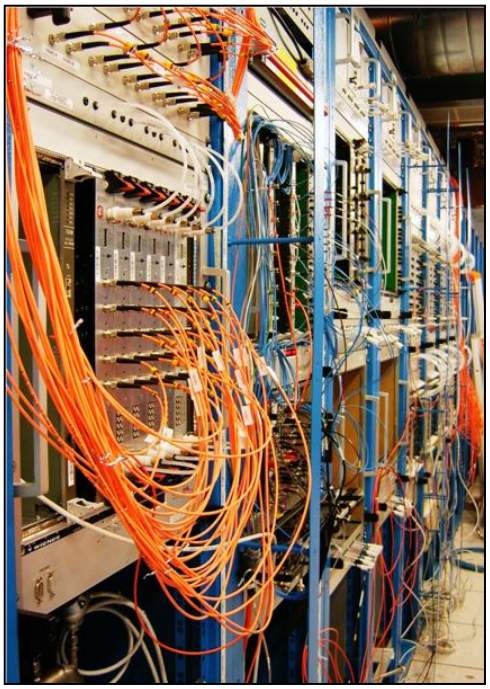
USC
UNIVERSITY
OF SOUTHERN
CALIFORNIA

Data Management and Acquisition



Office for the Protection of Research Subjects (OPRS)

Dalar Shahnazarian, MSW Candidate, IRB Student Mentor;
Susan L. Rose, Ph.D.; Jennifer Hagemann, MS; Monica Aburto

This series of booklets is adapted from the Collaborative Institutional Training Initiative (CITI) Responsible Conduct of Research online course available at www.citiprogram.org

Also available in the RCR Series:

Using Animal Subjects in Research

Collaborative Research

Peer Review

Conflicts of Interest and Commitment

Human Subjects Research

Mentoring Student Researchers

Research Misconduct

Responsible Authorship and Publication

About the Source Material

The Collaborative Institutional Training Initiative (CITI) web based education program, developed by the University of Miami and the Fred Hutchinson Cancer Research Center, offers training in Human Subjects Research, the Responsible Conduct of Research, and Good Clinical Practice. CITI is currently used by over 1130 participating institutions and facilities from around the world and offers online course material in more than seven different languages. CITI RCR was developed with public funds and thus allowed access to material used to create these booklets.

Introduction to Data Management and Acquisition

Data are the foundation of research and scientific advancement. Accordingly, data integrity is paramount. The first step in good data management is designing research that creates meaningful and unbiased data, that will not waste resources and that will appropriately protect human and animal subjects.

Proper data collection, retention, and sharing are vital to the research enterprise. If data are not recorded in a fashion that allows others to validate findings, results can be called into question. This booklet will discuss circumstances in which parts of data must be protected and not shared, in order to secure the intellectual property inherent in new inventions and protect the confidentiality of human research subjects. Who actually owns data collected in an academic environment for a research project is a legal/ethical issue that will be explained. Case studies and references have also been provided.

What is Data?

Data is any collection of facts, measurements, or observations used to make inferences about the world in which we live.

Different disciplines have different notions of what constitute data. Data can range from material created in a wet laboratory, such as an electrophoresis gel or a DNA sequence, to information obtained in social-science research, such as a filled-out questionnaire, video and audio recordings, or photographs. Data can be astronomical measurements, microscope slides, climate patterns, cell lines, field notes, soil samples, or results of statistical analyses.

Legal and Regulatory Concerns

Much research has legal and/or ethical requirements that must be adhered to for the safety of subjects, to establish intellectual property rights, and protect the public and the institution.
A researcher must be mindful of the following:

- Requirements of statutes and regulations, as imposed by federal, state or local government authorities.
- Certificatory standards of private bodies (e.g., JCAHO).
- Policies of one's own organization.
- Ethical standards set by one's profession.
- Broader social norms.

Compliance with laws and organizational policies constitute the "minimum necessary" for responsible conduct, and every researcher must be aware of them.

Privacy and Confidentiality

Data is protected by the rules about who can access information, and under what conditions – sometimes called the "privacy" or "confidentiality" rules. It also refers to the technical, physical and administrative safeguards and controls that fall under the heading of information security. Absent appropriate information security practices, privacy/confidentiality protections are empty promises.

Privacy refers to a research participant's willingness to allow access to themselves and their personal information. **Confidentiality** refers to the process of protecting and using private data or specimens. Plans for managing data in a confidential manner must be appropriate to the study being proposed

The compliance rules that apply to a particular kind of data are conditioned by many factors. One of them is the nature of the data itself. Personal Identifiable Information (PII) such as names and social security numbers, are protected by many states' laws and are subject to federal regulation.

"Health" data is protected by the federal Health Insurance Portability and Accountability Act (HIPAA), while "education" data is safeguarded by the federal Family Education Rights and Privacy Act (FERPA), as well as most state's laws. Financial data is protected by the federal Financial Services Modernization Act (FSMA), as well as almost all state's laws. Most states also have passed laws to support these specific federal statutes.

If you are conducting research outside of the United States, you must be aware of other countries' data protection requirements. Countries in the European Union, for example, generally have much stricter data protection requirements than does the U.S.

Collection and Use of Data

Many laws and regulations apply, depending on the kind of data, from where the data are obtained, and the purposes for which the data are sought. Because of the complexity of data handling, investigators are always urged to make use of the expertise of their organization's compliance and/or legal departments, in order to be certain exactly what constraints apply.

Substantial fines and other sanctions are possible for failing to meet requirements, another good reason to find out exactly what rules apply to your research.

Data Acquisition

Data Selection

Data selection is the process of determining appropriate data types and sources, as well as suitable instruments for the collection of data. This determination is often discipline-specific and is primarily driven by the nature of the investigation, existing literature, and accessibility of desired data sources.

What Data to Collect and When

Selection of "appropriate" data is constrained by issues of cost and convenience, not only the data's ability to adequately answer research questions. During the design phase of a study, it is important to consider the regulatory requirements and financial costs associated with collecting, handling, and storing the proposed data.

Investigators must always assess the risk that cost/convenience factors might compromise the integrity of the research endeavor.

Data *selection* should precede actual data *collection*. Clear selection standards set in advance work to prevent selective data reporting later – that is, selectively excluding data that are not supportive of a research hypothesis.

How Should Data be Collected?

Different disciplines have preferences for different approaches, and for what constitutes acceptable "rigor" for reliability and validity of results. This is one reason why prior review of the existing literature on a topic is imperative when designing a research protocol. For example, a key component of most protocol designs will be the sample size (or “n”). From a purely methodological perspective, determining the sample size depends on how large an error one is willing to tolerate in estimating population parameters or, put differently, what effect size will be required for the result to be considered significant. Research designs with too small of an “n” (n= number of subjects, be it humans or animals) are unethical because they waste resources and present results that are unrepresentative of the population being studied. Similarly, if a proposed "n" is greater than what is needed to adequately test the hypothesis, this is also a waste of resources. These must be determined in advance of commencing data collection. But statistical explanatory power must be balanced against time, cost and other practical considerations, just like every other element of the protocol.

Data Collection

Data collection is the process of gathering and measuring information on variables of interest in an accepted systematic fashion.

Data collection methods vary by discipline and data type of but the responsibility to conduct accurate and honest collection applies in all cases. Both the selection of appropriate data collection instruments (existing, modified, or newly developed) and clearly defined instructions for their correct use reduce the likelihood of errors occurring.

Consequences from improperly collected data include:

- Inability to answer research questions accurately.
- Inability to repeat and validate the study.
- Distorted findings
- Wasted resources.
- Misleading other researchers to pursue fruitless avenues of investigation.
- Compromising decisions for public policy or private decision-making.
- Causing harm to human participants and animal subjects.

Training the Research Team

An important component of assuring quality data collection is developing a rigorous and detailed recruitment and training plan for everyone who participates in the investigative effort. The training aspect is particularly important to address the potential problem of staff unintentionally deviating from the research plan. In addition, the structure of communication between study staff and participants must be clearly delineated in the protocol, so transmission of any change in procedures to staff members will not be compromised. The Principal Investigator has overall responsibility for providing adequate training and oversight of study conduct.

Quality Control of Data

Regular reviews and audits of records, whether the data are quantitative or quantitative, will allow investigators to verify that collection is proceeding according to procedures established in the research design. Where possible, researchers should try to build checks-and-balances into the collection process, in order to spot small problems that are otherwise destined to grow into large ones.

Storage and Security

Research data must be stored in a safe and secure manner during and after the conclusion of a research project. Reliable security policies and procedures to safeguard data handled electronically as well as through non-electronic means, such as paper files, journals, and laboratory notebooks are essential. Data must be protected while "at rest" and in transit, during collection, analysis, and disclosure to others.

Protection of Data

There are three core goals for information protection: **confidentiality, integrity and availability**.

Confidentiality

Confidentiality refers to limiting information access and disclosure to authorized users and preventing access by or disclosures to unauthorized persons. Federal regulations, such as the Common Rule and FDA regulations require attention to the privacy of research subjects. HIPAA adds its appropriate safeguards requirements for most research data derived from health care records.

Availability

Availability refers to the accessibility of information to authorized users. Also, technical failures risk making data unavailable to anyone. It is essential to make periodic backup copies of a data collection, and store these copies in a secure,

secondary location that is protected both from intruders and environmental threats.

Integrity

Integrity refers to the trustworthiness of information. Namely, integrity means that data have not been changed inappropriately after recording, accidentally or deliberately. The information must reflect the actual circumstances (validity) and be able to generate identical data (reliability).

Data Retention and Disposal

How long the data will be kept after a project is over usually depends on:

- The nature of the project, including potential ongoing interest in or need for the data.
- Costs of maintaining the data in a secure state over the long run.
- The research sponsors' requirements and guidelines.
    - Under current federal Department of Health and Human Services requirements, for example, research records must be maintained for at least three years after the last expenditure report. Other federal regulations or institutional guidelines may require that data be retained for longer periods.
    - Note that litigation "stops the clock" for data retention - materials must be retained until all investigation activities have concluded.

Retaining data on paper files and electronic media long past the end of a project can increase the chances of unauthorized access. Risks increase all the more when investigators leave the project, transfer to another institution, retire or die without establishing proper data management procedures, including disposal or archival storage.

Disposal of Data

Disposal of sensitive data requires care and technical expertise to ensure that the information could not be reconstructed from the storage media. When disposing of magnetically recorded data stored on computer hard drives, flash drives or floppy disks, multiple-pass erasures are required. Optical media need to be over-written (if of a re-writeable variety) or otherwise shredded.

If an investigator lacks the requisite expertise and tools, appropriate technical resources from his/her organization should be sought, or a third-party disposal contractor should be hired.

Data Analysis

Statistical Examination

Data analysis refers to the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data. The form of the analysis is determined specific type of the data and whether a qualitative or quantitative approach was taken.

The Appropriate Method

Researchers must be aware of a number of data analysis issues. "Scientific misconduct" becomes likely when researchers operate beyond the frontiers of their methodological expertise by not receiving sufficient analytic training. Review of proposed protocols – particularly proposed analytic methods – is critical when an investigator has any doubts about his or her level of competence. While analysis methods vary among scientific disciplines, the optimal stage for determining appropriate analytic procedures is early in the research process, preceding data collection. Waiting until later in the research process increases the risk that analytic decisions will be driven by consideration of which produces the most favorable results.

Publication and Reporting

Data publication and reporting is the process of preparing and disseminating research findings to the scientific community. Scholarly disciplines advance through dissemination and review of research findings at professional meetings and in publications in discipline-related journals. For the process to be maximally productive there must be a trust relationship between the author and readers regarding the accuracy and truthfulness of any publication.

Data Sharing and Ownership

Ownership Issues

Data "ownership" generally refers to both the possession of and responsibility for information. As a legal concept, it embraces the range of rights and obligations with respect to a data collection, including rights and obligations to share. Information control is conditioned by both technical capabilities – to access, create, modify, or package it – and also by legal-regulatory constraints.

All investigators and research staff should review the institution's policies with respect to data ownership to ensure that their understanding matches the institution's. Graduate students and postdoctoral fellows involved in research projects can mistakenly operate under the belief that that they own the data collected. However, they are considered employees of the university and the

institution usually owns the rights to the data, but this may vary by organization. When faculty members perform research on their own, at their own initiative and direction, the copyright typically belongs to them. If a specific third-party sponsor is involved, the sponsor / granting agency may set out the terms of copyright.

The Bayh-Dole Act of 1980

The Bayh-Dole Act of 1980 allowed universities to have control of the intellectual property, such as patents, generated from federally funded research. With a patent in hand, universities could exclusively license the patent to businesses. Many universities have benefited from the licensing revenue that has flowed from this arrangement. The law has encouraged new collaborative relationships between academic researchers and companies.

Rights and Obligations to Share

Professional norms and a long scientific tradition promote the principal of "openness" – that is, that investigators will publish significant research results and engage in the free exchange of information.

Sharing data has a number of benefits to society including protection of the integrity of scientific data. It permits (re)analyses:

- To verify or refute reported results.
- To refine results.
- To check if the results will remain valid under varying assumptions.

Recently, because of controversies about the secrecy of results from clinical trials for drugs, the pharmaceutical industry is beginning to offer its research data in online databases (de-identified, aggregated). Sharing has benefits for individual researchers, in that it can lead to collaborations. However, some information cannot be released because of privacy and human-subject protection concerns. Also, the release of research data/results before publication can jeopardize the ability of an investigator to be the first to publish findings.

In 2003, the National Institutes of Health instituted a new Policy on Data Sharing. The new policy applies to investigator-initiated one-year $500,000 grants and may have an impact on smaller grants too. The goal of the policy is to expedite the timely release and sharing of final data to enhance the research enterprise. The NIH is requiring that investigators asking for this or greater funding levels include with their grant applications information about how they plan to share the data generated from their research. If a grant is awarded, the data-sharing plan must be enacted.

The National Science Foundation also has a data-sharing policy: "NSF expects significant findings from research and activities it supports to be promptly submitted for publication, with authorship that reflects the contributions of those involved. It expects investigators to share with others at no more than incremental cost and within a reasonable time, the data, the samples, physical collections and other

supporting materials created or gathered in the course of the work. […] Exceptions may be allowed to safeguard the rights of individuals and subjects, the validity of results or the integrity of collections."

Conclusion

An important issue for researchers is how scientists balance the free exchange of some sensitive and/or proprietary scientific data and information with the possibility of unauthorized use.

Publishing in peer-reviewed journals or presenting in scholarly meetings is the primary mechanism for investigators to disseminate their findings to the research community. This community relies on authors to report the events of a study honestly and accurately. All researchers should be aware of the issues that compromise the integrity of data acquisition and collection.

Investigators must learn to negotiate the delicate balance that exists between an investigator's willingness to share data in order to facilitate scientific progress, and the obligation to sponsors, collaborators and subjects to preserve and protect data. Signed agreements of nondisclosure between investigators and their corporate sponsors can circumvent efforts to publish data or share with colleagues.

Case Studies

I. Who Owns Research Data?

Background

Jessica Banks, a Ph.D. student working with Professor Brian Hayward, a noted immunologist studying the functions of T-cell subsets in function GVH, has recently defended her dissertation and is now ready to file it and leave for her new job. During her second year, when starting research with Hayward, Banks divided her time among three projects. Then, in her third year, after consultation with Hayward, she decided to continue and expand upon one of the three lines of investigation for her dissertation research. This was also the project most closely related to Hayward's NSF grant at the time. Later, Banks' experimental plan and early results were included in Hayward's grant renewal. The other two promising lines of research were left incomplete. Banks' new job is a tenure-track position in a midsize private university on the west coast.

The data

Shortly before leaving for her job, **Banks comes to Hayward's office to make copies of research data** stored only on Hayward's computer using special software, which she also plans to copy. Although her new faculty position will place a heavy emphasis on teaching, she is looking forward to continuing to do some research as well. In particular, she is eager to pick up where she left off with the two incomplete projects she worked on earlier. Hayward comes in as Banks is downloading her material, and asks her what she is doing. She tells him, and he then says to her that she cannot take the data. "They belong to me," he says. Banks is confused. "But I did the work, and I wanted to follow up on it. **I can't do that without the data.**" Hayward is adamant. "I'm sorry, but you should understand this. Our research project was a joint enterprise, and all the work you did was funded by money I brought in via grants. The data do not belong to you or to me; they actually belong to the university, and the work will be continued with other students. I've already talked to one of the new students about working on those projects this fall." Banks, seeing her plans fall apart around her, protests, but Hayward is implacable.

Making a copy of the data

After a few minutes, she stalks away. Later that afternoon, Banks gets together with her classmate Paul Larson, and she tells him about her run-in with Hayward. "Look," Larson says. "Hayward has no right to deny you access to data. You did the work that generated all the data." "I know!" Banks says. "But Hayward wouldn't listen to that argument when I made it." "Here's my suggestion," Larson says after some reflection. "Just stop by his office and copy it sometime during the weekend. I happen to know Hayward will be out of town, so he'll never know. That's the fair thing to do." **Banks seems uncertain**, but she says she'll think about Larson's suggestion and decide before the weekend.

II. Share and Share Alike?

Background

Jim is a graduate student in the department of genetics. For his thesis research, he is mapping a gene involved in blood-sugar homeostasis. His work is part of a larger, multi-center study of the genetics of obesity. The larger study involves several thousand patients and includes information such as socioeconomic class, self-identified ethnicity, activity level, weight, and other medical data. Blood and DNA samples are maintained in Jim's lab along with a database that links unique identifiers but not patient names with the data. The study coordinator at each site has access to the encryption key; however, the students and other researchers working on the project do not. Researchers may use the database to retrieve and enter data pertaining to the samples, but they cannot learn the identity of the individuals in the study. **The study is said to be annonymized**.

Informed consent

The subject/patients involved in the study were recruited at various study sites. On first contact with a potential participant, a genetic counselor explains the study and arranges for a meeting to begin the informed-consent process. During this meeting, participants learn about the aims of the project, their role as subjects, and the risks and benefits involved in participation. **The consent forms** state that blood and DNA samples and the resulting data will be **anonymized**, that subjects may withdraw at any time, and that samples will be used exclusively for this study. If individual participants' samples are to be used in unrelated research, they must be re-contacted and they must go through a second consent process, specific to the new study.

Let's share the sample

Jim's project involves a subset of several hundred samples from the obesity study. One day, Renee, one of the other graduate students in the lab, approaches Jim and starts asking questions about the samples he's working with. She explains that for her work on sickle-cell anemia and mutations in a hemoglobin gene in African-Americans she needs 50 ethnically matched control samples. Since Jim has access to such a large collection of samples, Renee asks if she can take small aliquots of some of his samples from the obesity study. She tells Jim that she will not be looking at disease in these patients and is not really doing a "study" on them. **She just needs them as controls,** and she doesn't even need that much DNA. "Which box are they in?" Renee asks, as she heads for the freezer. Renee was standing at the freezer with the door open when Jim said, "I'd be happy to tell you more about our samples, Renee, but you had better talk to Jane, the study coordinator, about getting consent from the obesity-study participants if you really want to use them for your study." He went on, "Another option, which might be faster, is to just order a set of **anonymous samples from a commercial DNA bank.** It would really be a pain to re-contact all of those people just for a set of controls."

Resources

USC Statistical Consultation and Research Center (SCRC)
http://www.usc.edu/schools/medicine/departments/preventive_medicine/divisions/biostatistics/research/scr.html

USC High Performance Computing Center
http://www.usc.edu/hpcc/systems

USC Geographic Information Systems Laboratory
http://uscgislab.net/incEngine/?content=main

USC Center for Excellence in Research
http://www.usc.edu/research/about/vp/cer/index.html

CITI Program:
www.citiprogram.org

USC Contacts

Office for the Protection of Research Subjects
3720 South Flower Street, Third Floor
Los Angeles, CA 90089-0706
Tel (213) 821-1154
Fax (213) 740-9299
E-mail: oprs@usc.edu
https://oprs.usc.edu/

Health Sciences Institutional Review Board
General Hospital, Suite 4700
1200 North State Street
Los Angeles, CA 90033
Tel (323) 223-2340
Fax (323) 224-8389
E-mail: irb@usc.edu
https://oprs.usc.edu/hsirb/

University Park Institutional Review Board
Credit Union Building (CUB), Suite 301
3720 S. Flower Street
Los Angeles, CA 90089
Tel (213) 821-5272
Fax (213) 821-5276
E-mail: upirb@usc.edu
https://oprs.usc.edu/upirb/

Office of Research
Credit Union Building, Suite 325
3720 S. Flower Street
University of Southern California
Los Angeles CA 90089-4019
Tel (213) 740-6709
Fax (213) 740-8919
E-mail: vice.president.research@usc.edu
http://www.usc.edu/research/

CITI Helpdesk
Tel (213) 821-5272
E-mail: citi@usc.edu
https://oprs.usc.edu/education/citi/

iStar Technical Help
Tel (323) 276-2238
E-mail: istar@usc.edu
Web: http://istar-chla.usc.edu

Office of Compliance
3500 Figueroa Street
University Gardens Building, Room 105
Los Angeles, CA 90089-8007
Tel: (323) 740-8258
Fax: (213) 740-9657
E-mail: complian@usc.edu
http://www.usc.edu/admin/compliance/

USC Stevens Institute for Innovation
3740 McClintock Ave. Hughes EEB 131
Los Angeles CA 90089
Tel: (213) 821-5000
Fax:(213) 821-5001
http://stevens.usc.edu/

Health Research Association (HRA)
1640 Marengo Street, 7th Floor
Los Angeles, CA 90033
Tel (323) 223-4091
Fax (323) 342-0947
Web: http://www.health-research.org/

IRB Student Mentor
Tel (213) 821-1154
E-mail: irbgara@usc.edu
https://oprs.usc.edu/education/mentor/

Office of Contracts and Grants-UP
Credit Union Building (CUB), Suite 303
3720 S. Flower Street
Los Angeles, CA 90089
Tel: (213) 740-7762
Fax: (213) 720-6070
http://www.usc.edu/research/dcg/

Office of Contracts and Grants-HSC
1540 Alcazar Street, CHP 100
Los Angeles, CA 90033-9002
Tel: (323) 442-2396
Fax: (323) 442-2835
http://www.usc.edu/research/dcg/