

SEINE: Methods for Electronic Data Capture and Integrated Data Repository Synthesis with Patient Registry Use Cases

Bhargav Adagarla¹, Daniel W. Connolly¹, Tamara M. McMahon¹, Manikandan Nair¹, Lisa D. VanHoose³, Priyanka Sharma², Linda J. D'Silva³, Lemuel R. Waitman¹

¹Division of Medical Informatics, ²Department of Internal Medicine, ³Physical Therapy and Rehabilitation Science, University of Kansas Medical Center, Kansas City, KS

Abstract

Integrated Data Repositories (IDR) allow clinical research to leverage electronic health records (EHR) and other data sources while Electronic Data Capture (EDC) applications often support manually maintained patient registries. Using i2b2 and REDCap, (IDR and EDC platforms respectively) we have developed methods that integrate IDR and EDC strengths supporting: 1) data delivery from the IDR as ready-to-use registries to exploit the annotation and data collection capabilities unique to EDC applications; 2) integrating EDC managed registries into data repositories allows investigators to use hypothesis generation and cohort discovery methods. This round-trip integration can lower lag between cohort discovery and establishing a registry. Investigators can also periodically augment their registry cohort as the IDR is enriched with additional data elements, data sources, and patients. We describe our open-source automated methods and provide three example registry uses cases for these methods: triple negative breast cancer, vertiginous syndrome, cancer distress.

Introduction

Integrated data repositories allow clinical researchers to see patient care processes across disparate data sources. The University of Kansas Medical Center's Healthcare Enterprise Repository for Ontological Narration¹ (HERON), based on informatics for integrating biology and the bedside² (i2b2), was established in 2010 to simultaneously advance clinical and translational science initiatives and strengthen the university's pursuit of National Cancer Institute designation (i.e. award of a NCI Cancer Center Support Grant). The data incorporated into the repository now includes patient demographics, laboratory results, diagnoses, procedures, vital statistics from the Social Security Death Index, the hospital tumor registry, multidisciplinary flowsheets, medications, the university bio-specimen repository, and University Healthsystem Consortium Clinical³ Database. Currently, the repository has over 1.25 billion facts for over 1.9 million patients. The i2b2 software has been adopted by a majority⁴ of NIH Clinical and Translational Science Award recipients, other academic medical centers nationally and internationally, and also pharmaceutical companies.

Electronic data capture applications optimized for clinical data are widely used for data collection in research and operational settings. Web-based EDC systems see wide adoption due to their easy-to-use interfaces, data security, portability and rapid deployment. REDCap⁵ (Research Electronic Data Capture) is one of the most popular, web-based applications designed to support data capture for research studies. Currently in use by over 798 institutions in over 61 countries across the world⁶, it provides an intuitive user interface, basic statistical tools, and comprehensive auditing. Use of REDCap at KUMC has grown steadily since its installation in November 2010, to include over 1200 users and 1000 projects. The projects include clinical trials, registries and quality of care databases, among others.

Motivation

Clinical registries can be powerful tools to gain insights into the course of a disease⁷. They can be used by clinicians to understand the course of the disease and effectiveness of care; by clinical organizations to evaluate the care delivered by the clinicians; by clinicians to compare their treatment strategies with those of their peers⁸. But chart review and other traditional techniques for building and maintaining them are labor intensive and prone to errors and redundancy in data capture. A goal of integrated data repositories is to bridge these gaps between registries and systems such as EHRs.

Successful cohort discovery from an integrated data repository results in a researcher getting access to a relevant data set. Delivery by file transfer is traditional but raises disclosure risks. Using an Entity Attribute Value (EAV)⁹ data model in an IDR accommodates a wide variety of data sources and types, but traditional statistical analysis techniques¹⁰ are designed to use standard row-modeling structures.

To address the above issues we developed methods that 1) integrate data from an EDC system into an IDR to enrich registries by linkage with EHRs and other data sources to reduce chart review and abstraction burden and 2) deliver data from the IDR securely in EDC system containing a summary of the data in ordinary row-modeling CRFs that can be used as the basis of a registry. Integrating EDC data into IDRs also helps researchers leverage the data visualization, analysis, and manipulation tools that are common in data repositories. Likewise, delivering IDR data requests in EDC systems enables researchers to exploit the data collection, export and user management capabilities provided by EDC systems.

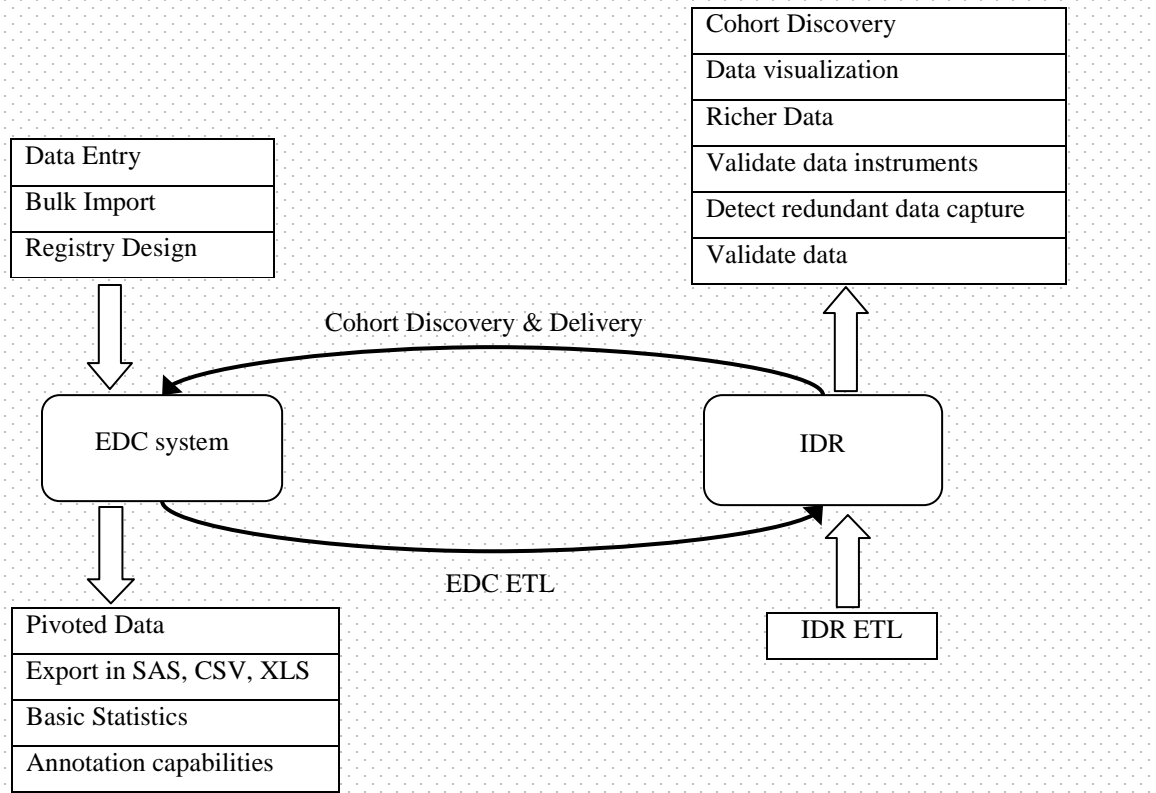


Figure 1. Integration between Electronic Data Capture and Integrated Data Repository systems (SEINE)

Methods

Inspired by seine fishing nets, SEINE (Synthesizing EDC IDR Network Exchange), was developed using REDCap and i2b2. REDCap's data model is encapsulated by a metadata table that defines all the study fields with accompanying data types and represents the patient data either as a flat table with one row per subject (familiar to end users) or as an Entity Attribute Value (EAV) model (commonly used in electronic medical records). In contrast, i2b2 uses a star schema common in data warehouses¹¹, as its underlying data model with a large observation fact table complemented by patient, visit, and concept dimensions. The concept dimension is analogous to the REDCap metadata table but also provides hierarchical paths or organizing terminologies. Both REDCap and i2b2 have adjunct tables to store user profiles and auditing.

Loading REDCap registries into i2b2: Data in HERON is refreshed monthly using an extract/transform/load (ETL) process. Data from various sources (e.g. EPIC® Electronic Health Record Clarity database, the hospital tumor registry using North American Association of Central Cancer Registries (NAACCR) format) are staged in Oracle database schemas on a server designated for identified data¹². Then, an automated task control program written in python that wraps Structured Querying Language (SQL) scripts specific to each data source, transforms the staged data into the i2b2 star schema. Next, the transformed data is de-identified and transferred into an identical i2b2 star schema on the server accessed by the i2b2 application. To add REDCap as a data source in this process, we researched existing solutions and discovered the ODM (Open Model for Exchanging Power System Simulation Data) to i2b2 importer tool¹³ built by Informatics, Children's Hospital Boston. It is designed to allow for importing

of study data from EDC, but our evaluation concluded that it would not be feasible to integrate into our automated ETL process.

Our design begins with staging tables from REDCap's MySQL database related to data, metadata, event metadata and logs on the identified server. Aside from "multiple choice" categorical fields, the discrete data in REDCap can be directly transformed into discrete i2b2 observations as both systems use an EAV data model. But in the REDCap metadata table, options for categorical fields are combined into one cell and to facilitate transformation, SEINE splits such fields into multiple rows. To de-identify data from REDCap, our ETL scripts omit any fields marked in the REDCap data dictionary as identifiers. Further, only fields that are either dropdowns or text fields with proper data validation are loaded. Currently we make no attempt to process free-text fields in order to avoid the risk that such fields may contain patient identifiers. The data from the remaining fields is subject to the general HERON de-identification design involving date-shifting etc.

As a final step, term trees are built into the concept paths to emulate REDCap structures. The projects are listed as folders under a top-level REDCap folder in i2b2. The projects in turn comprise of folders representing each of the data collection instruments in the project. The fields are each listed under their parent case report form. Dropdown categorical fields are again folders with each discrete choice being a leaf, whereas all numeric or validated text fields are listed as leaves. The subjects in the REDCap data are linked to rest of their data in HERON via the health systems enterprise master patient index medical record number (MRN). The SEINE REDCap transformation SQL and python code is available under an open source license as part of the HERON codebase¹⁴.

Access rights to REDCap data in HERON: HERON encapsulates i2b2 within a regulatory web interface that manages single sign-on authentication, verifies human subjects training requirements, and manages system access agreements and authorization. We have integrated the REDCap project access control model into this interface. After login to HERON, the system queries REDCap to determine user rights and customizes the REDCap projects visible within i2b2. As a result, investigators can share their registries in HERON with other investigators by simply granting them rights using web forms in REDCap and access is granted immediately upon login; there is no dependency on the monthly ETL process.

Extracting data from i2b2 into REDCap: HERON uses a self-service model for data exploration and data extraction. Previously, investigators would create an i2b2 query to define their patient cohort and place all the variables (i2b2 concepts or terms) they wished to extract into a final "terms" panel in the user interface. After approval by the HERON Data Request Oversight Committee, the honest broker would then provide the data to the investigator in several comma separated value (CSV) files created by a SQL script that took the i2b2 query name as an argument.

SEINE's methods now give the honest broker automated tools to provide the data in the form of a REDCap project. The first tool, the R Data Builder, extends i2b2 via the analysis tool plug-in interface. The input is a patient set (resulting from an i2b2 query) and a list of variables (i2b2 concepts). The R Data Builder queries the data repository for these variables and writes the results to an R¹⁵ data file. The R Data Builder¹⁶ was previously developed to allow integration with the R statistical platform.

Building the CRFs to hold i2b2 observation facts was a particular challenge. In the i2b2 EAV model, entities are patients (or encounters), attributes are observations (diagnoses, lab results, medications, etc.) and values include date of observation, numeric lab result, and so on. Investigators are familiar with row modeling where each row represents a patient or case, each column represents a variable or observation, and cells represent values. In such a structure, answering "what is the average age of males in this cohort who have taken aspirin?" is supported by typical spreadsheet features: filter the sex column to males, the aspirin column to yes, and compute the average of the age column. If each attribute were included at most once for each entity, pivoting the data from EAV to row-modeling would be straightforward. But aspirin observations of multiple sorts (order, administration, dose, etc.) may be recorded at multiple times. Our approach is to provide a summary of the data: for each entity and attribute, we provide a count and the first and last values. Given an approved data use request, the honest broker:

1. Creates a REDCap project for the data set.
2. Re-runs the investigators query to copy the investigator's cohort into a patient set.
3. Uses the HERON R Data Builder analysis tool with the concepts contained in the investigator's query to extract the data into an R data file.
4. Schedules an R script that pivots patient data to generate the data import file and data dictionary for the REDCap project. The REDCap data dictionary defines one CRF for demographics from the i2b2 patient

dimension (age, sex, race, etc.) and other CRFs that group observations by the top-level i2b2 folders (e.g. Diagnoses, Medications, Lab Results). The project can be organized by either encounter or patient.

5. After uploading the data dictionary, schedules a process to upload the data.
6. Assigns user rights for the REDCap project based upon the approved data use request.

Results

The September 2013 HERON refresh includes approximately 266K facts about 4510 patients from six REDCap registries related to breast cancer, distress measurements in cancer patients and sepsis management, among others. While our methods continue to evolve, we highlight three cases exercising different portions of SEINE.

Case 1: Our initial integration of REDCap registries into HERON was piloted with a breast cancer registry¹⁷. The Triple Negative Breast Cancer (TNBC) Registry collects over 10 years of clinical data, tumor characteristics, and follow up information for patients with a rare¹⁸ and aggressive subtype of breast cancer. This registry is modeled as a longitudinal REDCap project with over 600 fields. Figure 2 shows one of the CRFs and resulting i2b2 hierarchy.

Event Name: Registration	
Baseline Demographics CRF	123456
Baseline Tumor Characteristics CRF	Upload document
Date of Biopsy	2013-10-01
Location of Biopsy	KU
Surgical Pathology Number	0000000
Pathological Type	Ductal
TNM Staging: T	Medullary
TNM Staging: N	Ductal
TNM Staging: M	Lobular
Stage	I
Triple Negative?	Yes
ER%	22
PR%	24
HER2: IHC	2+
HER2: FISH	22
EGFR%	32
KI67%	46
P53%	34
Others	
Form Status	

- REDCap [265,590 facts; 4,510 patients]
 - Distress2 [3,675 facts; 3,675 patients]
 - Triple Negative BRCA Registry- TNBC only [37,336 facts]
 - 66-120 month follow ups [467 facts; 15 patients]
 - BRCA Status [1,684 facts; 186 patients]
 - Baseline Demographics [1,302 facts; 186 patients]
 - Baseline Medical History [11,307 facts; 186 patients]
 - Baseline Tumor Characteristics [1,849 facts; 186 patients]
 - Complete? [186 facts; 186 patients]
 - Date of Biopsy [184 facts; 184 patients]
 - HER2: IHC [182 facts; 182 patients]
 - Location of Biopsy [186 facts; 186 patients]
 - KU [56 facts; 56 patients]
 - Other [130 facts; 130 patients]
 - Pathological Type [181 facts; 181 patients]
 - Ductal [176 facts; 176 patients]
 - Lobular [<10 facts]
 - Medullary [<10 facts]
 - Other [<10 facts]
 - Stage [186 facts; 186 patients]
 - TNM Staging: M [186 facts; 186 patients]
 - TNM Staging: N [186 facts; 186 patients]
 - TNM Staging: T [186 facts; 186 patients]
 - Triple Negative? [186 facts; 186 patients]
 - No
 - Yes [186 facts; 186 patients]

Figure 2. Triple Negative Breast Cancer Registry: REDCap data collection instrument with mock data is shown on the left with corresponding i2b2 hierarchy on the right.

The integration was helpful in validating the design of the registry data collection instruments in REDCap. For example: ‘age at diagnosis’ was modeled as a drop down field with one option for each number instead of being a text box with number validation on it. This led to a browsing experience in HERON where ranges of ages could not be selected easily. Also, some of the CRFs were replicated for follow-up visits instead of using REDCap’s longitudinal database feature that can be translated into encounters in i2b2. The clinical team that constructed the registry then rectified these issues. The ability to query across the TNBC registry and the Biospecimen Repository and Tumor Registry is of particular interest for research. Queries again revealed that 61% of the registry patients had a specimen in the Biospecimen Repository while 56% had data in Tumor Registry. Queries also show that 2% had been reported as deceased by the Social Security Death Index and the EMR.

Case 2: Our earlier approach of delivering the i2b2 data in a long EAV file was sometimes intimidating or time consuming for investigators without extensive data analysis backgrounds. SEINE methods for data extraction were applied for a data use request for patients with vestibular dysfunction (i.e. vertigo). The investigator used i2b2 to

define cohorts with and without type 2 diabetes and extracted information regarding specific vertiginous diagnoses along with diabetes related clinical variables. Providing the data within REDCap saved considerable time and much of the preliminary analysis was accomplished by REDCap's built in "Graphical Data View & Stats" module as well as export into files compatible with standard statistical programs.

Case 3: Finally, an investigator exercised all portions of SEINE in the developing and supplementing a Cancer Distress Registry for investigating the relationship between problem list and distress scores on distress thermometer¹⁹ in cancer patients. Investigators had developed a registry of 3675 patients from records manually maintained in their clinic. This initial cohort was uploaded into a REDCap project which was then incorporated into HERON as a part of the monthly ETL process. During the HERON ETL process, the patients were linked to rest of their EHR data and hospital tumor registry data via their medical record numbers. The investigator used i2b2 to refine queries that identified an additional 1042 patient; expanding their registry. Finally, the researcher placed a request for identified data for cumulative 4717 patients (146,284 facts) from the EHR, tumor registry, and vital statistics.

HERON both provided additional demographic variables that were not initially available in the manually maintained database and allowed the researcher to conduct interrater reliability assessments of variables across multiple sources (EHR, Registry, manually maintained registry). Investigators were able to clearly identify gaps in data collection processes. The investigator plans to make additional data requests as more patients are accumulated in clinic maintained registry and as patient volumes increase within the HERON integrated data repository.

Discussion

Our choice of platforms and technologies is intended to facilitate research by the clinical research community: i2b2 is open source and installed at dozens of sites, and REDCap is used by thousands of researchers and carries no license fees. Our HERON source code (available at <https://informatics.kumc.edu>) is open source, written in SQL and python, which are ubiquitous programming technologies **as well as the R statistical package**. As the codebase for SEINE directly binds with the underlying data structures of i2b2 and REDCap, which are rarely modified, we expect the maintenance burden to be minimal. For example, accommodating newer field validation types in REDCap involves just a few changes to the relevant ETL code.

When a REDCap registry is integrated into HERON, the resulting alternative view of the data collection instruments, an i2b2 folder hierarchy with facts and patient counts at each level, provides feedback on the registry design that can lead to early detection and correction of design flaws such as undiscovered patient identifiers, redundancy, and inappropriate use of data types. To maximize the value of data collected in registries, a plan for integration with larger data contexts should be part of data collection instrument design. Without careful planning, our experience shows that with small patient populations and very diverse designs, data integration problems naturally arise.

Feedback on SEINE has been sufficiently positive that we plan to complete features left out of the initial releases. We plan to add a convention to identify encounter dates in the case of longitudinal REDCap projects. We also plan to enhance SEINE to fulfill identified data requests by researchers where currently it only delivers de-identified data. We have also learned that the de-identification strategy for incorporating EDC data within the IDR has scalability concerns as it requires the informatics team to review and coordinate with researchers to ensure that fields that contain PHI are properly identified before extraction and loading on a project by project basis.

Conclusion

SEINE methods quickly incorporate user-designed registries and clinical trial case report forms into an enterprise information architecture, though their diverse designs can pose problems during integration and highlight the benefits of advanced planning and standardized instrument design. These tradeoffs are more than offset by reducing the data entry burden, complimentary data visualization and manipulation tools, and responsive investigator-centered access control. While REDCap and i2b2 are each excellent tools in their own right, the cases presented above demonstrate that they are considerably more powerful when brought together with SEINE. The resulting platform provides full life-cycle data management for a variety of research needs.

References

1. Waitman LR, Warren JJ, Manos EL, Connolly DW. Expressing Observations from Electronic Medical Record Flowsheets in an i2b2 based Clinical Data Repository to Support Research and Quality Improvement. AMIA Annual Symposium Proceedings. 2011;2011:1454-63.
2. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association. 2010;17(2):124-30.
3. UHC Clinical Data Base/Resource Manager. [10/10/2013]; Available from: <https://www.uhc.edu/11536.htm>.
4. i2b2: Informatics for Integrating Biology & the Bedside. [10/10/2013]; Available from: https://www.i2b2.org/work/i2b2_installations.html.
5. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. Journal of biomedical informatics. 2009;42(2):377-81.
6. REDCap Consortium. [10/10/2013]; Available from: <http://project-redcap.org>.
7. Gliklich R, Dreyer N, editors. Registries for evaluating patient outcomes: a user's guide. 2nd ed. Registries for evaluating patient outcomes: a user's guide 2nd ed2009. p. 377-81.
8. Kennedy L, Craig A-M. Global registries for measuring pharmaco-economic and quality-of-life outcomes: focus on design and data collection, analysis and interpretation. Pharmacoeconomics. 2004;22(9):551-68.
9. Stead WW, Hammond WE, Straube MJ. A Chartless Record--Is It Adequate? Proceedings of the Annual Symposium on Computer Application in Medical Care. 1982:89-94.
10. Dinu V, Nadkarni P, Brandt C. Pivoting approaches for bulk extraction of Entity-Attribute-Value data. Computer methods and programs in biomedicine. 2006;82(1):38-43.
11. Kimball R. The data warehousing toolkit. Indianapolis, IN: Wiley Pub. ; 1997.
12. HERON – University of Kansas Medical Center's Healthcare Enterprise Repository for Ontological Narration. [10/10/2013]; Available from: <https://informatics.kumc.edu/work/wiki/HERON>.
13. ODM to i2b2 importer. Available from: <https://community.i2b2.org/wiki/display/ODM2i2b2/Home>.
14. HERON Source Code. [10/10/2013]; Available from: https://informatics.kumc.edu/work/browser/heron_load.
15. Team RDC. R: A Language and Environment for Statistical Computing. Vienna, Austria2008.
16. Connolly D. R Data Builder. 2013 [10/10/2013]; Available from: <https://informatics.kumc.edu/work/wiki/HeronStatsPlugins>.
17. Adagarla B, Connolly D, Nair M, Sharma P, Waitman LR. Integrating REDCap Patient Registries within an i2b2 Integrated Data Repository. 2012 AMIA Summit on Clinical Research Informatics. San Francisco, CA, USA2012.
18. Foulkes WD, Smith IE, Reis-Filho JS. Triple-negative breast cancer. The New England journal of medicine. 2010;363(20):1938-48.
19. Lynch J, Goodhart F, Saunders Y, O'Connor SJ. Screening for psychological distress in patients with lung cancer: results of a clinical audit evaluating the use of the patient Distress Thermometer. Supportive Care in Cancer. 2010;19(2):193-202.

Acknowledgements

This work was supported by a CTSA grant from NCATS awarded to the University of Kansas Medical Center for Frontiers: The Heartland Institute for Clinical and Translational Research # UL1TR000001 (formerly #UL1RR033179). The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or NCATS.