Reproducibility of Electronic Health Record Research Data Requests

Ramkiran Gouripeddi, MBBS, MS[1,2], Rishi Deka, PhD[3], Thomas Reese, PharmD[1], Ryan Butcher, MS, MBA[2], Bradley Martin, PharmD, PhD[4], Jeff Talbert, PhD[5,6], Bernie LaSalle, BS[1,2], Julio C. Facelli, PhD[1,2], Diana Brixner, RPh, PhD[3]

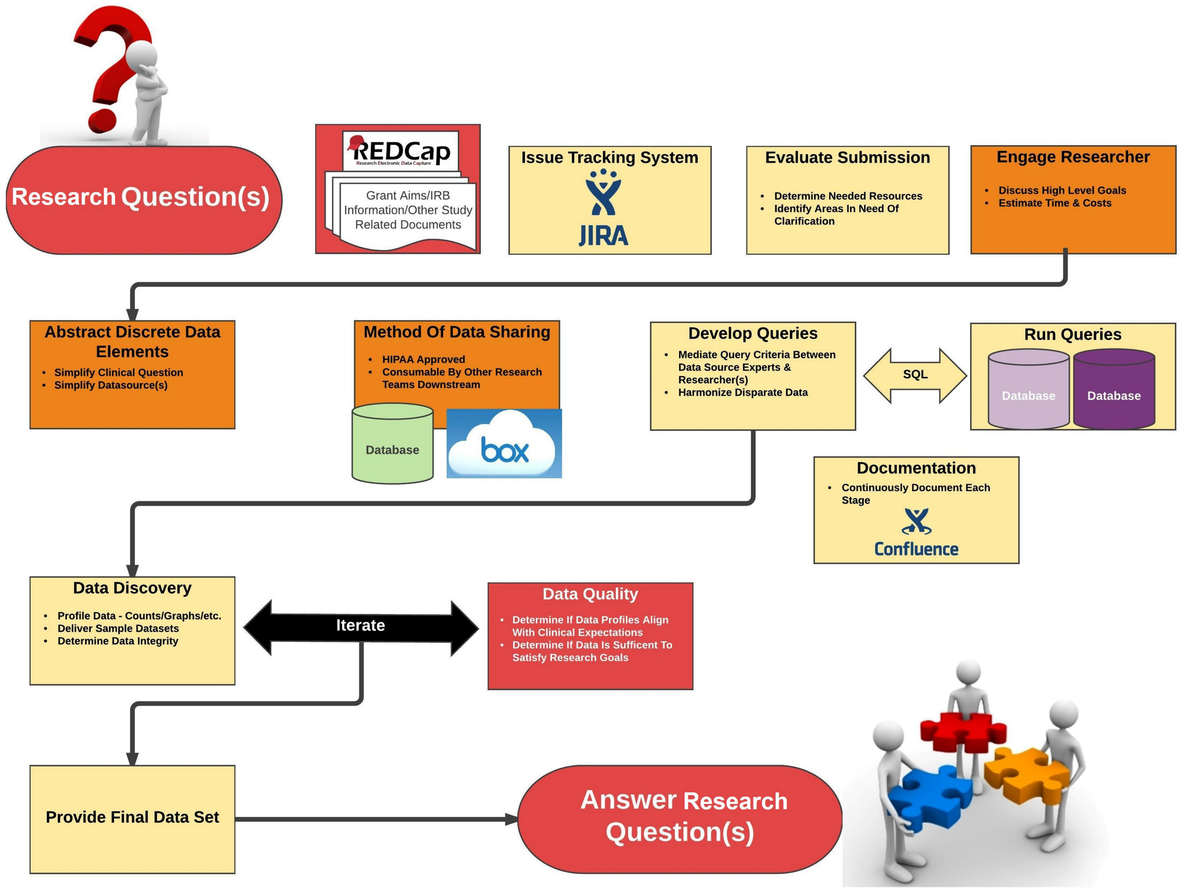
[1]Department of Biomedical Informatics, [2]Center for Clinical and Translational Science, [3]Department of Pharmacotherapy Outcomes Research and Health Policy, University of Utah, Salt Lake City, Utah, USA, [4]College of Pharmacy, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA, [5]Institute for Pharmaceutical Outcomes and Policy, [6]Center for Clinical and Translational Science, University of Kentucky, Lexington, Kentucky, USA

Introduction

- Translational research is dependent on secondary use of existing data including electronic health record (EHR), as a resource for knowledge discovery.
- Researchers provide natural language descriptions of their data needs, which research data teams use to develop queries and results from EHR systems..
- Data teams and researchers usually engage in complex written and verbal communication processes including:
  - Mediate details within natural language descriptions.
  - Abstract understood requests to the complexities present within EHR systems.
  - Development of appropriate query scripts, extraction, transform and provisioning of query results and analysis.
  - Iterate over these steps multiple times in order to develop final data deliverables.
- Goal of this study: Analyze reproducibility of current processes using natural language descriptions for acquiring research data from EHR.

Methods

- Provided a natural language description of an Upper Respiratory Tract Infection (URTI) study to data teams of three CTSA sites.
- Teams were blinded to the true nature of the study, which was to understand the reproducibility of the natural language description.
- Asked teams to describe their data query process to ascertain possible reasons for discrepancies.



A Typical Research Data Request Process Consisting of Multiple Steps with Complex Communication[1]

Patients eligible for enrollment between July 1, 2012 and September 30, 2015. Data will be reviewed up to six months pre-index to identify baseline characteristics and exclusion criteria. Subjects identified with an ICD-9-CM diagnosis code for an URTI during an outpatient patient encounter will be included. The index date is defined as the first ICD-9-CM documentation for an URTI. A 6-month pre-index period will be used to identify exclusion criteria, and to observe baseline characteristics. Patients will be examined for outcomes of interest within 24-hours of the index clinic date and time. In addition, we exclude patients with a positive rapid antigen detection test (RADT) for group A streptococcal pathogens at the initial visit (results available within 24-hours) as this is an instance when antibiotic prescribing is appropriate.

Inclusion criteria:

1. *Age ≥18 years old*
2. *Diagnosis (ICD – 9 code list provided) of a URTI in the outpatient setting*

Exclusion criteria:

1. *AIDS/HIV (ICD – 9 code list provided)*
2. *COPD/Asthma (ICD – 9 code list provided)*
3. *Cancer (ICD – 9 code list provided)*
4. *Conditions for which antibiotic prescribing may be appropriate*
   a. *An URTI diagnosis within the 180 day pre-index period*
   b. *Additional infectious diseases (ICD – 9 code list provided)*
   c. *A positive rapid antigen detection test for group A streptococcal pathogen (LOINC code list provided).*

Upper Respiratory Tract Infection Study Data Request Description

Step	Criteria	Site A	Site B	Site C
1.	Patient with outpatient visits between July 1, 2012 and September 30, 2015	684,478	460,159	412,942
2.	Patients of Step 1 with URTI	18.7%	1.7%	3.1%
3.	Patients of Step 2 > 18 years of age	17.5%	34.2%	39.9%
4.	Patients of Step 3 remaining after applying exclusion criteria	55.8%	62.1%	23.4%
5.	Patients of Step 4 receiving an antibiotic prescription within 24 hours of index visit	9.4%	0.3%	36.5
6.	Patient of Step 4 receiving an antibiotic prescription within 8 days of index visit	11.2%	0.3%	40.3%

Variations at Each Step of the Data Request across the Three Sites

Discussion

- We found that contextual, organizational and data analyst specific issues played a significant role in how the data team constructed their queries.
- Differences in EHR data transformation, storage and presentation to data teams played an important role.
- Inability to successfully reproduce data requests across different sites should be an important consideration.
- Reproducing research data requires effective communication between the research and data teams.
- Need for:
  - Structured or semi-structured data requisition methods using templates.
  - Context-sensitive and metadata-driven workflows supporting entire life cycle of research data requisition including the development of the natural language description of the research data, query mediation, data abstraction, data extraction and provisioning of results and analysis.
  - Use FAIR[2] (Findable, Accessible, Interoperable, Reusable) principles within translational pipelines (e.g. Research Data Alliance – Health Data Interest Group[3]).

References

1. R. Butcher, R. K. Gouripeddi, R. Madsen, P. Mo, and B. LaSalle, CCTS Biomedical Informatics Core Research Data Service, Research Reproducibility Conference, Salt Lake City, 2016.
2. M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data, Mar. 2016.
3. Health Data IG, RDA, 21-Dec-2015. Available: https://www.rd-alliance.org/groups/health-data.html.

Acknowledgements: This work is supported by the Utah Center for Clinical and Translational Science funded by NCATS award UL1TR001067. Computer resources were provided by the University of Utah Center for High Performance Computing. **Contact Information:** Ram Gouripeddi (ram.gouripeddi@utah.edu)

HEALTH UNIVERSITY OF UTAH

Utah CCTS