86</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA_000025</td>
<td>CRMS-0005</td>
<td>Carotid Artery</td>
<td>Active</td>
<td>08/09/2011</td>
<td>VAS</td>
<td>86</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA_000025</td>
<td>CRMS-0006</td>
<td>Carotid Artery</td>
<td>Active</td>
<td>08/09/2011</td>
<td>VAS</td>
<td>86</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA_000025</td>
<td>CRMS-0007</td>
<td>Carotid Artery</td>
<td>Active</td>
<td>08/09/2011</td>
<td>VAS</td>
<td>86</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA_000025</td>
<td>CRMS-0008</td>
<td>Carotid Artery</td>
<td>Active</td>
<td>08/09/2011</td>
<td>VAS</td>
<td>86</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA_000025</td>
<td>CRMS-0009</td>
<td>Carotid Artery</td>
<td>Active</td>
<td>08/09/2011</td>
<td>VAS</td>
<td>86</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA_000025</td>
<td>CRMS-0010</td>
<td>Carotid Artery</td>
<td>Active</td>
<td>08/09/2011</td>
<td>VAS</td>
<td>86</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA_000025</td>
<td>CRMS-0011</td>
<td>Carotid Artery</td>
<td>Active</td>
<td>08/09/2011</td>
<td>VAS</td>
<td>86</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA_000025</td>
<td>CRMS-0012</td>
<td>Carotid Artery</td>
<td>Active</td>
<td>08/09/2011</td>
<td>VAS</td>
<td>86</td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA_000025</td>
<td>CRMS-0013</td>
<td>Carotid Artery</td>
<td>Active</td>
<td>08/09/2011</td>
<td>VAS</td>
<td>86</td>
<td>M</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### EPR2020 Research Lab Results View (Exportable!)

**Research Support:**
- Research Lab Results
- Cohorts
- Clinical Detail
- Info

<table>
<thead>
<tr>
<th>Result Date</th>
<th>Patient</th>
<th>MNIN</th>
<th>Study Title</th>
<th>Result</th>
<th>Test Name</th>
<th>Result Value</th>
<th>Reference Range</th>
<th>Result Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
<tr>
<td>05/07/2011</td>
<td>Aroni 1</td>
<td>638912</td>
<td>Leukocyte Count</td>
<td>Neutrophil</td>
<td>Neutrophil Percentage</td>
<td>490</td>
<td>500-1000</td>
<td>Final</td>
</tr>
</tbody>
</table>
Researcher App View: Diagnosis

“Find my patients who had a cardiac diagnosis in the 1st quarter of 2010”

Provides assistance gathering:

• Preliminary, anonymous data for feasibility, grant applications and statistical sample-size estimates
• IRB-approved case-finding-for study enrollment (mailings, phone solicitation), chart review, and cohort/case-control studies
• Research data extracts-monthly/quarterly integrated extracts from EPR, POE, ORMIS, lab/PDS, billing systems, vaccination/transfusion/culture data, etc.

Electronic Patient Record (EPR) and Research

Can look at statistics for race, gender, age, diagnosis description
EPR2020 Search patient’s documents for key term

For Access to EPR2020 Contact:

Diana Gumas: dgumas1@jhmi.edu
3. Center for Clinical Data Analysis (CCDA)

Provides periodic (monthly/quarterly) bulk data extracts

- **Preliminary, anonymous data** for feasibility, grant applications and statistical sample-size estimates
- **IRB-approved case-finding**—for study enrollment (mailings, phone solicitation), chart review, and cohort/case-control studies
- **Research data extracts** - monthly/quarterly integrated extracts from EPR, POE, ORMIS, lab/PDS, billing systems, vaccination/transfusion/culture data, etc.

How CCDA works

**For IRB-approved research:**
- Provide full protocol + IRB approval
- Meet to discuss query methods, format
- Iterate, then schedule production (email extracts, Jshare)
- Hourly cost

**For non-IRB projects (exploratory analyses, QI)**
- Same process, cost subsidized by ICTR/JHM (?)
- Do NOT implicitly morph QI into IRB

For more information, contact:
David Thiemann - dthiema1@jhmi.edu
Structuring a data request

Who?
List of MRNs or cohort defined by characteristics?

What?
Procedures? PCP encounter? Diagnosis? Free Text? Lab results?
Repeat: Procedures? Diagnosis? Free Text Search? Lab results?

When?
Age of individual? Date/range of event?
Does another thing happen in a certain time frame?

Where?
Zip code? Unit? Floor? Hospital? PCP location?

Why or How?
Structuring your data request clarifies your thinking and improves communication with the data analyst

Data Sources

- EMR Data
- Financial Data
- Registry Data
- State / National / Public data sets
- Research study data sets

JHM has a great deal of data:
- It is structured for source production systems
- Challenging to data mine
- Very difficult to join across systems.
- Inappropriate access can lead to significant legal implications
### Johns Hopkins Medicine Clinical Systems Environment

<table>
<thead>
<tr>
<th></th>
<th>JHH</th>
<th>JHUBMC</th>
<th>HGH</th>
<th>Suburban</th>
<th>JHCOP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient Portal</strong></td>
<td>EPIC MyChart</td>
<td>EPIC MyChart</td>
<td>EPIC MyChart</td>
<td>EPIC MyChart</td>
<td>EPIC MyChart</td>
</tr>
<tr>
<td><strong>Registration / Scheduling</strong></td>
<td>EPIC</td>
<td>Meditech / Epic, GE (IDX)</td>
<td>Meditech</td>
<td>McKesson</td>
<td>GE (IDX)</td>
</tr>
<tr>
<td><strong>ADT</strong></td>
<td>Keane</td>
<td>Meditech</td>
<td>Meditech</td>
<td>McKesson</td>
<td></td>
</tr>
<tr>
<td><strong>Inpatient EMR</strong></td>
<td>Allscripts</td>
<td>Meditech</td>
<td>Meditech</td>
<td>McKesson</td>
<td></td>
</tr>
<tr>
<td><strong>Ambulatory EMR</strong></td>
<td>EPIC</td>
<td>EPIC</td>
<td>EPIC</td>
<td>EPIC</td>
<td>EPIC</td>
</tr>
<tr>
<td><strong>Specialties</strong></td>
<td>Varied</td>
<td>Varied</td>
<td>Varied</td>
<td>Varied</td>
<td></td>
</tr>
<tr>
<td><strong>EPR2020</strong></td>
<td>Enterprise Clinical Data and Document repository</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hospital Billing</strong></td>
<td>Keane</td>
<td>Meditech</td>
<td>Meditech</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Professional Billing</strong></td>
<td>GE (IDX)</td>
<td>GE (IDX)</td>
<td>GE (IDX)</td>
<td>GE (IDX)</td>
<td>GE (IDX)</td>
</tr>
</tbody>
</table>

---

### JHM Data Source Systems

<table>
<thead>
<tr>
<th>System</th>
<th>Contact</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPR2020 / EPR</td>
<td>CCDA Diana Gumas</td>
<td>Some search functionality. Integration with CRMS. Data mining is SQL based hand work. Tables are very complex and require domain expertise. Usually requires an IRB request and funding source.</td>
</tr>
<tr>
<td>Allscripts (SCM)</td>
<td>CCDA Marty Hamburg</td>
<td>Data mining is SQL based hand work. Tables are very complex and require domain expertise. Usually requires an IRB request and funding source.</td>
</tr>
<tr>
<td>Meditech</td>
<td>Andy Frake</td>
<td>Reporting capability</td>
</tr>
<tr>
<td>Epic</td>
<td>Jim Ham</td>
<td>TBD process, outpatient and Community Division data</td>
</tr>
<tr>
<td>Casemix</td>
<td>David Plaut</td>
<td>Billing data, requires access permissions, some standard reports</td>
</tr>
<tr>
<td>Speciality System Owner</td>
<td>ORMS, Metavision, TheraDoc, EBB, etc</td>
<td></td>
</tr>
<tr>
<td>I2b2</td>
<td>Sam Meiselman</td>
<td>Counts, deidentified data, pre-IRB</td>
</tr>
</tbody>
</table>
Types of data

- Clinical Data – Labs, Radiology, Clinical Documentation
- Casemix
- Billing - what items were charged to the patient, quantity, charge $ (not cost), date and location at time of charge
- Payment data - $ paid against a bill by a paper, not the $ paid per item on the bill
- Costs (limited) – what did an item cost vs what we charged
- Pharmacy dispensing and administration detail (JHH only)
- Census & transfers (where are patients)
- Readmissions
- Outpatient Scheduling

Coding schemes

<table>
<thead>
<tr>
<th>Code Type</th>
<th>Inpatient or Outpatient</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-9 Diagnosis</td>
<td>In &amp; Out</td>
</tr>
<tr>
<td>ICD-9 Procedures</td>
<td>In</td>
</tr>
<tr>
<td>CPT</td>
<td>Out</td>
</tr>
<tr>
<td>MS DRG</td>
<td>In</td>
</tr>
<tr>
<td>APR DRG</td>
<td>In</td>
</tr>
<tr>
<td>CPC DRG (Charge/Case, MD only)</td>
<td>In</td>
</tr>
<tr>
<td>APG</td>
<td>Out</td>
</tr>
<tr>
<td>CPV APG (Charge/Visit, MD Only)</td>
<td>Out</td>
</tr>
<tr>
<td>RPG (charge item catalog)</td>
<td>In &amp; Out</td>
</tr>
<tr>
<td>ICD-10 coming soon...</td>
<td></td>
</tr>
</tbody>
</table>
Financial vs. Clinical data

- All coding schemes are financially driven
- There may be a diagnosis code for a disease that the patient does not actually have, however, resources were expended to rule that diagnosis out.
- Coding is driven by physician documentation

Challenges

- Silos of Data implemented over time
  - Enterprise systems
    - EPIC
    - EPR / EPR 2020
    - Allscripts (Sunrise Clinical Documentation)
    - HMED (Emergency Room)
    - ORMIS (GE Centricity Periop Manager)
    - MetaVision (Anesthesia Documentation)
    - Casemix / Datamart / HDM
    - Theradoc
    - Electronic Bed Board (EBB)
    - OB Documentation (GE Centricity Perinatal)
    - Labs, Radiology and Pharmacy source system
- Patient case identification and matching between systems
Some Lessons Learned

Data Mining
1. Must have strong clinical leadership and pervasive buy-in for research project
2. Must understand clinical workflow and source documentation to translate to the research study
3. Must stay within scope of readily known data
4. Must audit for completeness and accuracy
5. Retrospective Data – Important Caveats
   - Data Dictionary / Version Control
   - Documentation Consistency over time and source systems
   - Documentation Purpose (Billing vs Clinical Documentation)
   - Discovery of new diagnosis – recoding historical data
   - Clinical Relevance and Interpretation (time series data, lab and medication data)

Other Internal Sources for Research Data

National Registries
- Maryland Trauma Registry
- Tumor Registry
- UNOS Registry (Transplants)
- NSQIP
- VQI – Vascular Quality Initiative

National CV Registries
- STS Adult Cardiac Surgery
- STS Congenital Cardiac Surgery Registry
- STS Thoracic Registry
- STS Cardiac Anesthesia Registry
- STS/ACC TVT Registry
- INTERMACS VAD Registry
- PEDIMACS VAD Registry
- ELSO Registry (ECLS Registry)
- PC4 Registry (Pediatric Cardiac Critical Care)
- ACC NCNR Action Registry - GWTG
- ACC NCNR IMPACT (Ped & Adult Congenital)
- ACC ICD (Generator & Leads)
- ACC Cath PCI Registry

Specialty Research Databases
- School of Medicine
- School of Public Health
- Kennedy Kreiger Institute

Data Registries and Internal Database Initiatives
What will Epic do for researchers?

Research Encounter Scheduling
Better Research Billing
Electronic research orders/preference lists
Other studies for which your patient is a participant
Report that patient admitted or arrived at ED
Data querying tools accessible by Epic user
Patient Reported Outcomes via MyChart patient portal
Registries from data collected in Epic
Access to codified data across the enterprise
Epic Data

- Ambulatory, HCGH, Sibley are on-line
- Some historical data migrated to Epic. TBD research utility
- JHBMC and JHH remain with legacy systems for near term future
- Challenge will be to join inpatient and Epic-data
- Data access for researchers discussions are underway
- Several proposed points of access:
  - Department based – TBD readiness
  - I2b2 – pilot available now [http://i2b2.johnshopkins.edu/](http://i2b2.johnshopkins.edu/)
  - Center for Clinical Data Analysis
  - Epic reporting team – TBD readiness

How will researchers get access to Epic data for research?

Still under discussion...

Likely to be services provided by CCDA
Clinical Data Environment

Legacy Data
• J-HH
• J/HBMC

Epic Ambulatory Data
Opportunities
• Detailed inpatient data
• Medications
• Readmission notification
• Practice variation

Risks
• Not integrated
• Transitions of care

Opportunities
• Registries for Chronic Disease and Wellness
• Tools for clinicians, PCP staff, and care coordinators
• Data via Clarity export

Care data – accessible via Epic

Epic Community Division Data
Opportunities
• Inpatient, Meds and ED visits available for care management
• Readmission alert
• Practice variation
• Data via Clarity export

Risks

JHHC Systems ACO-like
Opportunities
• Care coordination
• Management of Risk
• "what if" analysis
• ACO reporting
• How are we doing?
• New business

Additional Data
• EAHPS / MCAHPS
• CMS data sets
• 3rd party or pilot data sets--ie High Value Health Collaborative
• CRISP
• Epic extensions to non JHM practices and hospitals

EMR Data – Chronic Disease – Planned Business Objects Universes

<table>
<thead>
<tr>
<th>Diagnoses</th>
<th>Chronic Disease Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Alcohol &amp; Tobacco use - Diabetes</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Comorbidities - diabetes</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Numeric - Comorbidities - Diabetes</td>
<td>Comorbidities - Diabetes</td>
</tr>
<tr>
<td>Communication - Diabetes</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Medications - Diabetes</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Numeric - Diabetes</td>
<td>Medications - Diabetes</td>
</tr>
<tr>
<td>Most Recent Labs - Diabetes</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Pregnancy - Diabetes</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Visit History - Diabetes</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Vital - Diabetes</td>
<td>Diabetes</td>
</tr>
</tbody>
</table>

628 Objects / 47 tables / 1528 Rows across CHF, Diabetes, Hypertension Registries

**Epic will be a phenomenal tool, and . . .**

- **Set up required**: We need to check for Clarity, BOE, and future Data Warehouse connections between tables
- **Alignment**: We need to align our data access policies with JHM-available tools
- **Deidentified Access**: I2b2 can act as an interim bridge to access deidentified Epic data if interface connections are approved
- **Interim Data Warehouse**: I2b2 is currently acting as a synthetic data warehouse until the Epic EDW is funded and brought on line.
And finally, some wise words from Dr. David Thiemann

- Real work, not ad hoc/bootstrap
- Need $$$ and FTE(s)
- Smart analyst(s) who know database technology and understand (or can learn) nuances of the sources and content domain
- Hands-on PI management/guidance
- Statistical liaison early, before database schema and ETL methods are set in stone

The Extract-Transform-Load process: Getting Clinical Data into Research DB

- Raw clinical/administrative data is limited and must be used with caution in research
- Build an intermediate (staging) database
  - Don’t do data management in SAS/Stata/Excel
- Data dictionary—derivation for each field
- Templated, tested, documented cleanup scripts/routines.
- Intermediate tables: Log each step/modification
  - For each batch, be able to re-create data transform from scratch
  - Version control, change control and documentation are vital
  - Build data versioning into the database
Transforming Data is Non-Trivial

- Raw data typically string (char/text) fields
- Unanalyzable characters (* < >, comments) still have meaning
  - Put non-numeric data in separate field. Avoid numerical recoding (999)
- ~3% of pts have multiple/non-preferred MRNs
  - Need 1-to-many link table
- Assays/reference ranges/coding changes
  - Avoid using raw codes (CPT/ICD) in research db
  - Map clinical codes to research terms
- Defer analytic assumptions. When recoding data, anticipate problems. Keep options open.

Understanding Business Logic

Trust but verify: Test coding accuracy
- Providers may habitually use imprecise/inaccurate diagnosis codes (especially in profee data)
- ICD9 procedure indications often a billing fiction
- Trained coders may make systematic errors
- Different content domains may have different standards (inpt vs outpt coders)
- Don’t infer/assume dependencies unless enforced by source system.

Run min/max queries, aggregates, outer joins
- Confirm date ranges, data ranges, relative proportions by year
- For ex. Lab results on deceased patients may be legitimate

Don’t assume that null rows actually are empty. Maybe the query missed something.
4. caTissue for Collecting & Accessing Biospecimen Data

What caTissue can do for you
• Track collection and storage of specimens
• Clinical annotations
• Track distribution of specimens
• Derivation and Aliquoting of specimens
• Reports

Cost
• No Charge for repositories using basic services
• Hourly charge for legacy data migration, special reports & features

For more information: Diana Gumas dgumas1@jhmi.edu

---

i2b2 BackGround

www.i2b2.org
• Informatics for Integrating Biology & the Bedside
• Simple Tool Yielding Accelerated Data Access
• Deidentified Cohort Builder
• Query “Articulator”
I2b2

I2b2 pilot is a proven path –
- https://www.i2b2.org/
- Cohort Finding Tool
- Some basic data visualization tools
- Flexible and simple Data Mart Schema
- Customizable Ontologies
- Many open source additional modules
- SHRINE as a possible link to non-JHM institution data sets

Shorten cycle time from idea to data for analysis

Self Service + Collaboration

i2b2 at JHMI

Landing Page http://i2b2.johnshopkins.edu/
- Request Access
- Client Launch
- System Information and Announcements
- Help Documentation and Support Links
i2b2 at JHMI

Architecture
  – Database
    • Fed from JHMI Source Systems
      – EPR2020
      – SCM
    • Deidentified for i2b2 Users
    • Re-"Identifiable" for DBAs
  – Web Application
    • JHED authorized security
    • Web application is only way for Users to access data

i2b2 at JHMI Data Content

Dimensions
  – Patients/ Demographics
  – Visits/ Encounters
  – Providers

“Facts” (clinical findings)
  – Numeric Labs
  – Inpatient and Outpatient Diagnosis*
  – Inpatient Procedures
  – Inpatient Medications (in progress)
### i2B2 inventory

<table>
<thead>
<tr>
<th>Data</th>
<th>Status</th>
<th>Q</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Mapping and Dimension</td>
<td>Completed</td>
<td>2.5M</td>
<td>Patients and visits based on EPR2020 identity backbone including multi-institution and multi-MRN references.</td>
</tr>
<tr>
<td>Visit mapping and dimension</td>
<td>Completed</td>
<td>42M</td>
<td>Needs validation of inpatient v. outpatient flags</td>
</tr>
<tr>
<td>Provider dim.</td>
<td>Completed</td>
<td>0.35M</td>
<td>Contents of EPR/EPR2020 (CDT provider)</td>
</tr>
<tr>
<td>Provider Ontology</td>
<td>In progress</td>
<td></td>
<td>JHH and Bayview providers</td>
</tr>
<tr>
<td>Demographics</td>
<td>Completed</td>
<td>2.5M</td>
<td>Includes: Age, zipcode, race, gender</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Completed</td>
<td>40M</td>
<td>Based on ICD9 codes from datamart and IDX via EPR2020. Some gaps identified</td>
</tr>
<tr>
<td>Inpatient Procedures</td>
<td>Completed</td>
<td>2M</td>
<td>Based on ICD9 codes from datamart and IDX via EPR2020.</td>
</tr>
<tr>
<td>Outpatient Procedures</td>
<td>In Progress</td>
<td></td>
<td>Requires a CPT lookup table and procedure fact records</td>
</tr>
<tr>
<td>Lab Data</td>
<td>In Progress</td>
<td>3Years</td>
<td>Ontology and Numeric results complete. Standard Text results, complex test results in progress</td>
</tr>
<tr>
<td>Meds</td>
<td>In Progress</td>
<td>TBD</td>
<td>Dispensed meds from SCM (JHH only for now). Current plan is to use Multim ontology until rxNorm can be added.</td>
</tr>
</tbody>
</table>

### i2b2 at JHMI

<table>
<thead>
<tr>
<th>Category</th>
<th>Record Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Findings (lab, diagnosis etc)</td>
<td>196156286</td>
</tr>
<tr>
<td>Inpatient and Outpatient Visits</td>
<td>45329952</td>
</tr>
<tr>
<td>Patients</td>
<td>4781047</td>
</tr>
<tr>
<td>Providers</td>
<td>35073</td>
</tr>
<tr>
<td>Ontology Concepts</td>
<td>29689</td>
</tr>
</tbody>
</table>
i2b2 and CCDA

Customer Self Service via i2b2
  – Feasibility
  – Query Articulation

CCDA Extracts/ Data Support Extend Reach
  – Ongoing or One Time Extracts
  – Variety of Outputs
  – Process of Intake, Requirements, Development

Case Studies

• Richard Moore HIV Studies
  – Ongoing Database Extracts
  – Support Large Clinical Review Studies
S. Melzer ACS Review: Barrett’s Esophagus

- 4.4 M patients and 26M visits
- 12.8 M documents / 7.2 M Radiology Reports
- 28M Lab Results

<table>
<thead>
<tr>
<th>Key Group</th>
<th>Patients</th>
<th>Visits</th>
<th>Documents</th>
<th>Radiology Reports</th>
<th>Lab Results</th>
<th>Access – Future Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Diagnoses</td>
<td>7775</td>
<td>26034</td>
<td>752,130</td>
<td>6,168,130</td>
<td>2,580,990</td>
<td></td>
</tr>
<tr>
<td>Barrett esophagus (E51.9)</td>
<td>4875</td>
<td>16,351</td>
<td>1,522,846</td>
<td>12,328,846</td>
<td>5,688,246</td>
<td></td>
</tr>
<tr>
<td>Barrett in DPO Problem List</td>
<td>1800</td>
<td>7753</td>
<td>64,300</td>
<td>528,300</td>
<td>234,400</td>
<td></td>
</tr>
<tr>
<td>Intestinal metaplasia of stomach (E51.9)</td>
<td>54</td>
<td>17</td>
<td>3,854</td>
<td>31,264</td>
<td>14,110</td>
<td></td>
</tr>
<tr>
<td>Intestinal metaplasia of small intestine (E52)</td>
<td>2485</td>
<td>12,140</td>
<td>1,540,510</td>
<td>12,910,510</td>
<td>4,751,310</td>
<td></td>
</tr>
<tr>
<td>Intestinal metaplasia of esophagus (E51.9)</td>
<td>2466</td>
<td>1,246</td>
<td>48,527</td>
<td>392,527</td>
<td>191,237</td>
<td></td>
</tr>
<tr>
<td>Intestinal metaplasia of lower third of stomach (E51.9)</td>
<td>786</td>
<td>525</td>
<td>20,000</td>
<td>160,000</td>
<td>60,000</td>
<td></td>
</tr>
<tr>
<td>Intestinal metaplasia of middle third of stomach (E51.9)</td>
<td>1500</td>
<td>73</td>
<td>500</td>
<td>4,100</td>
<td>1,500</td>
<td></td>
</tr>
<tr>
<td>Intestinal metaplasia of other sites (E51.9)</td>
<td>718</td>
<td>4,200</td>
<td>154,000</td>
<td>1,230,000</td>
<td>500,000</td>
<td></td>
</tr>
<tr>
<td>Intestinal metaplasia of thoracic esophagus (E51.9)</td>
<td>1680</td>
<td>86</td>
<td>1,000</td>
<td>8,000</td>
<td>2,000</td>
<td></td>
</tr>
<tr>
<td>Intestinal metaplasia of upper third of esophagus (E51.9)</td>
<td>1480</td>
<td>64</td>
<td>150</td>
<td>1,200</td>
<td>300</td>
<td></td>
</tr>
</tbody>
</table>

Access – Future Concept

Preparatory for research
Investigator obtains # of subjects in cohort by querying CRDW managed data mart

De-Identified data
Coded data
Limited data set
Contact Information
Identified cohort

Investigator requests data
Data request deemed feasible by CRDW staff

<subject>, <children, or HIV/AIDS data
IRB approval

Individual consent for HIV or AIDS data
Individual consent

Data Use Committee approval and prioritization
Data received by Investigator
Data received by Investigator in Secure Data Zone
Patients + Data = Insight

Data / Insight

JHM Patients who have a diagnosis of Hypertension or CHF or family history

Non-JHM Patients who have a similar diagnosis, CMS or research partner data

Outpatient visit-Diagnoses

Cardiology Procedures

Cost / LOS / CMI

Genomic Data – JHM patients

Genomic Data – Reference / Non-JHM patients