Scientific Reproducibility using the Provenance for Healthcare and Clinical Research Framework

Satya S. Sahoo
Collaborators/Co-Authors: Joshua Valdez, Michael Rueschman, Matthew Kim, Susan Redline

Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University
Department of Medicine, Brigham and Women’s Hospital and Beth Israel Deaconess Medical Center, Harvard University

BD2K All Hands Meeting 2016
Scientific Rigor and Reproducibility

- Reproducibility is a cornerstone of scientific advancement

NIH guidelines for “Rigor and Reproducibility”

1. Scientific Premise
2. Scientific Rigor
3. Biological Variables
4. Authentication
Role of Provenance: Translating Guidelines to Practice

• The **Reproducibility Cycle**: available and missing components

  - **Public Datasets** (available)
  - **Contextual Metadata** (missing)

  **World Wide Web Consortium (W3C) PROV Specifications**: Standard for provenance metadata modeling

---

**Provenance Metadata** describes the history of information resources.
Provenance for Clinical and Healthcare Research (ProvCaRe)

• What provenance metadata elements are needed for reproducibility?

• ProvCaRe Framework (S3 Model)
  1. Study Method: Design and method used in the experiment
  2. Study Tools: Instruments and tools used in experiment
  3. Study Data: Data elements, cutoff threshold, minima/maxima
Provenance for Clinical and Healthcare Research (ProvCaRe)

- Objectives of the ProvCaRe framework:
  - Develop a provenance model for scientific reproducibility: ProvCaRe ontology
  - Automated extraction of provenance from published scientific studies: ProvCaRe text processing pipeline
  - Provenance query and exploration interface for users: ProvCaRe research repository

- Technologies used in ProvCaRe
  - PROV specifications for provenance modeling
  - Web Ontology Language (OWL2)
  - Resource Description Framework (RDF) for provenance graphs
  - RDF data store
  - Clinical Natural Language Processing
ProvCaRe Ontology

- New post-coordination syntax for provenance expressions
- Re-uses terms from existing biomedical ontologies

ProvCaRe ontology:
- Extends W3C PROV Ontology
- Represents S3 Model
ProvCaRe-NLP Pipeline

- Extraction of provenance metadata from published scientific studies
  - Selected 50 articles that generated or used datasets available from the National Sleep Research Resource (NSRR)
  - NSRR is largest sleep medicine data repository: 50,000 sleep studies involving 36,000 participants
ProvCaRe Research Repository

- ProvCaRe repository for querying provenance of scientific studies
- Scenario 1:
  - Find prospective cohort studies of sleep disordered breathing and hypertension
- Scenario 2:
  - Develop rigorous study design using provenance of published studies

- User interface for ProvCaRe research repository
- Provenance Ranking scientific studies based on ProvCaRe framework compliance

---

**Search ProvCaRe Repository:**

*Compose Query*

hypertension

**Results:**

<table>
<thead>
<tr>
<th>Study Reference</th>
<th>Study Methods</th>
<th>Study Data</th>
<th>Study Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Katz et al., 2014</td>
<td>nonrandomized uncontrolled studies, limiting inferences, baseline, follow-up</td>
<td>physical activity, behavioral outcomes, sleep activity, intervention</td>
<td>eAT intervention, RAI, covariate regression analysis</td>
</tr>
<tr>
<td>O’Connor et al., 2009</td>
<td>baseline, follow-up, cohort study</td>
<td>person, association, study population, correlation data</td>
<td>multivariate regression method, average, median</td>
</tr>
</tbody>
</table>

- Provenance extracted from studies categorized into the S3 model

- Simple Query Composition Functionality
Next Steps: ProvCaRe and BD2K Centers

- CEDAR: Develop and publish ProvCaRe metadata templates for reproducible studies
- Validate published studies using CEDAR API services

- bioCADDIE Common Data Elements for dataset identification and query
- bioCADDIE search API integrated into ProvCaRe research repository
Acknowledgements

Big Data to Knowledge (BD2K) Data Provenance grant (1U01EB020955)

http://provcare.case.edu satya.sahoo@case.edu

References: