Data Tools and Strategies for the Resource-Constrained

Christopher Callahan MS PMP

National Oral Health Conference

April 29, 2006

The findings and conclusions in this presentation are those of the author and do not necessarily represent the views of the Centers for Disease Control and Prevention

SAFER • HEALTHIER • PEOPLE™

Introduction and Background

- Christopher Callahan, MS PMP
  - ccallahan@cdc.gov
- Background in social science and addictions research
- Graduate of the CDC Public Health Informatics Program

Overview

- Data Management as a system
- Resources
- Planning
- Implementation
- Maintenance
- Security
- Software

Purpose

- Promote an understanding of an engineering approach to data management that maximizes utility and minimizes risks
- To increase awareness of software options that are available to you
- Improve your ability to make better choices regarding the tools you use

Data Management Past and Present

- The major data management tool of the past was the scientific notebook
  - Held observations in a structured format
  - Required discipline to use effectively
- The computer has replaced the scientific notebook
  - Facilitates analysis of larger volumes of data
  - Requires discipline to use effectively

Requirements for a Research Data Management Approach

- Need to be able to access and process data stored “somewhere”
- Need to make the best use of available resources
- Need to be able to reuse data
- Need to minimize (if not eliminate) rework
- Need to minimize training
- Need for a simple system to do all of this

Data management system basics

- Plan and document the data management system
- Implement and document processes
- Monitor and document activities
- Back up data
- Revise processes and update documentation
- Archive and document completed activities

Data Management Issues

- Comprehensive planning is mandatory
  - Monitor collection, reduce data transitions, designing database structure
- Documentation
  - Perhaps the single most important management factor
- Data Protection
  - Risks are diverse (corruption, loss, access)
- Continual correction of the system
  - Continuous progress towards strategic goals and objectives

Resource Profile

- Five resources that will affect your RDM system development
- Design towards strengths and aim to reduce risks
- Leadership (on research activities)
  - Strategic: decisions on planning, priorities and budget
  - Operational: recommendations for systems, implementing plans and maintaining RDM processes

Resource Profile (2)

- Personnel
  - People involved in the data management activities
  - Quantity and overall quality
- Budget
  - Quantity
  - Availability
- Timeframe
- Technology Environment

Plan Components

- Desired result of planning is:
  - Data management activity overview
  - Understanding of the resource profile
  - Documenting the trade-offs between alternatives
  - Documentation guidelines for each phase
  - Identified methods for preventing, detecting, and resolving potential problems

Design Considerations

- Alternative processes (and monitoring)
- Schedule (time and dependencies)
- Costs (estimated)
- Document the recommended processes as well as an alternative and consider their pros and cons, trade-offs
- Write final plan for approval

General Phases of Data Management

- Data collection and verification
- Data Inventory
- Data validation and cleaning
- Data transformation and analysis
- Reporting
- Process monitoring

Implementation

- Standards can help a lot with:
  - Documentation
  - Naming entities
  - Programming
  - Backup and Archiving

Implementation (2)

- Monitoring all data management processes from the start of the system is important for early detection of problems
- Address data quality
- Expect to find issues.

Implementation (3)

- Documentation is the heart of the system
- Information needed for effective management
- Best way to demonstrate that a finished project was executed correctly and that the results were based on valid data

Implementation (4)

- Staff training promotes
  - Understanding of standards and documentation needs
  - Communication
  - Clarification of roles and responsibilities
  - Understanding of technical aspects of activities
  - Understanding of global vs immediate priorities
- Leadership
  - Successful implementation needs the full support of everyone directly or indirectly involved

System Maintenance

- Maintenance introduction
- Importance of monitoring
- Data quality
- Documentation updates
- Data security

Maintenance Introduction

- Purpose – continued operation of system
- Enforce standards and procedures
- Schedule trainings and orientations
- Enhance teamwork
- Avoid rapid or unexplained changes
- Maintain a continuous record of documentation
- Keep everyone informed of changes

Importance of Monitoring

- Monitoring is essential to early discovery of problems
- Monitor processes to verify system operation
  - Simple and practical
- Look for:
  - Error rates
  - Lost records
  - Inactive processes (informal workaround?)

Data Quality

- Five aspects
  - Integrity
  - Accuracy
  - Precision
  - Validity
  - Usability

Data Qualities - Integrity

- Does the stored data represent the original reported value?
  - At risk from transitions
    - Physical or conceptual
  - Compare subset of records to their collection values

Data Quality - Accuracy

- How well does the stored data represent reality?
- Can be difficult to monitor directly
- Other quality components often are used as surrogates

Data Quality - Precision

- Can your process record the same value on repeated evaluation of the same item?
- Measurement limited to reproducible processes

Data Quality - Validity

- “Fitness for use”
- Qualitative measure of the synthesis of known information about a data representation
  - Combination of integrity, accuracy and precision
- Tools
  - Logic checks
  - Check digits or checksums
  - Range checks
  - Consistency checks (internal and external)

Data Quality - Usability

- Is the data in a format that the researcher can use?
- Valid, but unusable data doesn’t count in terms of getting results
- Use standard file formats

System Maintenance

- Importance of monitoring
- Data Quality
- Documentation updates
- Data security

Documentation Updates

- Poorly maintained documentation keeps data from being considered high quality
- Four types of system data need to be updated to maintain quality
  - Process descriptions (changes to plan, who, what, when, how and why)
  - Data descriptions (data always changes)
  - Program descriptions (external)
  - Research Report descriptions (reproducibility)

System Maintenance

- Importance of monitoring
- Data Quality
- Documentation updates
- Data security

Data Security

- Expect failure
- Backups (data, programs, related files)
  - Maintain two sets of onsite backups and one offsite
  - Test restoration before you really need it
- Backups (personnel)
  - Cross-train personnel on data management activities
  - Skill redundancy helps protect the system from disruption due to sickness or turnover
- Archiving
  - Organized storage of materials that are no longer needed

Data Security (2)

- Disaster Planning (fire, flood, etc)
  - Minimizes decision making and confusion
- Considerations
  - Do you need to relocate?
  - What materials are essential to restoring your system?
  - Where are you going to get them?
  - In what order should they be replaced?
  - Identify key people

Statistical Software

- Different software has different strengths
  - Have an analysis plan
- Consider using more than one program to enhance your capabilities
  - Analytic
  - Data management

Statistical Software Choices

- ArcGIS
- EpiInfo
- R
- SAS
- SPSS
- Stata
- SUDAAN

Considerations for Choosing Statistical Software

- Cost
  - Licenses, Availability, Maintenance and support
- Learning Curve
  - Training, “Online” help, communities of practice, time
- Expertise (Statistical capabilities)
  - Basic vs advanced
  - Is the software itself limited?

Comparison of Software

	Budget	Expertise	Training	Complex Sample Surveys
ArcGIS	H	H	H	-
EpiInfo	☺	L	M	
Minitab	M	M	L	
R	☺	H	H	
SAS	L/Free	H	H	Y
SPSS	H*	M	L	Y*
Stata	L	M	L	Y
SUDAAN	M**	H*	H	Y

Scenario 1: Adequate Budget, but No Expertise or Time for Training

	Budget	Expertise	Training	Complex Sample Surveys
EpiInfo	☺	L	M	
Minitab	M	M	L	
SAS	M/Free	H	H	Y
SPSS	H*	M	L	Y*
Stata	L	M	L	Y
SUDAAN	M**	H	H	Y

Scenario 2: Expertise Only

	Budget	Expertise	Training	Complex Sample Surveys Surveys
EpiInfo	☺	L	M	
Minitab	M	M	L	
SAS	M/Free	H	H	Y
SPSS	H*	M	L	Y*
Stata	L	M	L	Y
SUDAAN	M**	H	H	

Scenario 3: Learning Time Only

	Budget	Expertise	Training	Complex Sample Surveys Surveys
EpiInfo	☺	L	M	
Minitab	M	M	L	
SAS	M/Free	H	H	Y
SPSS	H	M	L	Y*
Stata	L	M	L	Y
SUDAAN	M**	H*	H	

Summary

- To achieve your scientific and programmatic goals a systematic approach to data management can help maximize your resources and minimize the risks
- Good, up-to-date documentation helps demonstrate that your analyses were executed correctly and were based on valid data
- Data management systems should be practical and simple
- One size does not fit alltools to ensure breadth of techniques and a minimum of effort

References

- Calvert WS, Ma JM. Concepts and Case Studies in Data Management. Cary: SAS Institute Inc; 1996.
- Mitchell MN. Strategically using General Purpose Statistics Packages: A Look at Stata, SAS and SPSS. Los Angeles (CA): UCLA Academic Technology Services, Statistical Consulting Group; 2005 Jan. Technical Report Series, Report Number 1, Version Number 1.