Clinical and Translational Research Needs as an Exemplar of Data Warehousing for Informatics

Russ Waitman, PhD
Director of Medical Informatics,
Associate Professor, Department of Biostatistics
Director, Frontiers Biomedical Informatics
Assistant Vice Chancellor, Enterprise Analytics
University of Kansas Medical Center
Kansas City, Kansas

This project is supported in part by NIH grant UL1TR000001.
Outline

• A Fishing Story
• What is Biomedical Informatics?
• What are the Clinical Translational Science Awards?
• Informatics Aims with a Integrated Data Repository as Center
• Technical Details and Development Approach
• Questions/Demo
Biomedical Informatics Can Help Your Research

• We have tools and expertise to manage data and convert it into information

• **REDCap** and **CRIS** – enter and manage data

• **HERON** – **fish for data** from the hospital/clinic

• **Biweekly** Frontiers Clinical Informatics **Clinics**
  – Tuesday 4-5 pm in 1028 Dykes Library.
  – Next session December 11.
You’re that fisherman: wanting to land data to answer your research hypothesis

Bennett Spring Trout Park, Lebanon Missouri
http://mdc.mo.gov/regions/southwest/bennett-spring
The Fish: Diagnoses, Demographics, Observations, Treatments
Why so many fish? Medical Informatics

**Current Goal:** Build Hatchery, Manage the Fishery
Second Goal: If you need help fishing, hire a guide from Medical Informatics

Photo Credit: HuntFishGuide.com
http://www.flickr.com/photos/huntfishguide/5883317106/
Prepare and Analyze data with Biostatistics
Our shared goal: a tasty publication

Photo Credit: Steve Velo
http://www.flickr.com/photos/juniorvelo/259888572/
Nightmare: looks like a nice river, but can’t catch fish

- I’ll just enter everything in Excel…
- What if I lose or accidentally sort my spreadsheet?
- How to I let students only review de-identified data?
- Prevent the wrong people (statistician/student) from entering/changing data?

- Hospital/Clinic is making me use this Electronic Medical Record and I get nothing in return...

Little White Salmon River, Washington State, last Summer in July
Sometimes, You’re willing to enter data/buy fish:

**REDCap**: Research Electronic Data Capture

- [https://redcap.kumc.edu](https://redcap.kumc.edu)
  - It uses the same username and password as your KUMC email.
    - Non-KUMC researchers can request an affiliate account through Frontiers CTSA office
  - Check out the training materials under videos
  - Case Report Forms and Surveys

- For consultation and to move project to production: Register your project with us so we can keep track of your request.
  - [http://frontiersresearch.org/frontiers/biomedical-informatics](http://frontiersresearch.org/frontiers/biomedical-informatics)

- Check out other institutions using REDCap and possibly borrow from the master library.
  - [http://www.project-redcap.org/](http://www.project-redcap.org/)
REDCap Case Report Form Example
REDCap Survey: Think SurveyMonkey

Vexed by a complicated protocol in a vulnerable population?
Want to go beyond compliance hurdles to sleeping well at night?

Frontiers CTSA brings you an exciting opportunity to obtain thoughtful ethics consults to further your research.
Please take a minute to fill our your request so we can direct the best ethicist to work with you on your urgent moral quandary.

<table>
<thead>
<tr>
<th>First name:</th>
<th>[ ] must provide value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last Name</td>
<td>[ ] must provide value</td>
</tr>
<tr>
<td>email</td>
<td>[ ] must provide value</td>
</tr>
</tbody>
</table>

Where are you?
[ ] UMKC  [ ] Children’s Mercy  [ ] KCUMB
[ ] KUMC-W  [ ] KUMC-KC  [ ] KSU-L  [ ] UM
[ ] Lukes  [ ] VA  [ ] Other

Role
[ ] Principal investigator
[ ] Co-investigator
[ ] Study Coordinator
[ ] IRB member
[ ] Other

Is your question regarding any of the common ethics topics:
[ ] Informed Consent
[ ] Inclusion/Exclusion criteria
[ ] Recruitment
[ ] Vulnerable populations

Add any additional notes or comments here

Upload your IRB protocol or a copy of your credit card here

Submit
REDCap: think Fish Tank you manage

http://www.flickr.com/photos/wiccked/185270913/lightbox/
I want to go fishing, not fill a fish tank (REDCap)
Use HERON: a managed fishery
Aim #2: Create a data “fishing” platform: HERON, https://heron.kumc.edu

- **Get a License:** Develop business agreements, policies, data use agreements and oversight.

- **Get a Fishing Rod and Bass Boat:** Implement open source NIH funded (i.e. i2b2 https://www.i2b2.org/) initiatives for accessing data.

- **Know what your catching:** Transform data into information using the NLM UMLS Metathesaurus as our vocabulary source.

- **Stock Different Tasty Fish:** link clinical data sources to enhance their research utility.
HERON: Getting a Fishing License

- Fill out System Access Agreements to sponsor students/staff
- Fill out Data Use Agreement to request data export
- **No Limit!!!** IRB Protocol Not Required to view or pull de-identified data
- Must be on campus or use VPN
- Check [http://informatics.kumc.edu/work/blog](http://informatics.kumc.edu/work/blog) for latest status
The i2b2 “Fishing Rod”: build Diabetes cohort

Types of “fish” in folders

Drag concepts from upper left into panels on the right
i2b2: **AND** in Frontiers Research Registry

Dragging over the second condition
i2b2: **AND** a high Hemoglobin A1C

When you add a numeric concept, i2b2 asks if you want to set a constraint.
i2b2 Result: 497 patients in Cohort

Run the Query
Query took 4 seconds
497 patient in cohort
The dream: landing the big one

http://www.oregon.com/columbia_gorge_attractions/bonneville_hatchery

Catch the data for JAMA, NEJM publication
Without getting bit

!!CAUTION!!

TROUT BITE

DO NOT PUT YOUR HANDS OR FINGERS IN THE WATER
• The Fundamental Theorem of Biomedical Informatics:
  – A person working with an information resource is better than that same person unassisted.

  – NOT!!
Background: Bill Stead

Evidence → Clinician → Patient Record → Synthesis & Decision

William Stead: http://courses.mbl.edu/mi/2009/presentations_fall/SteadV1.ppt
The demise of expert based practice is inevitable.

Human Cognitive Capacity

Facts per Decision

1990 2000 2010 2020

Decisions by Clinical Phenotype

- Structural Genetics: e.g. SNPs, haplotypes
- Functional Genetics: Gene expression profiles
- Proteomics and other effector molecules

Dan Masys, William Stead: http://courses.mbl.edu/mi/2009/presentations_fall/SteadV1.ppt
Biomedical Informatics Research Areas

- Machine learning
- Text interpretation
- Knowledge engineering

- Knowledge Acquisition
  - Biomedical Knowledge
  - Biomedical Data

- Data Acquisition
  - Knowledge Base
  - Inferencing System

- Data Base

- Real-time acquisition
  - Imaging
  - Speech/language/text
  - Specialized input devices

- Knowledge Engineering
  - Model Development
  - Information Retrieval
  - Diagnosis
  - Treatment Planning
  - Human Interface
  - Teaching
  - Image Generation

“It is the responsibility of those of us involved in today’s biomedical research enterprise to translate the remarkable scientific innovations we are witnessing into health gains for the nation.”
NIH Goal to Reduce Barriers to Research

- Administrative bottlenecks
- Poor integration of translational resources
- Delay in the completion of clinical studies
- Difficulties in human subject recruitment
- Little investment in methodologic research
- Insufficient bi-directional information flow
- Increasingly complex resources needed
- Inadequate models of human disease
- Reduced financial margins
- Difficulty recruiting, training, mentoring scientists
The purpose of this initiative is to assist institutions to forge a uniquely transformative, novel, and integrative academic home for Clinical and Translational Science that has the consolidated resources to:

1) captivate, advance, and nurture a cadre of well-trained multi- and inter-disciplinary investigators and research teams;

2) create an incubator for innovative research tools and information technologies; and

3) synergize multi-disciplinary and inter-disciplinary clinical and translational research and researchers to catalyze the application of new knowledge and techniques to clinical practice at the front lines of patient care.
1. Provide a HICTR portal for investigators to access clinical and translational research resources, track usage and outcomes, and provide informatics consultative services.

2. Create a platform, HERON (Healthcare Enterprise Repository for Ontological Narration), to integrate clinical and biomedical data for translational research.

3. Advance medical innovation by linking biological tissues to clinical phenotype and the pharmacokinetic and pharmacodynamic data generated by research cores in phase I and II clinical trials (addressing T1 translational research).

4. Leverage an active, engaged statewide telemedicine and Health Information Exchange (HIE) effort to enable community based translational research (addressing T2 translational research).
Aim 1 Example:

**REDCap: Research Electronic Data Capture**

- Added REDCap self service model to complement Velos Clinical Trial Management System
- For the last 2 quarters KUMC has some of the highest growth in new users nationally (> 700)
- Low barrier to entry, no extra passwords low management cost (<1 FTE)
- But, people are free to screw up their research
- **Much easier to manage than a Data Warehouse team**
Aim #2: Create a data “fishing” platform: HERON, https://heron.kumc.edu

• **Get a License:** Develop business agreements, policies, data use agreements and oversight.

• **Get a Fishing Rod and Bass Boat:** Implement open source NIH funded (i.e. i2b2 https://www.i2b2.org/) initiatives for accessing data.

• **Know what your catching:** Transform data into information using the NLM UMLS Metathesaurus as our vocabulary source.

• **Stock Different Tasty Fish:** link clinical data sources to enhance their research utility.
Develop business agreements, policies, data use agreements and oversight

- September 2010 the hospital, clinics and university signed a master data sharing agreement to create the repository.
  - Executive Committee – decides organization/systems expansion
  - Data Request Oversight Committee – guides implementation and approves/monitors use.

- Use Cases:
  - After signing a system access agreement, cohort identification queries and view-only access is allowed but logged and audited
  - Requests for de-identified patient data, while not deemed human subjects research, are reviewed.
  - Identified data requests require approval by the Institutional Review Board prior to data request review.
  - Contact information from the Frontiers Participant Registry have their study request and contact letters reviewed by the Participant and Clinical Interactions Resources Program
HERON Repository Architecture

Participants Clinical Systems (EPIC, IDX, VELOS)

Information in files from Source Systems

(Ex: archived database extracts or HL7 messages)

Extract, Transform & Load Processes or HL7 Listeners

Identified staging database (Night HERON)

Identified data server

De-identified staging database (Blue HERON)

De-identification & Transform Processes

Clinical/Translational Researcher

I2b2 clinical business intelligence application (JBoss, VMWare virtualized host managed by KUMC Information Resources)
Most Clinical Systems focus on transaction processing for workflow automation.
Technical Details

• Separate servers (HP DL180s) for identified and de-identified data.
  • In the same data center as clinical organizations systems.
  • Supported by KUMC Information Resources
• Using SUSE Linux (enterprise license & support)
• KUH Clarity is Oracle-based so we went Oracle and also because of site license and team expertise
• Upgraded storage to use FusionIO Duo 1.28 TB NAND cards
  • Solid state storage with direct PCI bus connection.
  • Impact has been dramatic on query performance and ability to support simultaneous queries.
• Continual Challenges though with Oracle Optimization as ad hoc queries and new functionality stretches system response
Constructing a Repository: Understanding Differing Data Models Used by Systems

Hierarchical databases (MUMPS), still very common in Clinical systems (VA VISTA, Epic, Meditech)


Star Schemas: Data Warehouses

Relational databases (Oracle, Access), dominant in business and clinical systems (Cerner, McKesson)
Moving to FusionIO storage memory platform (August 2011) improved performance, but after upgrading to i2b2 1.6, query times increased significantly.

**Average Query Time/Number of Queries by Month**

**Hardware/Software:** SUSE Linux Enterprise Server 11 (x86_64), Oracle 10g, 70G of RAM, 12 CPUs, Fusion-IO 1.28 TB IoDrive Duo storage tier
Software Development Process

• Team for HERON: Software Engineers (~1-3 FTE), Database Administrator (0.5FTE), Clinical Informatics Coordinator, Informatics Director (0.5->0.15 FTE)
• [http://informatics.kumc.edu](http://informatics.kumc.edu)
• Simple Wiki & Blog for documentation
• Checking in code into Mercurial version control
• View changes in TRAC
• Work with tickets, milestones and roadmap
  – New data added to monthly release milestones
  – Other milestones for functionality
  – Tickets for tasks, enhancements, end user problems & bugs
• **Share lessons and experiences with others**, what separates an academic institution from a business
• TRAC demo
• Goal: stable monthly process, minimal downtime
  • Complete rebuild of the repository, not HL7 messaging.
  • Two databases: create new DB while old DB is in use.
  • When the new DB is ready, switch over i2b2 to serve it.

• Initial Files from Clinical Organizations
  • Export KUH Epic Clarity relational database instead of Cache/MUMPS.
  • Monthly file from UKP clinic billing system (GE IDX).
    • Demographics, services, diagnoses, procedures, and Frontiers research participant flag.

• ELT processes largely SQL (some Oracle PL/SQL)
  • Wrapped in python scripts.
Engage contributing organizations — Russ and DBAs

Analysis — Director, Informaticists (CIC), Biomedical Informatics Software Engineers (BISE)

Set up databases - DBA

Extract Source Data to Staging Area — DBA

Transform into i2b2 Night HERON — BISE, DBA

De-id transform to Blue HERON — BISE, DBA

Regularize process — BISE lead develops, DBA operates

Publish release notes — CIC

   — HERON visits new lake or river

Whole process fairly “lights out”

   — Takes ~ 80 hours, switch over takes minutes
“Lazy” Load supports alternative views of reality

- Load with the local terminology first. Map concepts to standards secondarily in the concept space.
- Allows multiple ontologies for observations and works around mapping challenges with contributing organizations

Further technical details described at: http://informatics.kumc.edu/work/wiki/HERON
Constructing a Clinical Integrated Data Repository: Ethical and Regulatory Concerns

• Who “owns” the data? Doctor, Clinic/Hospital, Insurer, State, Researcher... perhaps the Patient?
  – Perception/reality is often the organization that paid for the system owns the data.
  – My opinion: we are custodians of the data, each role has rights and responsibilities

• Regulatory Sources:
  – Health Insurance Portability and Accountability Act (HIPAA)
  – Human Subjects Research

• Research depends on Trust which depends on Ethical Behavior and Competence

• Goals: Protect Patient Privacy (preserve Anonymity),
Will the released columns in combination with publicly available data re-identify individuals?

What if the released columns were combined with other items which “may be known”?

Sensitive columns, diagnoses or very unique individuals?

New measures to quantify re-identification risk.

---

• HIPAA Safe Harbor De-identification
  – Remove 18 identifiers and date shifting by 365 days back
  – Resulting in non-human subjects research data but treated as a limited data set from a system access perspective. System users and data recipients agree to treat as a limited data set (acknowledging re-identification risk)

• To be addressed:
  – For now, we won’t add free text such as progress notes with text scrubbers (DeID, MITRE Identification Scrubber toolkit)
  – While de-identified, access to timeline functionality provides individualized patient “signatures”
  – Obtained Cross organizational consensus to expand functionality beyond obfuscated counts
    • No sets < 10 and sets randomly perturbed ± 3 patients
Data re-identification risk and sensitivity for different data access possibilities with HERON/i2b2

- We’ve committed to offer this manually with a DUA and IRB approval; Currently provide contact info with HICTR participant committee approval
- Current Practice when giving researchers Epic access
- We’ve committed to offer this with a DUA
- Would be nice to offer this with a level of review between a SAA and DUA
- We are currently here and require SAA, faculty or sponsorship, and HSC training

Author: Russ Waitman, KUMC
Last modified: July 22, 2011
<table>
<thead>
<tr>
<th>HERON Cimarron (data through June 2012)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer Cases [9,325,899 facts; 63,779 patients]</td>
</tr>
<tr>
<td>Demographics [18,190,532 facts; 1,915,719 patients]</td>
</tr>
<tr>
<td>- Age [1,891,822 facts; 1,891,822 patients]</td>
</tr>
<tr>
<td>- Ethnicity [1,915,719 facts; 1,915,719 patients]</td>
</tr>
<tr>
<td>- Frontiers Research Participant Registry [11,637 facts; 11,637 patients]</td>
</tr>
<tr>
<td>- Gender [1,915,719 facts; 1,915,719 patients]</td>
</tr>
<tr>
<td>- Language [1,915,719 facts; 1,915,719 patients]</td>
</tr>
<tr>
<td>- Marital Status [1,915,719 facts; 1,915,719 patients]</td>
</tr>
<tr>
<td>- Place: School District [213,881 facts; 213,881 patients]</td>
</tr>
<tr>
<td>- Place: State [1,273,103 facts; 1,273,103 patients]</td>
</tr>
<tr>
<td>- Place: distance from KUMC [1,205,390 facts; 1,205,390 patients]</td>
</tr>
<tr>
<td>- Race [1,915,719 facts; 1,915,719 patients]</td>
</tr>
<tr>
<td>- Religion [1,915,719 facts; 1,915,719 patients]</td>
</tr>
<tr>
<td>- Vital Status [2,100,385 facts; 1,915,719 patients]</td>
</tr>
<tr>
<td>- Deceased [21,691 facts; 21,691 patients]</td>
</tr>
<tr>
<td>- Deceased per SSA [184,666 facts; 184,666 patients]</td>
</tr>
<tr>
<td>- Deferred</td>
</tr>
<tr>
<td>- Living [203,073 facts; 203,073 patients]</td>
</tr>
<tr>
<td>- Not recorded [1,690,955 facts; 1,690,955 patients]</td>
</tr>
<tr>
<td>- Diagnoses [20,461,541 facts; 617,972 patients]</td>
</tr>
<tr>
<td>- Flowsheets [490,945,281 facts]</td>
</tr>
<tr>
<td>- Laboratory Tests [75,783,302 facts; 267,529 patients]</td>
</tr>
<tr>
<td>- Medications [100,395,527 facts; 262,717 patients]</td>
</tr>
<tr>
<td>- Medications by VA Class/Clinical Dose Form (DRAFT) [89,758,747 facts; 249,754 patients]</td>
</tr>
<tr>
<td>- Procedures [10,099,551 facts; 554,882 patients]</td>
</tr>
<tr>
<td>- REDCap [15,922 facts; 123 patients]</td>
</tr>
<tr>
<td>- Specimens [31,261 facts; 3,056 patients]</td>
</tr>
<tr>
<td>- UHC DRAFT 2008 Q4 to 2012 Q1 [3,898,600 facts; 56,930 patients]</td>
</tr>
<tr>
<td>- UHC Agency for Healthcare Research and Quality [381,163 facts; 56,861 patients]</td>
</tr>
<tr>
<td>- UHC Core Measures [308,324 facts; 12,134 patients]</td>
</tr>
<tr>
<td>- UHC Demographics [184,764 facts; 56,930 patients]</td>
</tr>
<tr>
<td>- UHC Diagnosis [1,952,136 facts; 56,930 patients]</td>
</tr>
<tr>
<td>- UHC Procedures [232,001 facts; 49,374 patients]</td>
</tr>
<tr>
<td>- UHC Visit Details [840,212 facts; 56,930 patients]</td>
</tr>
<tr>
<td>- Visit Details [2,242,318 facts]</td>
</tr>
<tr>
<td>- Clinical Services [2,242,318 facts]</td>
</tr>
</tbody>
</table>

**HERON: Current Contents**

- >800 million facts
- 1.9 million patients but...
  - Most are just old administrative registrations
Richness of Phenotype is the Goal.
Example: Frontiers Participant Registry

*All Frontiers Participants have Diagnosis and Procedure Data.
Engagement and Review

- Dedicated Coordinator. Informatics Clinics held biweekly and one-to-one trainings and consultations offered
- Integrating HERON’s use into other research workflows
  - Finding patients for prospective trials: combining the Frontiers Participant Registry with the EMR data to find willing participants that meet study criteria.
  - Searching for samples: Biospecimen Repository combined with EMR to find tissues that meet research criteria.
- Auditing small queries
Supporting National Cancer Institute Cancer Center Designation

Incorporate Clinical, Administrative, Research Datasources

- Inpatient and outpatient electronic medical records (Epic)
- Professional Services Billing and Scheduling (GE IDX)
- KUCC Biospecimen Shared Resource Samples Database
- Hospital (KUH) Tumor Registry (NAACR format)
- Social Security Death Master File (NIST format)
- Technical Charges from hospital and clinics (UHC validated format)
- Research Data Capture (REDCap)
- Clinical Research Information System (Velos)

HERON's current contents with Cancer Center centric data in green

- Demographics (master patient index)
- Race/Ethnicity
- Laboratory Results
- Nursing observations/vital signs
- Clinical Diagnoses (ICD9)
- Medications (dispensed, ordered, home meds, administered)
- Physician Orders
- Procedure charges (CPT)
- Outpatient Billing diagnoses (ICD9)
- Inpatient visit/provider service

- Specimen collected
- Tumor Staging and Grade
- Diagnosis and Treatment
- Survival and Progression
- Site Specific Factors (e.g. ER positive)
- Death per Social Security Administration
- MSDRG, APDRG, LOS, Readmissions
- Technical Charge Diagnoses ICD9
- Service line, AHRQ quality and JCAHO core measures

- Triple Negative Breast Cancer Registry initial pilot completed

Status as of September 19, 2012

http://informatics.kumc.edu/work/wiki/HeronProjectTimeline#March2012Planning
- contains current plan for next several monthly releases
Idealized HERON Cancer Center Research Workflow

1. See what we have
2. Define a cohort
3. Conduct Analysis
4. Resulting Plot