

Good Data Practices

- Series Recap
 - Session 1: Early Data Planning for Research
 - Session 2: Managing and Documenting Data Workflow

Session 3: Planning for Data Re-use

- Data activities at project close
- Why make research-generated data available for re-use?
- Policy on sharing data for re-use
- Planning for sharing data for re-use
- Documentation required for re-use

Poll Question #1

- Are you planning or already working on a project that will produce a dataset that might be shared for re-use?
 - Yes
 - No

Session 3: Outline

- **Data activities at project close**
- Why make research-generated data available for re-use?
- Policy on sharing data for re-use
- Planning for sharing data for re-use
- Documentation required for re-use

Data activities at project close

- Data handling at project close
 - Decision to share data for re-use
 - R&D Committee notification
 - IRB permission/notification
 - VHA records retention requirements for research data

Data activities at project close

- VHA records retention requirements for data
 - What to retain?
 - Analytic dataset
 - Unique source data
 - How long to retain?
 - VHA 1200.05 Requirements for the Protection of Human Subjects in Research:
 - Retain secured copy of select dataset(s) until a record control schedule for VA research data is created.

Data activities at project close

- Example: Project closure process –Hines VA Hospital
 - Turn in study data disks to the research office – if any exist
 - Discontinue using identifiable data
 - Request that IT staff remove research access & secure data on network server

Session 3: Outline

- Data activities at project close
- **Why make research-generated data available for re-use?**
- Policy on sharing data for re-use
- Planning for sharing data for re-use
- Documentation required for re-use

Why make research-generated data available for re-use?

“Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. “

August, 2012 – IBM, Bringing Big Data to the Enterprise

Why make research-generated data available for re-use?

- Reduce, Reuse, Recycle!
 - **Reduce** redundant (expensive) data preparation
 - **Reuse** existing research data
 - **Recycle** subsets of research data



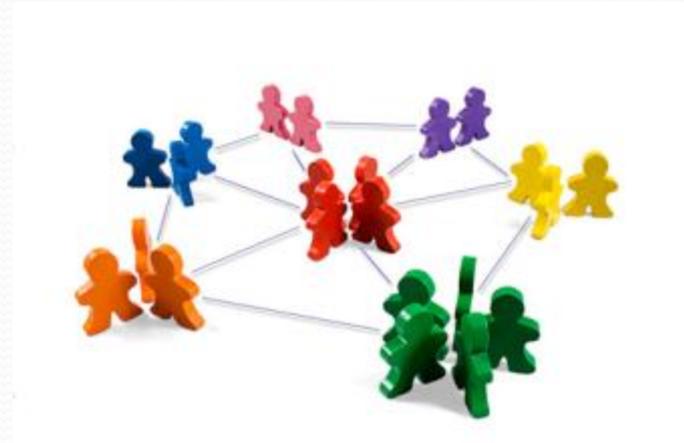
Why make research-generated data available for re-use?

National Institutes of Health: Goals of data sharing

- Permit access to unique data
- Expedite the translation of research into knowledge, products & procedures
- Permit testing of new alternative hypotheses & methods
- Support methods and measurement studies
- Facilitate the education of new researchers
- Reinforce open scientific inquiry
- Comply with funder requirements

Why make research-generated data available for re-use?

- Why make your data available for re-use?
 - To re-use the data yourself
 - Promote your research
 - Enable new discoveries with your data



Session 3: Outline

- Data activities at project close
- Why make research-generated data available for re-use?
- **Policy on sharing data for re-use**
- Planning for sharing data for re-use
- Documentation required for re-use

Policy on sharing data for re-use



- **NIH Data Sharing Requirements**

- All data should be considered for data sharing
 - Applications > \$500,000 must include data sharing plan
- Data should be
 - Widely & freely available
 - With timely release and sharing
 - While protecting patient/subject privacy

- NIH guidance:

http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

Policy on sharing data for re-use



VA | Defining
HEALTH CARE | **EXCELLENCE**
in the 21st Century

- **VHA Policy on Data Re-use for Research**
 - Permits re-use of research-generated data within the VHA in an IRB approved research data repository
 - VHA Handbook 1200.12 *“Use of Data and Data Repositories in VHA Research”*

Policy on sharing data for re-use

- Data sharing authorization checklist
 - ✓ HIPAA compliant patient authorizations
 - ✓ IRB approval of waiver of HIPAA authorization
 - ✓ VHA data owners/steward permission
 - ✓ Agreements with non-VHA data sources



Session 3: Outline

- Review of planning and documentation
- Why make research-generated data available for re-use?
- Policy on sharing data for re-use
- **Planning for sharing data for re-use**
- Documentation required for re-use

Poll Question #2

- Have you used public data from an external sources such as MEPS or ICPRS?
 - Yes
 - No

Planning for sharing data for re-use

- Key Components of Data Sharing Plans
 - What data are to be shared?
 - What is the authority to share for re-use?
 - How will data re-use be approved?
 - In what form will the data be shared?
 - How will data be provided?
 - How will data be protected?
 - Where will the data be stored?

Planning for sharing data for re-use

- What makes a good data repository host?
 - Data security
 - Scheduled back-ups
 - File recovery system
 - Adequate volume capacity
 - Compatible data formats
 - Long-term data retention capacity
 - Data access provisioning capability



ICPSR

Data Management & Curation

[Quality](#)[Preservation](#)[Access](#)[Confidentiality](#)[Citation](#)[Tools & Services](#)

Data Management & Curation

ICPSR stores, curates, and provides access to scientific data so others can reuse the data and validate research findings. Curation, from the Latin "to care," is the process that ICPSR uses to add value to data, maximize access, and ensure long-term preservation.

Data curation is akin to work performed by an art or museum curator. Through the curation process, data are organized, described, cleaned, enhanced, and preserved for public use, much like the work done on paintings or rare books to make the works accessible to the public now and in the future. With the modern Web, it's increasingly easy to post and share data. Without curation, however, data can be difficult to find, use, and interpret. Through curation, ICPSR provides meaningful and enduring access to data.

Quality



Data at ICPSR are enhanced with meaningful information to make it complete, self-explanatory, and usable for future researchers. As a repository, ICPSR adheres to standards that demonstrate it is organizationally, procedurally, and technologically sound as a trustworthy data custodian.

- [Preparing Data for Archiving](#)
- [Curating Data at ICPSR](#)

Preservation



Digital preservation is the proactive and ongoing management of digital content to lengthen the lifespan and mitigate against loss. ICPSR preserves its data resources for the long-term, guarding against deterioration, accidental loss, and digital obsolescence. ICPSR has a 50-year track record of reliably storing research data.

- [Trusted Digital Repositories](#)
- [Digital Preservation Policies and Planning at ICPSR](#)

Access



ICPSR hosts data in a repository with powerful search capabilities. Indexed by all the major search engines, ICPSR data are easily discoverable and widely accessible to the public.

Confidentiality



Data at ICPSR are screened for confidentiality and privacy concerns. Stringent protections are in place for securing and distributing sensitive data.

Planning for sharing data for re-use

- What makes a good VA data repository?
 - Data security
 - Scheduled back-ups
 - File recovery system
 - Adequate volume capacity
 - Compatible data formats
 - Long-term data retention capacity
 - Data access provisioning capability





VA
Informatics and
Computing
Infrastructure

Welcome to VINCI Workspace

Standard and Development Workspaces

[Standard Workspace](#)

[Development Workspace](#)

[Help](#)

Important Instructions

Click once on any icon below to start the application.

When prompted to enter your login, please enter it in the form of **DomainName\Login**.

This site is only supported on the Internet Explorer browser with Java Script enabled.

If you are not able to see the application icons below, [click here](#)  to access the VINCI Standard Workspace Desktop to access all applications.

									
Acrobat Distiller 9	Adobe Acrobat 9 Pro	Adobe Reader X	Altova XMLSpy	ArcMap 10	EndNote	Explorer	GATE 5.2.1 GUI	GuardianEdge Removable Storage	MATLAB R2011b
									
Microsoft Access 2010	Microsoft Excel 2010	Microsoft InfoPath Designer 2010	Microsoft InfoPath Filler 2010	Microsoft OneNote 2010	Microsoft PowerPoint 2010	Microsoft Word 2010	MySQL Query Browser	notepad++	NVivo 8
									
Protégé 3.3.1 (Knowtator)	Protégé 3.4.8 (Knowtator)	R x64 2.15.2	R x64 2.15.3	R x64 3.0.1	SAS & SAS Grid 9.2	SAS & SAS Grid 9.3	SAS Enterprise Guide 4.3	SAS Enterprise Guide 5.1	SAS Grid Enterprise Miner
									
SAS Power and Sample Size 3.1	SAS Universal Viewer	SAS XML Mapper	SPSS Modeler 15	SPSS Statistics 21	SQL Server Management Studio 2012	StataMP 11 (64-bit)	StataMP 12 (64-bit)	StatTransfer 9	TextPad
									
v3nlpClient	v3nlpDictionary	VinciFT	WinDiff	WinZip 10.0					

Poll Question #3

- If there were a central research data repository available in the VHA, how likely would you be to share data from one of your research projects for re-use?
 - 5 = Very likely
 - 4 = Likely
 - 3 = Maybe
 - 2 = Unlikely
 - 1 = Never

Session 3: Outline

- Data activities at project close
- Why make research-generated data available for re-use?
- Policy on sharing data for re-use
- Planning for sharing data for re-use
- **Documentation required for re-use**

Documentation required for re-use

- Importance of documenting as you go
 - Accurate, systematic record of the research process
 - Captures decisions and the reasoning behind them
 - Reduces mistakes, confusion and wasted time
 - Provides the basis for re-using data with confidence
- It's all in the details



ICPSR Deposit Data



Data Preparation Guide

Introduction

1. Proposal Development and Data Management Plans
2. Project Start-Up
3. [Data Collection and File Creation](#)
4. Data Analysis
5. Preparing Data for Sharing
6. Depositing Data

References

Related Links

Acknowledgments

[Download PDF version of the Guide](#)

Guide to Social Science Data Preparation and Archiving Phase 5: Preparing Data for Sharing

This chapter addresses the critical final steps researchers should undertake in preparing to archive and/or disseminate their data.

Respondent Confidentiality

Much of this guide has focused on on data preparation methods that can serve the research needs of both principal investigators and analysts of secondary data. In this chapter, however, we highlight one area of divergence necessitated by the responsibility to protect respondent confidentiality. Researchers must pay special attention to this issue. Once data are released to the public, it is impossible to monitor use to ensure that other researchers respect respondent confidentiality. Thus, it is common practice in preparing public-use datasets to alter the files so that information that could imperil the confidentiality of research subjects is removed or masked before the dataset is made public. At the same time, care must be used to make certain that the alterations do not unnecessarily reduce the researcher's ability to reproduce or extend the original study findings.

Below, we suggest steps that principal investigators can take to protect respondent confidentiality before submitting their data for archiving. But first, a quick review of why this is important.

The principles of disclosure risk limitation

Social scientists must demonstrate a deep and genuine commitment to preserve the privacy of the subjects whom they study in the course of their research. Most often applied to individuals who consent to be interviewed in surveys, this commitment extends also to groups, organizations, and entities whose information is recorded in administrative and other kinds of records.

"Institutions conducting research using human subjects funded by the federal department of Health and Human Services are responsible for compliance with the federal regulation on Protection of Human Subjects (45CFR46). Every such university and research institution must file an "assurance of compliance" with the HHS Office for Human Research Protections that includes

VI. Documentation Files. ICPSR prefers to receive documentation and other files associated with your data collection in electronic formats such as ASCII (text files) or Microsoft Word. The original source document is always preferred, but we will accept PDF files if they are PDF-A (image plus text). The codebook should specify the data position for each variable, describe the contents of each variable, provide full question text, and identify the range of possible codes and their meanings for each variable. Please provide the following documents, if available:

- Codebook
- Final project report, project summary, or other description of the project
- Bibliography of publications describing or resulting from the data
- Summary statistics (frequency distributions, means, etc.) of all variables
- Data collection instrument(s). An electronic version of each instrument, if applicable, including interview schedules, self-administered questionnaires, data collection forms for transcribing information from records, paper tests and scales, screening forms, and call-report forms, should be included along with a description of the circumstances in which each was used (study populations, time periods, etc.)

File name (include suffix, e.g., codebook1.doc)	File format (e.g., MSWord, ASCII text, WordPerfect)

VII. Data File Inventory. ICPSR prefers to receive data files in the following formats: SAS transport, SPSS portable, and Stata. It is also helpful to have value and variable labels and question text provided electronically. Provide an inventory of all data files below in the order they appear on your storage media.

Data File #1:	
File Name	
File Structure	<input type="checkbox"/> rectangular <input type="checkbox"/> hierarchical
# Cases	
# Variables	
File Format	<input type="checkbox"/> SAS transport <input type="checkbox"/> SPSS portable <input type="checkbox"/> Stata <input type="checkbox"/> ASCII <input type="checkbox"/> Other

Data File #2:	
File Name	
File Structure	<input type="checkbox"/> rectangular <input type="checkbox"/> hierarchical
# Cases	
# Variables	
File Format	<input type="checkbox"/> SAS transport <input type="checkbox"/> SPSS portable <input type="checkbox"/> Stata <input type="checkbox"/> ASCII <input type="checkbox"/> Other

Data File #3:	
File Name	
File Structure	<input type="checkbox"/> rectangular <input type="checkbox"/> hierarchical
# Cases	
# Variables	
File Format	<input type="checkbox"/> SAS transport <input type="checkbox"/> SPSS portable <input type="checkbox"/> Stata <input type="checkbox"/> ASCII <input type="checkbox"/> Other

[Copy, paste, and fill in additional tables as needed]

1. Are the data are compressed? yes no
2. If the data are compressed, what software was used to compress them?

Details: ICPSR Data Deposit Form

Note:

- Codebook
- Data collection instruments
- Data File Inventory

VARIABLE: COPAY

DESCRIPTION:

Three co_pay groups(no copay, some copay, all copay) based on veteran's priority category

Data Type: Numeric

Label: 'if PRI01_8 =1 then copay=0 if 2 <= PRI01_8 <= 6 then copay=1 if PRI01_8>6 then copay=2'

Program Location: hsrfiles\$\PartD\7 Programs\Jenny_huo\demog.sas

SOURCES:

Variable Name	File	Description
pri01_8	Name: VA Enrollment file(PSSG)	Three co_pay groups(no copay, some copay, all copay) based on veteran's priority category
ENRLEPRIO	Type: SAS	
ELIG	Location: hsrdata\$\PartD\Original_data\AA C_DATA\pssg04-10	

SELECTION CRITERIA or CALCULATION or CODES:

To determine priority category:

```
if pri01_8~=' ' then PRIORITY= pri01_8;
  else if pri01_8=' ' and ENRLEPRIO~=' ' then PRIORITY=ENRLEPRIO;

if PRIORITY ='1 ' then copay=0;
  else if '2 '<= PRIORITY <= '6 ' then copay=1;
  else if PRIORITY>'6 ' and priority not in ('99' '') then copay=2;
if priority in ('99' '') and elig~='' then do;
  if substr(elig,1,1)='A' then copay=0;
  else if substr(elig,1,1)='G' then copay=1;
  else copay=2;
end;
```

If listed as multiple priorities, assign the most frequently occurring

Details: Documentation of derived variable in an analytic dataset.

Note:

- Source
- Description
- SAS code

From: "Impact of Medicare Drug Benefit on VA Drug Use, Healthcare Use and Cost"; Funding Agency: Department of Veterans Affairs Health Services Research and Development; Study Number: IIR 07-165-2 - Kevin Stroupe, PI

Documentation required for re-use

Documentation checklist

- **Project Description**
 - Title
 - Creator
 - Identifier
 - Subject
 - Dates
 - Funders
 - Location



Checklist continued on next slide...

Documentation required for re-use

- **Data description**
 - File format
 - File structure
 - Variable list & description
 - Variable format
 - Code lists & definitions
 - Derived variables description



Checklist continued on next slide...

Documentation required for re-use

- **Methodology description**

- Cohort definition
- Sources of data
- Data processing description
- Data linkages
- Cleaning processes
- Data issues found & resolution
- Creation of derived variables
- The reasoning behind each decision that formed your data and analysis



Checklist continued on next slide...

Documentation required for re-use

- **Study Citation**

- PI & affiliation
- Descriptive title study & data
- Place & date of production
- Organizational name of data producer
- Sponsoring or funding agency and grant number
- Person/organization responsible for collecting data
- Special collaborator(s) (if applicable)
- Contact person for questions about the data.



- **Study Abstract**

Questions

Contact Information

Linda Kok, MA

VA Information Resource Center

A VA Health Services Research & Development Resource Center
working to improve the quality of VA research that utilizes
databases and information systems

Hines VA Hospital

VIReC@ va.gov

708-202-2413



Session 3: Planning for Data Re-use

- Recap
 - Data activities at project close
 - Why make research-generated data available for re-use?
 - Policy on sharing data for re-use
 - Planning for sharing data for re-use
 - Documentation required for re-use

Session 4: Research Application

- Preview
 - Proposal planning and development
 - Funding, IRB, and Study Initiation considerations
 - Documentation Content and Locations
 - Study Design and Implementation