Big Data and Health Record Privacy Proposals

May 25, 2016

Cavan Capps

Big Data Lead

Research and Methodology Directorate U.S. Census

Historic Opportunities

- Longitudinal Studies (health outcomes)
- Drug interactions
- Genetic interactions (increased granularity)
- Connecting the Dots:
  - Linking data like ER room visits and death certificates improves quality tracking and health practice innovation

Growing Privacy Challenges

- Commercial incentives to link data for marketing purposes
- Health Insurance has increasing incentives to identify “cherries to pick” *(if health outcomes are highly predictive, can a insurance based health payment system survive?)*
- Identity Theft
- **Public Trust is threatened**.

Three Proposed Access Models

w/Common Privacy Architecture

- Sandbox Enclave
- Synthetic Data created using “Formal Privacy” techniques
- Secure Multi-party Computing

- Employing a Common Privacy Architecture

Common Privacy Architecture Proposal

- Dataset level enforced Provenance
- Separation of PHI using Pseudo-IDs
- Unalterable Logging of all operations

- Results in a mechanism for independent “Privacy Auditing” much like independent “Financial Auditing”.

Common Architecture: Provenance

- Dataset level, not just patient level
- Machine/Human Readable (XML like HL7)
- Includes all Human agreed usage agreement
- Includes “Creation Characteristics” and “Chain of Custody”
- Includes all data transformations & usages
- Creates a new valid provenance for all extracts
- Is digitally finger printed to the data

Common Architecture: PHI replaced by Pseudo-ID

- PHI stored separately and encrypted (reduces Identity Theft Risk)
- Linking record attributes to PHI is done by random “Pseudo-Ids”, any re-linkage can be logged.
- Pseudo IDs can be separate by Organization (not a universal Pseudo ID)
- Any Linking is done using Pseudo ID mapping
- All linkage using Pseudo IDs are logged
- All linkage of PHI to attribute data are logged.
- Creates a new valid provenance for all extracts
- Is digitally finger printed to the data

Common Architecture: Unalterable data manipulation logging

- All data interactions are logged.
- Log is set up cryptographically so that it supports only appending, and breaks if edited.
    - Enforcement using proper procedures can access the data, but will be forcibly logged.
- Machine Learning techniques can be run on logs to discover re-identification related activities.
- The log plus the provenance document can be used to conduct “independent privacy audits”.

Model 1: Semi-Trusted Analytical Sandbox Enclosure

- All researchers must be semi-trusted (background check and face penalties)
- Individual records (consisting of pseudo-Ids and attributes) are restricted from being matched to data that includes PHI. *(this is fundamental)*
- Research the ability to do analysis only using distributions of records (of limited attributes) rather than individual records should be explored.
- Public outputs should be “Formally Privatized”.

Model 2: Publically Released Synthetic Differentially Private Dataset

- When created, the creator must pre-determine which data elements have the least amount of noise added to them. There is a limit to the number of questions that you can ask a particular “formally privatized” dataset.
- We need more research on how well a synthetic dataset can provide the same analytical answers that the confidential data can provide.

Model 3: Secure Multi-party Computing

Distributed computing over encrypted data

- Can use the common privacy architecture
- Individual records are not visible
- Pre-limit the type of analysis permitted
- Currently up to 100 times slower
- Currently being explored by the DARPA “Brandeis Project”
- Currently being explored by Census to get near-real-time business data for Economic Statistics