8-16-2012

Protecting Human Subjects and Their Data in Multi-Site Research

Clete Kushida
Stanford University School of Medicine

Harold S. Luft
Palo Alto Medical Foundation Research Institute

Keith Marsolo
Cincinnati Children's Hospital Medical Center

Follow this and additional works at: http://repository.academyhealth.org/webinars

Part of the Bioethics and Medical Ethics Commons, Health Services Research Commons, and the Law Commons

Recommended Citation
http://repository.academyhealth.org/webinars/4

This Video/Media is brought to you for free and open access by the Events at EDM Forum Community. It has been accepted for inclusion in Webinars by an authorized administrator of EDM Forum Community.
Protecting Human Subjects and Their Data in Multi-Site Research

Clete Kushida, M.D., Ph.D.; Keith Marsolo, Ph.D., Harold Luft, Ph.D.

August 16, 2012
Welcome!

→ Erin Holve, Ph.D., M.P.H., M.P.P.
   – Director of Research & Education, AcademyHealth
   – Principal Investigator of the EDM Forum
Improving Health and Health Care
As the professional society for health services researchers and health policy analysts, our mission is three-fold:

1. Support the development of health services research
2. Facilitate the use of the best available research and information
3. Assist health policy and practice leaders in addressing major health challenges

We work to both “push” the production of research and promote the “pull” by decision makers
Electronic Data Methods (EDM) Forum

→ Special Supplement with Medical Care
  – 14 commissioned & invited papers representing 41 investigators and 23 institutions
Connecting the Audio

Audio for this presentation must be accessed through the teleconference via your phone. Two options to connect:

1. Click on the phone symbol in the toolbar at the top of your screen. Enter your phone number and click “join,” the system will call you directly.

2. Call in directly:
   - Dial (866) 244-8528
   - Enter the access code 436690 and press pound (#)
Technical Assistance

➔ Live technical assistance:
  – Call Adobe Connect at (800) 422-3623

➔ Refer to the ‘Technical Assistance’ box in the bottom left corner for tips to resolve common technical difficulties.
Questions may be submitted at any time during the presentation.

To submit a question:
1. Click in the Q&A box on the left side of your screen.
2. Type your question into the dialog box and click the Send button.
Learning Objectives

→ To be able to identify practical approaches that can be used to facilitate IRB approval for multi-center research studies;
→ To gain an understanding of how these approaches have been used in existing research networks through case studies of 3 real-world examples;
→ To identify the need for optimal de-identification and anonymization strategies;
→ To describe whether current de-identification and anonymization strategies are effective; and
→ To discuss further work needed in this area.
Faculty

Keith Marsolo
Ph.D.
Cincinnati Children’s Hospital Medical Center

Clete Kushida
M.D., Ph.D.
Stanford Sleep Medicine Center

Harold Luft
Ph.D.
Palo Alto Medical Foundation Research Institute
Approaches to facilitate IRB approval of multicenter research studies

Keith Marsolo, PhD
Assistant Professor
Division of Biomedical Informatics
Cincinnati Children’s Hospital Medical Center
August 16, 2012
Motivation

• Multi-center studies critical for many types of research
  – Comparative effectiveness, rare disease, pediatrics, etc.

• For many/most NIH grants, IRB approval of human subjects research is required before receiving award

• As research becomes more complex and distributed, methods to speed IRB approval become increasingly important

• Much of existing literature focused on clinical trials or issues with IRB regarding implementation of HIPAA
  – Very little on distributed QI or research networks
  – Case studies serve as useful precedence
Drawbacks to multiple IRB reviews

• Consent form changes can increase complexity and decrease readability
  – Increases chance that errors will be introduced
  – Lowers study response rate

• Same study reviewed at different levels

• Duplication of effort
  – Every center in study must have IRB approval
  – However – centers can use the same IRB
Approaches to speed approval

• Working with IRBs
• Central IRB
• Federated IRB
• Umbrella protocol

• Spoiler: No “silver bullets”
Central IRBs (CIRBs)

- Not tied to a single institution
- Can focus on specific type of research or geographic area

- Examples
  - National Cancer Institute
  - Children’s Oncology Group
  - American Academy of Family Physicians (AAFP)
  - Biomedical Research Alliance of New York
  - Western IRB
Central IRBs

• Benefits
  – Faster review & approval
  – Lower administrative staff burden

• Drawbacks
  – Concerns from local IRBs about research integrity, review quality, local context
  – Hesitation about use without demonstration of efficiency
Federated IRBs

• Attempt to incorporate benefits of CIRBs while avoiding pitfalls

• Emerged from CTSA (Clinical & Translational Science Awards) Child Health Oversight Committee
  – First deployed during pilot phase of National Children’s Study

• 3 tiers of participation
  – Tier 1: Reliance on lead IRB (if all Tier 1 – lead IRB ~ CIRB)
  – Tier 2: Local IRB review of materials reviewed & approved by lead IRB
  – Tier 3: Reliance on local IRB (if all Tier 3 – traditional study)

• Mechanism for local IRBs to amend core consent material to comply with institutional/state requirements
Federated IRBs

• Benefits
  – Sites can choose level of participation
  – Build trust between institutions
  – Local IRBs can focus on local context
  – Lower administrative burden on local sites

• Drawbacks
  – Greater administrative burden & liability on lead site
  – Approach is new – little evidence
Case Study: DARTNet

• **Background**
  – Federated network of care centers established in 2008
  – Facilitate QI in primary care & compile data for CER

• **Architecture**
  – Sites assemble a database of patient-level data
  – Data is standardized, de-identified & linked through a web portal
    • De-identification performed by site or Clinical Information Networks of America (CINA)
  – Queries are broadcast to all federated databases
Case Study: DARTNet

• Umbrella protocol covers creation/use of databases
  – Site-level QI
  – Practice-level aggregate reports

• Each research study requires new protocol
  – Centers choose to participate
  – Approved by local IRB or AAFP IRB

• Studies also require completion of data use agreement
  – Language kept standard unless changes required by law
  – Minimize back-and-forth between lawyers
Case Study: SPAN

• Background
  – Federated, distributed query network
  – Conduct CER across several HMO Research Network (HMORN) sites & community partners

• Factors facilitating approval
  – Meeting with local IRBs before submission
  – Familiarity with HMORN model
  – Frequent meetings of HMORN IRB administrators & collective development of policies
  – IRB protocol covering development & governance of database
Case Study: ImproveCareNow

• Background
  – 36 centers focused on improving care of children with Inflammatory Bowel Disease (IBD)
  – Started in 2007 as QI network
    • Variety of IRB decisions – waive, obtain consent, not research
  – Shift to QI & research after awarded grant from AHRQ
    • Move to federated IRB
    • New patients: consent; Existing patients: maybe/maybe not

• Architecture
  – Before: populate central web-based registry through double-data entry
  – After: Mix of EHR-based data collection, double-data entry, central & distributed
Case Study: ImproveCareNow

- Initial findings (good)
  - Frequent check-ins with centers can help speed IRB process
  - Asking lawyers not to change legal agreements does work
  - More centers relying on lead IRB than expected

- Initial findings (not so good)
  - Discussion about reliance has required a one-on-one phone call (so far)
  - Ask too many questions and the IRB will become unresponsive
Conclusions

• Building trust with IRB is key
• Engage IRBs early

• Big unknown: changes to Common Rule
  – Proposed: all domestic sites in multi-center study use single IRB
  – Implies use of Central IRB, tier 1 Federated IRB
  – Techniques still applicable
Questions?
Strategies for De-Identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies

Clete A. Kushida, M.D., Ph.D.
Professor, Stanford University Medical Center
De-Identification and Anonymization Strategies
Why Are Such Strategies Important?
Database Sources, Key Words, Search Strategy
Case Examples (Text, Images, Biological Samples)
Discussion
  - Are Current De-ID Strategies Effective?
  - Which Strategies Are Best?
  - How Essential is Anonymization?
  - Do De-ID Strategies Alone Meet the Needs of Multicenter Research Studies?
  - What Approaches Can Be Used on a Multicenter Level to Ensure Privacy?
  - Further Work
Definitions

• De-identification and anonymization are strategies that are used to remove patient identifiers in electronic health record (EHR) data.

• *De-identification* of EHR data is the removal or replacement of personal identifiers so that it would be difficult to reestablish a link between the individual and his or her data.

• *Anonymization* refers to the irreversible removal of the link between the individual and his or her medical record data to the degree that it would be virtually impossible to reestablish the link.
Why Are Such Strategies Important?

• HIPAA Privacy Rule regulations (2000) permits covered entities to use/disclose data that have been removed of patient identifiers without obtaining an authorization and without further restrictions on use/disclosure. There are 18 “safe harbor” data identifiers under the Rule that constitute the minimal set of removed identifiers.

• Use of data removed of patient identifiers is one of three options available to investigators desiring to use medical data in research, besides obtaining informed consent from their patients or a waiver of informed consent from their IRB.
Why Are Such Strategies Important?

• As the use of EHRs has progressively increased, concerns have been raised about their utility to fundamentally improve the quality of patient care and the threat of unauthorized disclosure of PHI either unintentionally or by identity theft.

• Additionally, biomedical research is becoming increasingly dependent on the access, sharing, and management of EHR among clinical and research centers, especially those involved in observational and multicenter research studies.
Database Sources For Review

- **BIOSIS Previews** (via Thomson Reuters Institute for Scientific Information [ISI] Web of Knowledge, 1926-present)
- **CINAHL** (Cumulative Index to Nursing and Allied Health Literature, via EBSCOhost, 1937-present)
- **Inspec** (via Thomson Reuters ISI Web of Knowledge, 1898-present)
- **MEDLINE** (Medical Literature Analysis and Retrieval System Online, 1950-present)
- **SciVerse Scopus** (1823-present)
- **Web of Science** (via Thomson Reuters ISI Web of Knowledge, 1898-present)
Key Words and Search Strategy

- Key Words: deidentify, de-identify, deidentification, de-identification, anonymize, anonymization, data scrubbing, and text scrubbing
- Articles were included if they were published up to June 30, 2011 and there was no restriction on earliest date of publication (i.e., earliest date obtained in search was 1996).
- Through the combined database search, 1798 prospective citations were identified
The writing group chair conducted the review; however, five other members of the writing group independently reviewed the 120 full-text articles obtained after the abstracts review.

Differences between the reviewers’ judgments regarding inclusion or exclusion of articles were resolved by discussion; consensus was required from all six reviewers.

The full text of 120 articles were reviewed and resulted in a final sample of 45 articles that met inclusion criteria for review.
Flow Diagram of Search Results

- BIOSIS Previews N = 44
- CINAHL N = 48
- Inspec N = 420
- MEDLINE N = 171
- SciVerse Scopus N = 870
- Web of Science N = 245

Combined Total Citations N = 1798

Abstracts Reviewed N = 267

Exclusions Upon Abstracts Review N = 173
Primary reasons for exclusion:
- Not relevant to topic
- Outside medical records domain
- Non-English language article

Full-Text Articles Reviewed N = 94

Review of Additional Articles Extracted from References of Articles Obtained from Search N = 26

Articles Meeting Criteria N = 45

Exclusions Upon Citations Review N = 1532
Primary reasons for exclusion:
- Not relevant to topic
- Not relevant article type
- Duplicate citation

Exclusions Upon Articles Review N = 75
Primary reasons for exclusion:
- Not relevant to topic
- De-identification or anonymization strategy lacked sufficient detail to understand or interpret it
Case: De-Identification of Free Text

- Manual de-identification of PHI from free text in EHR can be tedious, costly, time-consuming, inaccurate, and unreliable.
- For example, resident clinicians can manually de-identify at a rate of about 18,000 words or 90 incidents of PHI per hour.

The automated software package, deid, scans the medical notes line-by-line, dividing them into individual words separated by whitespace.

deid identifies occurrences of PHI using dictionary-based look-ups and regular expressions.

deid replaces each PHI with a tag to indicate its corresponding category.

Case: De-Identification of Free Text

On a test corpus of 1,836 notes with 296,400 words, there was 90 instances of false negatives (missed PHI), or 27 per 100,000 word count, with a recall (sensitivity) of 94.3%.

Only one full date and one age over 89 were missed.

No patient names were missed.

Case: De-Identification of MR Images

- An automated “defacing” algorithm used models of non-brain structures to remove identifiable facial features from MR volumes of 342 T1-weighted datasets:
  - Did an effective job of removing facial features without sacrificing brain tissue (none removed)
  - Could be performed relatively quickly (approximately 25 min on a dataset of 342)
  - Did not interfere with subsequent data processing, and in some cases, improved the quality of subsequent automated skull-stripping by removing more non-brain tissue.

Case: De-Identification of MR Images

Case: De-Identification of Biosamples

Case: De-Identification of Biosamples

- This repository contains approximately 250,000 samples with an average influx of 90,000 samples per year of which approximately 80% need to be de-identified.
- This process differs from data scrubbing patient identifiers on the physical sample since a tube transfer procedure is used for sample de-identification.
- This is a manual de-identification procedure that is subject to human error.

Are Current De-ID Strategies Effective?

- Current de-identification strategies have impressive recall and precision rates.
- No existing system is perfect, and there is the possibility that certain PHI will not be de-identified.
- Limitations of many current systems include:
  - Inability to detect misspellings, typographical errors, and proper names that share characteristics with non-PHI
  - Restrictions in managing only certain types of data; algorithms that are not designed to handle diverse PHI (e.g., hard-coded PHI in output files)
  - Difficulty in compensating for variation in nomenclature
Which Strategies Are Best?

• For heuristic, lexical, and pattern-based systems, studies evaluating these systems have reported good performance (especially precision) but experienced domain experts must spend significant time and effort.

• For statistical learning-based systems, they are able to be used “out of the box” with minimal redevelopment time and learn how to identify PHI from the data itself rather than relying on precompiled, manually-constructed sets of data.
Which Strategies Are Best? (cont.)

• For both images and biological samples, there are too few studies with a paucity of quantitative data to judge the best approach.

• Biological samples have the added Common Rule anonymization requirements needed for IRB exemption that do not appear to be satisfactorily addressed by the current approaches.
How Essential is Anonymization?

• In theory, anonymization is important since it places the patient’s or research participant’s right to privacy as the top priority in any anticipated or unanticipated scenario, and dramatically minimizes the release of sensitive information that may discriminate or stigmatize the individual from a social or economic perspective.

• In practice, it still may be possible to identify an individual from supposedly anonymized data sets, especially with respect to rare diseases within a specific geographical area.
Do De-ID Strategies Alone Meet the Needs of Multicenter Research Studies?

• Besides the de-identification of individual documents, what can be done to ensure the privacy of data sets?

• What approaches can be used on a multicenter level to ensure patient or participant privacy?
What Approaches Can Be Used on a Multicenter Level to Ensure Privacy?

- De-identification and anonymization strategies are important, but are one component of an integrated data collection and management system.
- Some institutions use honest brokers, which collect and provide data to research investigators in a manner whereby it would not be reasonably possible for investigators to identify the participants directly or indirectly.
Further Work

• Management of identifiers for the protection of genetic information, particularly with respect to protecting the privacy of identities to which DNA sequences were derived.

• This area of genomic privacy is particularly challenging for the biomedical community, given the immense quantity of data that needs to be processed, stored, and shared, as well as the consequences that identifying genomic data may have on an individual’s health, employment, and insurance status.
Acknowledgements

- Deborah A. Nichols, M.S., Stanford University
- Rik Jadrnicek, Microflow DBMS Inc.
- Ric Miller, Microflow DBMS Inc.
- James K. Walsh, Ph.D., Sleep Medicine and Research Center, Chesterfield, MO
- Kara Griffin, Sleep Medicine and Research Center, Chesterfield, MO
- Pamela R. Hyde, M.A., Stanford University
- AcademyHealth
- Gurvaneet Randhawa, M.D., MPH and AHRQ

Funding was provided by a contract from AcademyHealth. Additional funding was provided by AHRQ 1R01HS1973 (Comparative Outcomes Management with Electronic Data Technology (COMET) Study).
Questions

To submit a question:
1. Click in the Q&A box on the left side of your screen
2. Type your question into the dialog box and click the Send button
Register for Upcoming Webinars: www.edm-forum.org

- Clinical Informatics, CER, and PCOR: Building Blocks for Meaningful Use of Big Data in Health Care
  - September 26, 2012; 1:00-2:30 pm Eastern Time
Thank You

*Please take a moment to fill out the brief evaluation which will appear in your browser.*