Acquisition, Management, Sharing, and Ownership of Data

Responsible Conduct of Research Training

UNIVERSITY OF NEVADA LAS VEGAS

Presented by: Joseph Lombardo, Executive Director
National Supercomputing Institute
&
Dedicated Research Network
lombardo@nscee.edu | 702-895-4153
What Type of Data Do You Use/Generate? What Do you Think are the Issues?
Data practices are central to all discussions of Responsible Conduct of Research (RCR)

- Fabrication and falsification (research misconduct) are typically data related.
- Sloppy data collection, management, or analysis practices can lead to biased results used by others.
- Incredible capacity for analysis and possibility of mistakes.
Data is a hot topic


- Authors point to the incredible amount of data now available and the opportunities and challenges related to data.
- The articles discuss many of the ethical dilemmas discussed in this presentation.
- **Big Data!**
What are data?

- NIH

“recorded information, regardless of the form or medium on which it may be recorded, and includes writings, films, sound recordings, pictorial reproductions, drawings, designs, or other graphic representations, procedural manuals, forms, diagrams, workflow charts, equipment descriptions, data files, data processing or computer programs (software), statistical records, and other research data.”
What are data?

In the Office of Management and Budget’s (OMB) Circular A-110, research data are defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues.
What are data?

- NSF

“What constitutes such data will be determined by the community of interest through the process of peer review and program management. This may include, but is not limited to: data, publications, samples, physical collections, software and models.”
What is Identifiable data?

- **Personal Health Information (PHI):** This is defined by HIPAA law and includes personal identifiers that are associated with medical information other than patient/subject self-reported information that may pertain to health. Information/data from medical records are considered PHI.

- **Personal Identifying Information (PII):** For the purposes of this policy, this includes information that identify a person including any or all of the following: (1) names; (2) social security numbers; (3) birth dates; (4) addresses; (5) IP addresses; (6) other data that could reasonably lead to discovering a personal identity *e.g.* criminal record, mothers maiden name

- **Limited Data Set** - Refers to PHI that excludes 18 HIPPA categories of direct identifiers! Commonly referred to as *de-identified data.*
1. Names.

2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP Code, and their equivalent geographical codes, except for the initial three digits of a ZIP Code if, according to the current publicly available data from the Bureau of the Census:
   a. The geographic unit formed by combining all ZIP Codes with the same three initial digits contains more than 20,000 people.
   b. The initial three digits of a ZIP Code for all such geographic units containing 20,000 or fewer people are changed to 000.

3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.
4. Telephone numbers.
5. Facsimile numbers.
6. Electronic mail addresses.
7. Social security numbers.
8. Medical record numbers.
9. Health plan beneficiary numbers.
10. Account numbers.
12. Vehicle identifiers and serial numbers, including license plate numbers.
15. Internet protocol (IP) address numbers.
16. Biometric identifiers, including fingerprints and voiceprints.
17. Full-face photographic images and any comparable images.
18. Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification.
Preliminary or “raw” data are not included for the purposes of access by the general public.

Investigators must retain this raw data in laboratory notebooks or records for purposes of validating research findings.

The raw data serves other purposes as well, such as patent applications, investigations of misconduct, or if the research results are used for public policy or regulatory purposes.

The definition provides the following exclusions: “This recorded material excludes physical objects (e.g., laboratory samples). Research data also do not include: (A) Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and (B) Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study.”
NSF now requires a data management plan for every award submitted. It must:

- Describe the data. Identify the data that will be measured, recorded, calculated, or modeled as well as the manner in which these tasks will be performed.
- Present the context of the data. Explain the rational for collecting or producing the data, as well as what insights could be gleaned from the data.
- Explain the nature of the data and identify the type of data collected.
- Describe the method for preserving and/or curating the data. Indicate how data will be backed up, as well as how it will be stored both on- and off-site.
NSF Data Management Plan Requirements Cont.: 

- Discuss the approach for accessing the data, if relevant. Be aware that when requested, data should be available for sharing within a reasonable period of time; understand that different communities may define the term “reasonable” in different ways.
- State how long the data will be preserved and/or curated. NSF requires data to be preserved at a minimum three years beyond the end of the award.
- Clarify ethical and/or privacy issues associated with the data, if relevant. Explain how these issues will be addressed.
- Detail intellectual property concerns associated with the data, if relevant. Explain how these concerns will be addressed.
- UNLV Sponsored Programs (http://research.unlv.edu/osp/resources.html)
Federal requirements: Managing and Sharing Data

- NIH (2003) requires a data sharing plan only when grant is $500,000 or more per year in direct costs.
- NIH’s policy encourages timely release and sharing of final research results
  - “Timely release and sharing” is defined as no later than the acceptance for publication of the main findings of the final data set.
- NIH supports the sharing of unique research resources or research tools under reasonable terms and conditions for dissemination and acquiring the tools.
  - No funding amount limit
# Elements of a Data Management Plan

- Data description
- Existing data
- Format
- Metadata (or data about data)
- Storage and backup
- Security
- Responsibility
- Intellectual property rights
- Access and sharing

- Audience
- Selection and retention periods
- Archiving and preservation
- Ethics and privacy
- Budget
- Data organization
- Quality Assurance
- Legal requirements
Examples

Data description

- **Generic Example 1:** This project will produce public-use nationally representative survey data for the United States covering Americans' social backgrounds, enduring political predispositions, social and political values, perceptions and evaluations of groups and candidates, opinions on questions of public policy, and participation in political life.

- **Generic Example 2:** This project will generate data designed to study the prevalence and correlates of DSM III-R psychiatric disorders and patterns and correlates of service utilization for these disorders in a nationally representative sample of over 8,000 respondents. The sensitive nature of these data will require that the data be released through a restricted use contract.
Example

Format

- **Generic Example 1:** Quantitative survey data files generated will be processed and submitted to the [repository] as SPSS system files with Data Documentation Initiative (DDI) XML documentation. The data will be distributed in several widely used formats, including ASCII, tab-delimited (for use with Excel), SAS, SPSS, and Stata. Documentation will be provided as PDF. Data will be stored as ASCII along with setup files for the statistical software packages. Documentation will be preserved using XML and PDF/A.

- **Generic Example 2:** Digital video data files generated will be processed and submitted to the [repository] in MPEG-4 (.mp4) format
**Example**

**Metadata**

- **Generic Example 1:** Metadata will be tagged in XML using the Data Documentation Initiative (DDI) format. The codebook will contain information on study design, sampling methodology, fieldwork, variable-level detail, and all information necessary for a secondary analyst to use the data accurately and effectively.

- **Generic Example 2:** The clinical data collected from this project will be documented using Clinical Data Interchange Standards Consortium (CDISC) metadata standards.
Example

Security

- **Generic Example 1:** The data will be processed and managed in a secure non-networked environment using virtual desktop technology.

- **Generic Example 2:** The main computer room that stores the program computer servers and network security equipment is protected by additional security features that require a separate authorization. Access to the data center is controlled by multiple biometric fingerprint readers. Any access to the computer by non-center persons is logged and requires a continuous escort by one of the center staff.

- **Generic Example 3:** The secure server environment that hosts the database is located within a hardened data center at the [repository], and is governed by standard Medical information security guidelines including the Health Insurance Portability and Accountability Act of 1996 (Public Law 104-191), also known as HIPAA.
**Budget**

**Generic Example 1:** Staff time has been allocated in the proposed budget to cover the costs of preparing data and documentation for archiving. The [repository] has estimated their additional cost to archive the data is [insert dollar amount]. This fee appears in the budget for this application as well.
Ethics and privacy

- **Generic Example 1:** For this project, informed consent statements will use language that will not prohibit the data from being shared with the research community.

- **Generic Example 2:** The following language will be used in the informed consent: The information in this study will only be used in ways that will not reveal who you are. You will not be identified in any publication from this study or in any data files shared with other researchers. Your participation in this study is confidential. Federal or state laws may require us to show information to university or government officials [or sponsors], who are responsible for monitoring the safety of this study.

- **Generic Example 3:** The proposed medical records research falls under the HIPAA Privacy Rule. Consequently, the investigators will provide documentation that an alteration or waiver of research participants' authorization for use/disclosure of information about them for research purposes has been approved by an IRB or a Privacy Board.
Federal requirements: Managing and Sharing Data

- Adequate management of federally-funded data
  - OMB stipulates 3 years or longer as required by granting agency (6 years by HIPAA, 2 years FDA)
  - If patent - life of patent - typically 20 years
  - If contested - until investigation is completed
  - Electronic copies are okay provided that there’s notification.
Who Owns the Data? (from HHS)

Case Vignette: Data Ownership

Dr. Smith works at The University and is the Principal Investigator on a large research project that is funded by the National Institutes of Health (NIH). However, while Dr. Smith wrote the original grant proposal, he does very little day-to-day work on the project. Instead, the Research Director, Betsy, oversees all aspects of the project, including staff supervision and all data management activities. In addition, Betsy has been lead author on several publications about the project's research findings.

Who owns the project and its data?

- The PI, Dr. Smith
- The Research Director, Betsy
- The University
- The National Institutes of Health
- No one person or organization
What are Some Ethical and Data Acquisition Issues?

- Various approvals may be needed prior to data collection (IRB, Institutional Animal Care and Use Committee, Land permits, etc).
- Dealing with unexpected data (outliers).
- Others need be able to reproduce/validate results, data may need to be available for this purpose (sharing).
- Proper data management takes a significant investment of time & effort.
Ethical and Data Acquisition Issues

- Must develop/use appropriate validated data collection methods
- Must have appropriate permissions (IRB, IACUC, Permits)
- Must protect data as it is collected
What are Some Data Management Concerns?

- If improperly managed, data may become invalid:
  - Human errors
  - Data transmission errors
  - Software malfunctions
  - Hardware malfunctions
  - Natural/structure disasters (fires, floods, etc)
- Data Storage: Original data stored safely; files backed up; samples saved so as not to degrade.
- Privacy and Confidentiality
- Retention: Generally must save data for 3 years after the completion of the project.
Data Management - Retention

There are many tales of early archaeologists burning wood from the ruins to make coffee. If we fail to curate the environmental archives we collect from nature at public expense, we essentially repeat those mistakes.

“Is there sufficient funding for your lab or research group for data curation?

8.8% YES, sufficient
10.9% YES, but not sufficient
80.3% NO

Science 11 February 2011:
Vol. 331 no. 6018 pp. 692-693
DOI: 10.1126/science.331.6018.692
INTRODUCTION: Challenges and Opportunities
Data Analysis Concerns

- How do you deal with outliers? If included in analysis may bias mean/variance, estimates, p-values, and conclusions. If removed should have good reason and describe when data is shared.
  - Check for recording error
  - “Rare” event syndrome (60 degree day in June in Las Vegas)
  - Transformation of data for analysis
  - Accommodation (nonparametric statistical methods)
- Statistical analyses
  - Proper analysis technique depends on type of data
  - Evaluation of sample size to ensure that it is valid
- Image manipulation
- Data selection
  - “Typical results”
Do you have the necessary expertise in your lab or group to analyze your data in the way you want?

The next few years [particularly in medicine] the volume of data we need to analyze will expand exponentially.
What are some Data Sharing Issues?

- Preliminary data usually should not be released.
- Confirmed or validated data - data should be confidential until accepted for publication.
- Published data - Once published there is an expectation that all the information about that experiment (including final data) should be freely available.
- MTA - Material Transfer Agreements (legal contracts) are often required. MTA’s are often required for sharing research materials also. Contact OSP for more information.
Case Vignette: Data Sharing

After completing the first phase of data analysis, 1 of the 3 main hypotheses of Dr. Smith and the research team was proven correct. However, the team also found some results from another facet of the project that they were not expecting. While these secondary results do not directly impact Dr. Smith's primary research questions, they may affect at least 3 other investigators' research. The results appear to be pretty definitive, but data analysis is still being conducted on other parts of the project.

The 2 Research Associates working on the project, Samantha and Enrique, are insistent that the team should immediately publish their findings in a journal, since the results may have implications on other PIs' work. Dr. Smith and Betsy, the Research Director, do not intend to publish any results for at least another year, since the research is ongoing and some questions are still unanswered.

What should the research team do?

_ They should publish the results in a journal as soon as possible.
_ They should tell the funding agency about the findings, and let the agency disseminate the information if it wants.
_ They should contact the other researchers to let them know the preliminary results.
_ They should do nothing; they aren't legally allowed to share their results until all data have been fully validated.
Privacy and Confidentiality

  - Authorization prior to the use of a subject’s individually-identifiable health information for research purposes.
  - Specific security requirements for health data access and storage.
- As of May 6, 2004, NSHE, a hybrid covered entity, designated its health care components of UNLV as follows:
  - Medical School and any associated clinics
  - Dental School and any associated clinics
  - Student Health Center & Pharmacy
  - Student Wellness
  - Center for Individual, Couple & Family Counseling
  - Center for Health Information Analysis
  - National Supercomputing Institute
Five Best File Encryption Tools

- VeraCrypt (Windows/OS X/Linux) - open source (free)
- Ax Crypt (Windows) - free, for the premium version $30
- BitLocker (Windows) – free
- GNU Privacy Guard (Windows/OS X/Linux) - GNU General Public License ... free
- 7-Zip (Windows/OS X/Linux) - open source (free)
Privacy and Confidentiality

- FDA, select agents, dual-use technologies (homeland security), patents, export controls (dissemination at foreign meeting), and classified research will affect these.
- State “sunshine” laws (e.g., open meeting) may grant access
- Industry sponsored research - pharmaceutical trials
- Biobanking and genomic data
Case Vignette: Research Team Responsibilities

After collecting data for about a year, Dr. Smith's research team revisited their original research questions. They decided to investigate an additional hypothesis related to a new issue that arose during the study. This change required adding about a dozen new questions to the self-administered questionnaire.

One day, the Research Assistant, Joel, realized that they had been administering the revised survey to subjects, but the Institutional Review Board (IRB) had not yet approved the changes.

Whose responsibility was it to make sure that data collection did not continue until the IRB approved the changes?

__ The PI, Dr. Smith
__ The Research Director, Betsy
__ The Research Associates, Samantha and Enrique
__ The Research Assistant, Joel
Who owns it?

- Traditionally it was the investigator
- More recent- the grantee (institution)
  - NIH and NSF both state the property belongs to the grantee but that they retain access.
  - State may actually retain possession if it’s a state school
    - Not in Nevada - given to Regents who in turn have given the intellectual property to the university.
    - Bayh Dole Act (or Patent and Trademark Law Amendments) Universities are given the right to retain title to inventions funded with federal $$ and right to control licensing – must share w/ inventor
- Rights of subjects to their “data”
Sources

- HHS Office of Research Integrity:  http://ori.hhs.gov/

- University of Idaho, Office of Research Assurances:  https://www.uidaho.edu/research/faculty/research-assurances

- NIH Data Sharing Plan Policy  

- NIH Data Sharing Workbook (pdf and MS word samples):  
  http://grants.nih.gov/grants/policy/data_sharing/


- NSF Data Sharing Policy:  
  https://www.nsf.gov/pubs/policydocs/pappg17_1/pappg_11.jsp#XID4

- NSF Data Management Plan Requirements:  
  https://www.nsf.gov/pubs/policydocs/pappg17_1/pappg_2.jsp#IIC2j