Meeting the Challenge
Semantically-based Provenance Metadata Tools
A Strong Team
Building on Existing Collaborations

Team (to date)
Co-PIs
• Marcy Harris – School of Nursing
• Carl Lagoze – School of Information

Team Members (to date)
• Ivo Dinov
  – MIDAS
• Jamie Estill
  – Medical School Information Services: Research
• Chuck Friedman
  – School of Medicine, Department of Learning Health Sciences
• Amitava Shee
  – Medical School Information Services: Research

Collaboration & Expertise
• National scale data research network using local health data
• Modeling, data curation, metadata, variety of relevant research workflows
• Implementation of semantic web technologies (e.g., RDF, OWL)
• Expertise in provenance metadata

Discussions with additional partners underway
Contemporary Research Environments Require New Types of Infrastructure

Information increase 10x every 5 years

Health care
• An information intensive industry
• Data characterized by all of the big data Vs

Expanding expectations
• Use the data and information to improve the human condition
• Cross the research - clinical care barrier
• Adopt open science principles
• Assure transparency, reproducibility

• Become a Learning Health System

More than 90% of the analysis pipeline has been devoted to time consuming data preparation steps such as normalization, calibration, outlier treatment, and provenance tracking. The practice of data science will benefit from new tools that can automate or semi-automate this process at large scales.
Provenance Metadata

• The core of transparency and scientific integrity
• Are these data trustworthy?
  – From where did data originate?
    • Trace the lineage based on the source data and transformations
    • Establish the copyright and ownership of data, enable its citation
  – How were the data generated and processed?
    • Trace the audit trail, determine resource use, detect errors
  – Are findings reproducible?
    • Repetition of data derivation, help maintain its currency, and be a recipe for replication.
  – How to cite data? What other data were used to calibrate, validate, and process these data?

• Traditional mechanisms break down when leveraging data from widely distributed and heterogeneous sources
Research Agencies are Paying Attention

Especially for repositories of health care data with genotype and phenotype data

- NIH genomic data sharing policy
- PCORI clinical data research networks
Aims

Link provenance metadata from origin to research data packages

• Design, develop and prototype models to express the provenance of transactional data across spectrum from origination in distributed, heterogeneous systems through workflows within research projects

• Develop tools and protocols to transmit and aggregate provenance metadata from distributed, heterogeneous systems into provenance-enhanced research data packages

• Develop tools that allow researchers to view and use provenance-enhanced research data packages that permit researchers to fluidly cross the boundaries between sensitive, confidential data, non-anonymized research snapshots, and safe harbor derivatives
What we Plan to Do (1)

- Leverage models of provenance based on Semantic Web technologies (RDF and OWL)
- PROV model
  - Based on a set of recommendations by W3C
  - Evidence of early uptake in healthcare standards communities
  - PROV does not yet include full representation of provenance for longitudinal metadata or life-cycle perspective
- DDI metadata format
  - Developed in social science community
  - Expressly longitudinal
  - RDF expression of DDI-Lifecycle
Virtual Transactional Tape
synchronize & concatenate data envelopes

1. system level transactional tape w/prov info
2. system-directed pull protocol
   (1/2) notion of 'conformant system'
3. extractor/aggregator
4. canonicalized transaction tape
5. snapshot prov aggregator
6. research data package in various formats w/common prov-enhanced DOI metadata
7. Anonymize/cloak for public use

ResourceSync
A NISO standard
What We Plan to Do (2)
Michigan PCORnet Node as Testing Environment

PCORnet: A national infrastructure for patient-centered clinical research

PCORnet
- A network of networks linking 75-100 million records
- Questions go to the data
- All have some LHS notion
- Enabled by a common data model and technologies
PCORnet Common Data Model v3.0

Fundamental basis

Data captured from processes associated with healthcare delivery, registry activity, or directly from patients

Data captured from healthcare delivery, direct encounter basis

New to v3.0

Associations with PCORnet clinical trials

Process-related data

Bold font indicates fields that cannot be null due to primary key definitions or record-level constraints.

http://www.pcor.net/resource-center/pcornet-common-data-model/
Under Construction
Implement Node in ARC environment
Under Consideration

Figure-5: Big-data based Analytics Platform
Related Initiatives – We have Colleagues!

- ONC HIT standards committee data provenance taskforce
- BD2K
- Michigan Data Science Initiatives: DSI + MIDAS
Summary

• Addressing an important challenge

• Have expertise, data, environment

• Looking for partners!

Questions? Comments?