Using the HERON Data Repository for Research

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
  – Tools for storing information: REDCap
  – Tool for viewing/getting information: HERON/i2b2
• Where does HERON fit in with Frontiers: our Clinical Translational Science Award?
• Re-identification Risk
• Demo
• 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 July 10.
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

Photo Credit: S. Klathill
http://www.flickr.com/photos/sklathill/505464990/
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 do 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

**University of Kansas Medical Center**

## KU-ADC

### NACC - A1 Subject Demographics

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADC Subject ID</td>
<td>1</td>
</tr>
<tr>
<td>Enrolled in NACC MDS</td>
<td>Yes</td>
</tr>
<tr>
<td>Primary Reason for coming to ADC</td>
<td>Other</td>
</tr>
<tr>
<td>Principal Referral source</td>
<td>Clinician</td>
</tr>
<tr>
<td>Presumed disease Status at enrollment</td>
<td>Control/Normal</td>
</tr>
<tr>
<td>Presumed Participation</td>
<td>Initial Evaluation only</td>
</tr>
<tr>
<td>ADC enrollment type</td>
<td></td>
</tr>
<tr>
<td>Subject's month of birth</td>
<td></td>
</tr>
<tr>
<td>Subject's year of birth</td>
<td></td>
</tr>
<tr>
<td>Subject's sex</td>
<td>Male</td>
</tr>
<tr>
<td>Does the subject report being Hispanic/Latino ethnicity (i.e., having origins from a mainly Spanish-speaking Latin American country), regardless of race?</td>
<td>Yes</td>
</tr>
<tr>
<td>What does subject report as his/her primary race</td>
<td></td>
</tr>
<tr>
<td>What additional race does subject report?</td>
<td></td>
</tr>
<tr>
<td>What additional race beyond what is indicated above in questions, does subject report?</td>
<td></td>
</tr>
<tr>
<td>Subject's primary language</td>
<td></td>
</tr>
<tr>
<td>Informant's years of education (report achieved level using the codes below; if an attempted level is not completed, enter the number of years attended). High school/GED = 12; Bachelors degree = 16; Master's degree = 18; Doctorate = 20 years)</td>
<td></td>
</tr>
</tbody>
</table>
**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 out 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></th>
</tr>
</thead>
<tbody>
<tr>
<td>* must provide value</td>
<td></td>
</tr>
<tr>
<td>Last Name:</td>
<td></td>
</tr>
<tr>
<td>* must provide value</td>
<td></td>
</tr>
<tr>
<td>email</td>
<td></td>
</tr>
<tr>
<td>* must provide value</td>
<td></td>
</tr>
</tbody>
</table>

Where are you?
* must provide value
- UMKC
- Children's Mercy
- KCUMB
- KU-MC
- KUMC-W
- Ku-L
- Lukes
- VA
- Other

Role
* must provide value
- 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]
Option Two: **CRIS/Velos**

**REDCap** Disclaimer

- For clinical trials, CRIS/Velos may be a better fit
  - Multiple years of experience
  - CRIS team builds for you with biostatistics review
  - Budget for CRIS team and biostatistics explicitly

- “Investigator driven” REDCap only works if you, the Principal Investigator, takes responsibility for your data
  - Scalability: informatics provides consultation and responsibility for technical integrity; *not your dictionary or data entry*.
    - Underwritten by CTSA, but you “feed and talk to your fish”
  - Middle model where informatics can build for you in REDCap.
    - Again, you budget for our team’s time
REDCap: think Fish Tank you manage

http://www.flickr.com/photos/wiccked/185270913/lightbox/
CRIS: Vancouver Aquarium. We feed fancy fish; Biostatistics cooks tasty meal
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://frontiersresearch.org/frontiers/HERON-Introduction](http://frontiersresearch.org/frontiers/HERON-Introduction) for more information, status, and training videos

Single sign-on using your email username
Real-time check for current human subjects training
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
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
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
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 2: Constructing a 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.

HERON Integrated Data 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)

De-identification & Transform Processes

De-identified staging database (Blue HERON)

De-identified data server

Identified data server

I2b2 clinical business intelligence application (JBoss, VMWare virtualized host managed by KUMC Information Resources)

Clinical/Translational Researcher
• HIPAA Safe Harbor De-identification
  – Remove 18 identifiers and **date shifting by 365 days back**
    • **Downside:** can’t do seasonal studies without IRB approval to go back and get actual dates
    • In general, tack on 7 months when wanting volume for the last year.
  – 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”
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
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.
This month, our tour of rivers and lakes in Kansas honors Melvern Lake.

The HERON repository contains approximately 856 million real observations from the hospital, clinics, and research systems:

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Observation</th>
<th>Patients</th>
<th>Source</th>
<th>Go-Live</th>
<th>Snapshot</th>
<th>Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>18.4M</td>
<td>1.94M</td>
<td>KUH Billing (O2 via SMS)</td>
<td>1980s</td>
<td>Oct 2012</td>
<td>various*</td>
</tr>
<tr>
<td>Diagnoses (ICD9)</td>
<td>186.2K</td>
<td>186.2K</td>
<td>Frontiers participant registry</td>
<td>Jun 2009</td>
<td>Oct 2012</td>
<td></td>
</tr>
<tr>
<td></td>
<td>34.2M</td>
<td>644K</td>
<td>Social Security Death Index</td>
<td>1962</td>
<td>Oct 2012</td>
<td></td>
</tr>
<tr>
<td>Medications</td>
<td>75.9M</td>
<td>289K</td>
<td>Organized by VA Class</td>
<td>Nov 2007</td>
<td>Oct 2012</td>
<td>various*</td>
</tr>
<tr>
<td>Lab Results</td>
<td>80.3M</td>
<td>282K</td>
<td>KUH/O2/Epic</td>
<td>2003</td>
<td>Oct 2012</td>
<td>various*</td>
</tr>
<tr>
<td>Procedure Orders</td>
<td>53.9M</td>
<td>431K</td>
<td>KUH/O2/Epic</td>
<td>?</td>
<td>Oct 2012</td>
<td>various*</td>
</tr>
<tr>
<td>Reports/Notes</td>
<td>24.7M</td>
<td>218K</td>
<td>KUH/O2/Epic</td>
<td>?</td>
<td>Oct 2012</td>
<td>various*</td>
</tr>
<tr>
<td>Cancer Cases</td>
<td>9.5M</td>
<td>65.7K</td>
<td>KUH Cancer Registry</td>
<td>1950s</td>
<td>Oct 2012</td>
<td>#1544</td>
</tr>
<tr>
<td>Hospital Quality Metrics</td>
<td>4.12M</td>
<td>60.9K</td>
<td>University Health System Consortium (UHC)</td>
<td>Q4 2008</td>
<td>June 2012</td>
<td></td>
</tr>
<tr>
<td>Triple Negative Breast Cancer Registry (BRCA)</td>
<td>17.8K</td>
<td>133</td>
<td>REDCap</td>
<td>July 2011</td>
<td>Sept 2012</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>816M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

https://informatics.kumc.edu/work/blog
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

<table>
<thead>
<tr>
<th>Inpatient and outpatient electronic medical records (Epic)</th>
<th>Ambulatory visit details, Ordersets &amp; Decision Support used <em>(pending)</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional Services Billing and Scheduling (GE IDX)</td>
<td>Provider/visits <em>(pending)</em></td>
</tr>
<tr>
<td>KUCC Biospecimen Shared Resource Samples Database</td>
<td></td>
</tr>
<tr>
<td>Hospital (KUH) Tumor Registry (NAACR format)</td>
<td></td>
</tr>
<tr>
<td>Social Security Death Master File (NIST format)</td>
<td></td>
</tr>
<tr>
<td>Technical Charges from hospital and clinics (UHC validated format)</td>
<td>Insurance Status, clinics, ED <em>(pending)</em></td>
</tr>
<tr>
<td>Research Data Capture (REDCap)</td>
<td>Patient Registry Annotation</td>
</tr>
<tr>
<td>Clinical Research Information System (Velos)</td>
<td>Clinical Trials Participation <em>(pending)</em></td>
</tr>
</tbody>
</table>

### 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
  - MDSRG, 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](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

Cancer Survival: Obese (BMI > 30) Diabetic Breast