Integrating REDCap Patient Registries within an i2b2 Integrated Data Repository

Summary
Cohort discovery and hypothesis generation for clinical trials eligibility and data extraction for patient registries are among the natural use cases for integrated data repositories which include electronic medical records (EMR).

- Bringing data from clinical trials and registries into the repository allows investigators to exploit business intelligence, visualization, and integration methods commonly included with data warehousing software but often not available within clinical trial management and electronic data capture systems.
- We describe an automated integration method that encapsulates both the patient data and the design of data collection instruments in a user-friendly ontology.
- The method was piloted by integrating a cancer registry maintained in the REDCap electronic data capture system (EDC) with an i2b2-based data repository.
- The resulting ontology supported rapid review of registry data quality and led to suggestions on how to improve design of the data collection instruments in the EDC to facilitate easier browsing and querying in the data repository.

REDCap at KUMC
At KUMC, REDCap was established in November 2010 and has now grown to include over 500 projects.

- This rapid growth can be attributed to:
  - an easy to use interface
  - KUMC’s low barriers to use: Any user with a KUMC affiliated enterprise account can log onto REDCap and build basic projects.

When building complex projects, users can:
- access extensive training materials available online
- attend biweekly clinics or request one-on-one training
- engage KUMC Bio-statistics to build and maintain their REDCap projects

Extract Transform & Load
- Tableau from REDCap MySQL database related to data, metadata, event metadata and logs are extracted. These data are staged in Oracle schemas on the identified server.

Transformation & Loading
- Metadata for drop-down fields is split using regular expressions and paired for ease of transformation:
  - Only fields that are either drop-downs or text fields with proper data validation are imported while file upload fields and text fields without any validation are ignored in accordance with the current take based on other data sources.
  - Any fields that were marked as containing protected health information (PHI) are also not loaded to facilitate de-identification.

- The subjects in the REDCap data are linked to rest of their data in HERON via their Medical Record Numbers (MRNs).
- Finally the data is transformed into the i2b2 schema, de-identified and loaded.

Observations
Looking at some of the objectives that motivated us to do this integration:

- Richer Data Picture & Validating Data
  - Over 50% of the registry patients had data in both Biospecimen Repository and Tumor Registry.
  - Two of the registry subjects were found to be reported as deceased by the Social Security Death Index.
  - Over 50% of the subjects also had data regarding Laboratory Tests, Medications, Procedures, Visit Details, Reports & Procedure Orders

- Optimizing CRF design & usability
  - CRFs were marked as ‘same as ‘age at diagnosis’ were modeled as a drop down field instead of using the text box with number validation on it.

- Eliminating redundant data collection
  - 1) Most of the data captured in Baseline Demographics CRF, such as gender, age and ethnicity are already being recorded in the EMR.
  - 2) Fields such as Stage in the CRF, Surgery Outcome can be obtained from the tumor registry.
  - 3) Some of the data in Physical Exam CRF such as Date of Blood Draw can be abstracted from the Biospecimen Repository, which is also part of HERON.

- Limitations
  - Currently this approach does not accommodate data from SQL fields or un-validated text fields.
  - It also does not consider the ordering of the fields in CRF’s when building the i2b2 ontology.

Triple Negative Breast Cancer Registry
- Triplenegative breast cancer (TNBC) is a rare and aggressive subtype of breast cancer. TNBC tumors are characterized by their negativity to the three receptors commonly associated with most other subtypes of cancer:
  - Progesterone receptor
  - Human epidermal growth factor receptor 2 (HER2)
  - Estrogen receptor.

- This registry collects cancer-related data and specimens from patients with triple negative breast cancer, over a period of 10 years.
- It has 807 Fields in 19 CRFs.

Motivation for Integration
- REDCap is a web-based electronic data capture system developed by Vanderbilt.

- Data in REDCap is organized into projects. A project can be tailored to represent anything from an entire clinical trial to a help desk system.

- The projects are made up of Data Collection Instruments i.e. Case Report Forms (CRFs).

- Data collection instruments are comprised of fields or data elements.

- Fields in REDCap broadly fall into three categories:
  1) File upload fields
  2) Multiple choice fields
  3) Text fields

- Text Fields in REDCap can be augmented by using data validations to enforce data types like date, number, text, phone number and zip code.

- Multiple choice fields can be true/false, yes/no, drop down or checkboxes.

- Fields can also be marked as containing PHI.

- Using events capability, data collection instruments can be designated to be completed at certain time points during the protocol.

Future Work
- Access control to REDCap data in HERON to limit only users who have been granted access to REDCap data via REDCap user interface

- Integrate other REDCap projects and data from other EDC systems, especially in responsive trials and frontline enrollment into HERON.

- Clinical trial data integration into data repository can give a better picture of enrollment targets and also to monitor and rectify under-enrollment problems that breast most clinical trials.

Conclusion
- With small patient populations and very diverse designs, clinical trials and registries can pose problems during integration. Hence, a plan for such integration should be part of the design of the data collection instruments for success. Users might have to be gently steered towards a more standard build earlier in the design process. This can be done either through training or one-on-one meetings with investigators before project design.

- The de-identification strategy for data also presents some challenges. The repository teams have to work with researchers so that fields that contain PHI can be properly identified before extraction and loading on a project by project basis.

Acknowledgements & References