Protecting Research Data

Scott Bradner
Harvard University
BC – 10/17/2011

Agenda

- types of research data
- de-identified data
- data guidelines
- data use agreements
- FISMA
- Harvard as an example

Research Data

- sources of research data
  - researcher generated data
  - data from another source
- protection requirements
  - none
  - from data or money source
  - from overarching regulations
  - from institution

Research Data, contd.

- sensitivity
  - public
    - from public sources or properly de-identified
  - confidentially promised but not sensitive
  - embarrassing
  - economic impact or job risking
  - life threatening

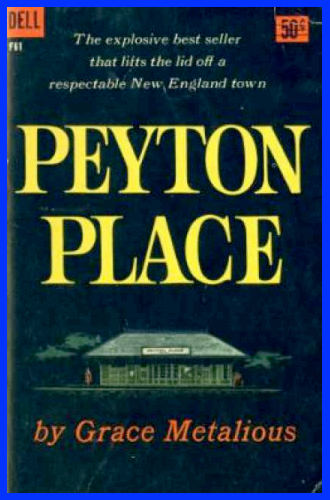
De-Identified Data

- identifiers that could link data to an individual have been removed

  e.g., DHS "safe harbor" (18 types of identifiers)
- changing understanding of possibility of re-identification

  DoB + zip code + sex = 87% match
- harder with context specific knowledge

  "Peyton Place problem"



De-Identified Data, contd.

- statistical method

  "apply generally accepted statistical and scientific principles and methods for rendering information not individually identifiable when determining if information is de-identified"

Data Guidelines

- use as little high-risk data as possible
  - e.g., need to have a very good reason to use SSNs
    - almost always name + last 4 digits of SSN fills need
- de-identify ASAP
  - note: mapping file must be protected at the level of the raw data
- but if you need to work on high-risk data then your protections must be up to the task

Data Use Agreements (DUAs)

- DUA is list of use, access & protection requirements imposed by data or funding source
  - different than the list of the protections you give to the data or funding source & say you will use
- core issue – signing authority
  - too often DUA requires a authorized institutional signature & researcher thinks they can sign
    - but generally do not have the authority to bind the institution – researcher could be personally liable

DUA Precedence

- a DUA defines the minimum set of protections & take precedence over institutional requirements
- but IRB can decide that the DUA requirements are not specific enough or are insufficient

  and require the researcher to adopt more stringent protections

  IRB has responsibility for ensuring protection of human subject data under US federal law

DUA quality

- wide variance in quality of DUA requirements
- from

  "protect the data"

  to

  specific checklists of protection techniques

- often specific requirements are not understood by DUA writer or recipient

  e.g., FISMA

FISMA

- Federal Information Security Management Act
- some push in federal agencies to include FISMA security requirements in grants & contracts
  - FISMA also shows up in DUAs from non-federal government entities
    - e.g., local school systems
- FISMA requirements in proposed DoD requirements to protect non-classified data

FISMA, contd.

- NIST 800-53 - revised July 2009
- 237 page document
- more than 200 individual requirements
  - high level generally requires automated mechanisms to meet requirements
  - moderate level sometimes requires automated mechanisms to meet requirements
  - low level generally does not require automated mechanisms

FISMA, contd.

- 3-level requirements
  - low - could be met by well run university data centers with some effort
  - moderate - possible to be met by well run university data centers with a lot of effort & expense
  - high - unlikely to ever be met by a university data center
- Amazon has received FISMA moderate certification for their cloud services
  - but user still responsible for a lot of requirements

Other Regulations

- many apply even if not mentioned in a DUA
- most states also have data protection requirements

  e.g., Mass 201 CMR 17
- Health Insurance Portability and Accountability Act (HIPAA)

  applies to medical records – some confusion for medical records not at a care giver

Failure Can Hurt

- Mass Gen agreed to pay a $1M penalty for misplacing medical records concerning 192 people
- UNC researcher demoted & pay cut after breach

  http://chronicle.com/article/Chapel-Hill-Researcher-Fights/124821

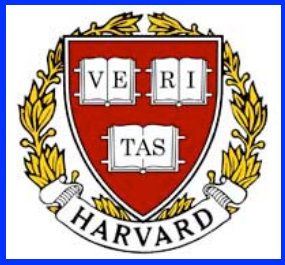
- industry estimate for cost of SSN breaches – about $214/record

  http://www.ponemon.org/blog/post/cost-of-a-data-breach-climbs-higher

Local Rules

- institution can develop their own rules
  - to be used for data generated or collected by institutional funded or unfunded researchers
  - to be used when IRB determines DUA requirements are too weak
  - to be offered to data sources or funders as list of ways to protect data
- note federal rules require “immediate” institution access to research data in case of an investigation

Harvard as Example



- Harvard developed a Harvard Research Data Security Policy (HRDSP) over about 1.5 years
- process driven by chair of Social Science Committee, Provost and Vice Provost for Research

  policy "owned" by VP for Research
- draft reviewed by IRBs, School CIOs, OGC, Social Science Committee, Provost, University Joint Committees on Inspection, ...

Harvard as Example, contd.

- (finally) approved October 2010

  http://www.security.harvard.edu/research-data-security-policy

- new post-IT reorganization effort to review & revise

  owner this time is joint research oversight committee

  aim to finish by next July

  make sure effort is not seen as “IT-driven” or “IT-owned”

Harvard Environment

- general university structure is distributed
  - "cloud education" (maybe 'quantum education')
  - informal associations among Schools
- recent (in the context of Harvard) push to change
  - e.g., new university CIO & IT reorganization effort
    - combining central IT and IT in some of the schools
  - also new university IT security officer & office

Harvard Research Environment

- research oversight is slightly less distributed
  - e.g., 3 Institutional Review Boards
    - coming under unified management
    - aim for common rules & forms
  - e.g., 3 Offices of Sponsored Projects
- one Vice Provost for Research
  - http://vpr.harvard.edu/
  - research policy, conflict of interest policy, IPR policy, etc.
  - one Chief Research Compliance Officer

HEISP

- Harvard Enterprise Information Security Policy
  - a set of University-wide policies to protect confidential information
  - annual training, etc
  - annual compliance assessment process
  - checked by Risk Management (Internal Audit) during audits
- also being reviewed & revised
  - university-wide committee led by chief security officer

HRDSP, Sections

- Introduction
- Research Information from Non-Harvard Sources
- Research Information from Harvard Sources
- Information Security Categories
- Legal Requests for Research Information

HRDSP, Introduction

- responsibilities: investigators:
  - disclose nature of data
  - prepare data security plans & procedures
  - implement plans & procedure
- responsibilities: IRB
  - ensure adequacy of investigators plans & procedures
- responsibilities: IT
  - assist investigators in determining proper levels
  - assist investigators in implementing security

HRDSP, Non-Harvard Data

- if data has a use agreement (DUA)

  protection must meet requirements in use agreement

  IRB can determine that DUA is too weak

    if so, treat as if data is from a Harvard source

- if research done in non-Harvard facility

  facility owner may define protection requirements

- otherwise

  treat as if data is from a Harvard source

HEISP: Data from Harvard Source

- human subjects research
  - research must be reviewed by a IRB
  - information used in research must be protected against inadvertent or inappropriate disclosure
  - IRB will confirm security level categorization
- other sensitive research
  - e.g. research with national security implications
  - researchers should work with school IT groups to determine data categories

HRDSP: Data Categories

- 5 levels of data about individually identifiable people
  - Level 5 - extremely sensitive information
  - Level 4 - very sensitive information (HEISP HRCI)
  - Level 3 - sensitive information about (HEISP other confidential information)
  - Level 2 - benign information
  - Level 1 - de-identified research information and other non-confidential research information

HRDSP: Why 5?

- started with HEISP - 3 levels
  - high risk confidential information (level 4)
  - other confidential information (level 3)
  - non-confidential information (level 1)
- added level 5
  - because non-network connected requirement is in some use agreements and is the right thing for some data
- added level 2
  - to deal with “minimal risk” data

HRDSP: De-Identification Key

- key for coded de-identified research information must be protected at the level that would have been applicable to the non-de-identified data
- what constitutes de-identification is not addressed in policy

HRDSP: Level 5

- description:

  Disclosure of Level 5 information could cause significant harm to an individual if exposed, including, but not limited to, serious risk of criminal liability, serious psychological harm or other significant injury, loss of insurability or employability, or significant social harm to an individual or group

- examples

  currently mostly requirement in data use agreements

  e.g., raw census data, some mental health records

HRDSP: Level 5 Protections

- stored in physically secure rooms under university control
  - not on janitor's key or building master key
- computers must not be connected to a network that extends outside the room

HRDSP: Level 4

- description

  Disclosure of Level 4 information could reasonably be expected to present a non-minimal risk of civil liability, moderate psychological harm, or material social harm to individuals or groups

- examples

  HEISP high risk confidential information (HRCI)

  e.g., subject's SSNs

  medical research records

  information with national security implications

HRDSP: Level 4 Protections

- do not store on user computers or devices
  even if encrypted (too much risk of error)
- servers in physically secure environment
  card based access best - create access log
- local network-based firewalls
- access limited to IRB approved individuals
- media: encrypt or store in a locked safe
- separate networks using private addressing
- regular vulnerability testing

HRDSP: Level 3

- description

  Disclosure of Level 3 information would could reasonably be expected to be damaging to a person's reputation or to cause embarrassment.

- examples

  most non-de-identified human research information

  student record information (FERPA)

  some commercial data

  employment records

HRDSP: Level 3 Protections

- encrypt laptops and portable devices
- use automatic patching
- virus protection
- encrypt all transfer over networks and on portable media
- limit access to those doing the research
- host-based firewalls
- lock up all non-electronic records

HRDSP: Level 2

- description

  Disclosure of Level 2 information would not ordinarily be expected to result in material harm, but as to which a subject has been promised confidentiality.

- examples

  data from reaction time experiments

  customer satisfaction survey data

HRDSP: Level 2 Protections

- good computer hygiene
  - secret complex passwords
  - not shared accounts
  - regular patching
  - avoid dangerous web sites
  - don’t respond to phishing

HRDSP: Level 1

- description
  - de-identified research information about people and other non-confidential research information
- examples
  - de-identified research information
    - but might be private until publication
  - student directory information
    - except for FERPA blocks
  - research information where no anonymity promised

Legal Requests for Research Info.

- forward any legal request of information (e.g., a subpoena, national security request or court order demanding disclosure of information in researcher possession) to OGC
- researchers not authorized to provide the information
- consider obtaining a Certificate of Confidentiality

  allow refusal to disclose

HRDSP: Other Information

- policies include specific guidance on how to do data collection in the field for each level data
- web site also includes:

  requirements when working with vendors

  process for responding to Freedom of Information Act (FOIA) requests (send to OGC)

  classified work (can not do)

  advice for travelers

  rules about paying subjects (i.e., tax requirements)

Implementation

- specific protection requirements for each level
  - existing HEISP level protection requirements well understood
  - Levels 5 and 2 will take some work
    - special facilities for Level 5
    - researcher cooperation for Level 2
- communications to researchers
  - annually by Deans
  - day-to-day by IRBs
- enforcement is an open question

Facility Certification

- facilities can get certified for particular level use
- IRB can rely on the certification for all research done in facility

  no need to review security plan for each project
- OK to use higher level facility for lower level research

  as long as higher level requirements followed

Web-Based Surveys

- rules in HEISP
- policy:

  Confidential information resulting from a survey or used as part of a data collection project must be protected from unauthorized access and improper sharing.

Web-Based Surveys: Rules

- prior IRB review required
- survey must have clear statement of purpose, data retention and access
- no level 4 or 5 data (e.g. no SSNs)
- encrypted connection required if confidential information is collected or displayed
- researcher can not have access to web log if anonymity promised

Commercial Web Survey Services

- generally OK for surveys that do not ask for confidential or identifying information
- usually have no individual contracts

  so no way to ensure proper protection for confidential information
- easy for researcher to mess up

  e.g., not knowing that IRB approval is required

  e.g. asking for email address for a iPod drawing on a survey that is supposed to be anonymous

The Cloud

- cloud computing services have the same problems as do web survey companies
  - e.g., no personal contracts, no default security
- additional cloud problem
  - no specific knowledge of where the data is stored
- advice – never use cloud services for level 4 or 5 data and think very carefully for level 3 data
  - except, maybe, the FISMA moderate compliant Amazon cloud service offering

Thank You