

Good Data Practices

- Series Recap
 - Session 1: Early Data Planning for Research

Session 2: Managing and Documenting Data Workflow

- Getting started
- Importance of documentation
- Data management workflow
- Analysis workflow

Poll Question #1

- What would you say is your level of research experience?
 - 1 (Novice)
 - 2
 - 3
 - 4
 - 5 (Expert)

Session 2: Outline

- **Getting started**
- Importance of documentation
- Data management workflow
- Analysis workflow

Getting started

- Formalizing your data management plan
 - Description of the project
 - Description of the data to be collected
 - Standards to be applied for formats, metadata, etc.
 - Plans for short-term storage and data management: e.g., file formats, local storage and back up procedures, and security
 - Description of legal and ethical issues: e.g., intellectual property, confidentiality of study participants
 - Access policies and provisions: i.e., how will you make data available to others, any restrictions to data reuse, etc.
 - Provisions for long-term archiving and preservation
 - Assigned data management responsibilities: i.e., which persons will actually be responsible for ensuring data management; compliance monitoring over time

Poll Question #2

- How much experience have you had developing a data management plan?
 - 1 (None)
 - 2
 - 3
 - 4
 - 5 (Extensive)

Data Management for the Sciences

Tags: data curation, data management plan tool, data repository, dataup, merriitt, nsf data management

A guide to best practices for management of research data, including links to data services from the University of California.

Last Updated: Aug 9, 2013

URL: <http://guides.library.ucla.edu/data-management>

[Print Guide](#)

[RSS Updates](#)

[SHARE](#)

[Overview](#)

[Managing Data](#)

[Creating a Data Management Plan](#)

[Funding Agency Requirements](#)

[Data Deposit and Sharing](#)

[Resources](#)

Funding Agency Requirements

[Comments\(0\)](#)

[Print Page](#)

Search:

This Guide

[Search](#)

NSF

The NSF strengthened its data sharing policy in January 18, 2011, when it began requiring all grant proposals to include a two-page data management plan. Guidelines are available [online](#). Specific NSF directorates, offices, divisions, programs, or other units may impose [additional data management requirements](#).

NSF Resources



DMPTool

You can use the **Data Management Plan (DMP)** tool created by the California Digital Library (CDL) to create a data management plan that will satisfy NSF-directorate specific requirements. Read more about the tool both at CDL and on our [Creating](#)

Links to Funding Agencies Guidelines



- [National Science Foundation: Dissemination and Sharing of Research Results](#)
- [National Institutes of Health: Data Sharing Policy](#)
- [Centers for Disease Control and Prevention Policy on Releasing and Sharing Data](#)
- [Department of Defense Principles and Operational Parameters of the DoD Scientific and Technical Information Program](#)
- [Environmental Protection Agency Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated](#)
- [NASA Earth Science Statement on Data & Information Policy](#)
- [National Institute of Standards and Technology \(NIST\) Guidelines, Information Quality Standards, and Administration Mechanism](#)
- [United States Department of Agriculture USDA Cooperative State Research, Education, and Service \(CSREES\)](#)
- [National Oceanic and Atmospheric Administration \(NOAA\) Data Submission Policies and Guidelines](#)
- [National Endowment for the Humanities \(NEH\): Data Management Guidelines](#)
- [Institute of Museum and Library Services \(IMLS\): Specifications for Projects that Develop Digital Products](#)
- [The Gordon and Betty Moore Foundation: Data Sharing and Plan](#)

NIH

The [NIH Public Access Policy](#) ensures that the public has access to the published results of NIH funded research. It requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive [PubMed Central](#) upon *acceptance for publication*. To help advance science and improve human health, the Policy requires that these papers are accessible to the public on PubMed Central no later than 12 months after publication.

For more information on how to comply with the NIH Public Access policy, please see the [NIH Public Access Policy research guide](#).

[NIH Public Access Resources](#)

UCLA Library Website

<http://guides.library.ucla.edu/content.php?pid=385860&sid=3182780&preview=200eac9a6bb823f4503d8413245dae7>

Data Management Plan: Sample Plan Created at the DataONE Best Practices Workshop - Santa Fe NM 7/2011 Atmospheric CO2 Concentrations, Mauna Loa Observatory, Hawaii, 2011-2013

Example of DMP Plan produced by the DMP Tool

1. Types of data produced

Air samples at Mauna Loa Observatory will be collected continuously from air intakes located at compass quadrants. Raw data files will contain continuously measured references standards, daily check standards, and blanks. The sample lines located at cor influence of source effects associated with wind directions. In addition to the CO2 data, w direction, temperature, humidity, precipitation, and cloud cover). Site conditions at Mauna retained. The final data product will consist of 5-minute, 15-minute, hourly, daily, and mont

2. Data and metadata standards

Metadata will be comprised of two formats: contextual information about the data in a text based document and ISO 19115 standard metadata in an xml file. These two formats for metadata were chosen to provide a full explanation of the data (text format) and to ensure compatibility with international standards (xml format). The standard XML file will be more complete; the document file will be a human readable summary of the XML file.

3. Policies for access and sharing

The final data product will be release to the public as soon as the recalibration of standard gasses has been completed and the data have been prepared, typically within six months of collection. There is no period of exclusive use by the data collectors. Users can access documentation and final monthly CO2 data files via the Scripps CO2 Program website (<http://scrippsco2.ucsd.edu>). The data will be made available via ftp download from the Scripps Institution of Oceanography Computer Center. Raw data (continuous concentration measurements, weather data, etc.) will be maintained on an internally accessible server and made available on request at no charge to the user.

Generated by the DMPTool

dmp.cdlib.org

10/06/11 12:00 AM

4. Policies for re-use, redistribution

Access to databases and associated software tools generated under the project will be available for educational, research and non-profit purposes. Such access will be provided using web-based applications, as appropriate.



2. Data and metadata standards

Metadata will be comprised of two formats; contextual information about the data in a text based document and ISO 19115 standard metadata in an xml file. These two formats for metadata were chosen to provide a full explanation of the data (text format) and to ensure compatibility with international standards (xml format). The standard XML file will be more complete; the document file will be a human readable summary of the XML file.

3. Policies for access and sharing

The final data product will be release to the public as soon as the recalibration of standard gasses has been completed and the data have been prepared, typically within six months of collection. There is no period of exclusive use by the data collectors. Users can access documentation and final monthly CO2 data files via the Scripps CO2 Program website (<http://scrippsco2.ucsd.edu>). The data will be made available via ftp download from the Scripps Institution of Oceanography Computer Center. Raw data (continuous concentration measurements, weather data, etc.) will be maintained on an internally accessible server and made available on request at no charge to the user.

4. Policies for re-use, redistribution

Access to databases and associated software tools generated under the project will be available for educational, research and non-profit purposes. Such access will be provided using web-based applications, as appropriate.

Data Management for the Sciences

Tags: data curation, data management plan tool, data repository, dataup, merritt, nsf data management

A guide to best practices for management of research data, including links to data services from the University of California.

Last Updated: Aug 9, 2013

URL: <http://guides.library.ucla.edu/data-management>
 Print Guide

 RSS Updates

 SHARE
 



Overview

Managing Data

Creating a Data Management Plan

Funding Agency Requirements

Data Deposit and Sharing

Resources

Creating a Data Management Plan

 Comments(0)

 Print Page
Search:

This Guide

 Search

Overview

In 2011, the [National Science Foundation \(NSF\)](#) began requiring that grant applicants include a data management plan in their proposals. The [National Institutes of Health \(NIH\)](#), [National Endowment for the Humanities \(NEH\)](#), the [Gordon and Betty Moore Foundation](#) and others have similar policies in effect.

By creating a data management plan, you will not only satisfy funding agencies but also have an opportunity to think through how to manage your data for your own use as well as any future use by fellow researchers.

Common Requirements for a DMP

Although funding institutions have different specific requirements, a data management plan should generally contain the following components:

- Description of the project: e.g., purpose of the research, organization(s) and staff involved
- Description of the data to be collected: e.g., the nature and format of the data, how it will be collected, and overview of secondary data available on the topic
- Standards to be applied for formats, metadata, etc.
- Plans for short-term storage and data management: e.g., file formats, local storage and back up procedures, and security
- Description of legal and ethical issues: e.g., intellectual property, confidentiality of study participants
- Access policies and provisions: i.e., how will you make data

Data Management Plan Tool

Use the Data Management Plan (DMP) Tool to create ready-to-use data management plans for specific funding agencies.



DMPTool

Online data management planning tool to help guide researchers through the process of creating a data management plan.

The list of funding agencies for which the DMP Tool can create customized plans include

- [National Science Foundation \(NSF\)](#): plans for different divisions such as Biological Sciences, Chemistry, Engineering, Physics are included. An NSF Generic plan is also offered.
- [National Institutes of Health \(NIH\)](#)
- [National Oceanic and Atmospheric Administration \(NOAA\)](#)
- [Institute of Museum and Library Services \(IMLS\)](#)
- [National Endowment for the Humanities \(NEH\)](#)
- [The Gordon and Betty Moore Foundation](#)

The DMP Tool saves you time by recognizing the points that need to be addressed in a data management plan for a particular funding agency. It will prompt you to answer questions to satisfy these requirements and then compile your answers into a formatted data management plan.

Getting started

- Defining roles and responsibilities
 - Assigning project team responsibilities
 - File management (project & data)
 - Data management
 - Modeling & analysis

Poll Question #3

- What is your primary research role?
 - Investigator
 - Data analyst/programmer or statistician
 - Research coordinator or assistant
 - Student, trainee, or fellow

Session 2: Outline

- Getting started
- **Importance of documentation**
- Data management workflow
- Analysis workflow

Importance of documentation

- **What** should be documented?
 - Data management methods
 - Data analysis methods
 - What you decided to do & why
 - What data were created & how



Data Documentation Initiative

- What is DDI?
- DDI Alliance
- DDI At Work
- Resources
- Specification
- RDF Vocabularies
- Controlled Vocabularies
- Community

Last Updated: Thu, 2011-12-08 06:01 — Sam Spencer

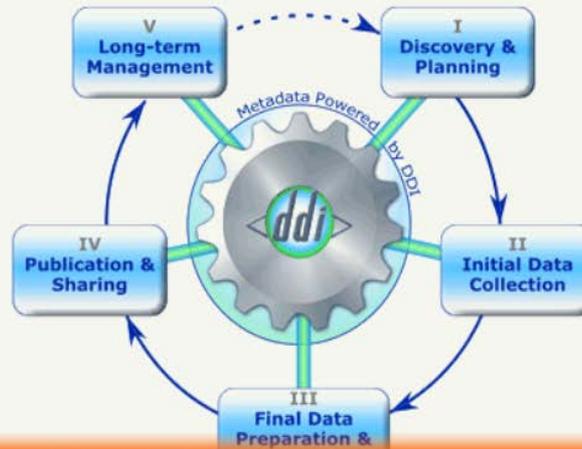
Welcome to the Data Documentation Initiative

A metadata specification for the social and behavioral sciences

Use DDI to:

- Document your data across the life cycle
- Interoperate with others
- Do Data Intelligently (DDI)!

Find out how others have put [DDI to work](#) in their organizations, explore [resources](#) for learning more about and using the DDI, or join the [DDI Community](#).



Dublin Core® Metadata Initiative

Making it easier to find information.

Specifications and documentation

- Home
- Metadata Basics
- DCMI Specifications
- Community and Events
- Join/Support
- About Us

DDI Lifecycle - Latest version: 3



DCMI Upcoming Events



DC-2013 in Lisbon on 2-6 September 2013 will explore questions regarding the persistence, maintenance, and preservation of metadata and descriptive vocabularies. The need for stable representations and descriptions spans all sectors including cultural heritage and scientific data.

Importance of documentation

- Key components to document
 - Sample and sampling procedures
 - Weighting
 - Date and geographic location of data collection, and time period covered
 - Data source(s)
 - Unit(s) of analysis/observation

Importance of documentation

- What to document about variables
 1. The exact question wording or exact meaning of the datum
 2. The text of the question integrated into the variable text
 3. Universe information, i.e., who was actually asked the question
 4. Exact meaning of codes
 5. Missing data codes
 6. Unweighted frequency distribution or summary statistics
 7. Imputation and editing information
 8. Details on constructed and weight variables
 9. Location in the data file
 10. Variable groupings

Importance of documentation

- Summary documents to maintain
 - Technical information about data files
 - Data collection instruments
 - Flowchart of the data collection instrument
 - Index or table of contents
 - List of abbreviations and other conventions
 - Interviewer guide
 - Recode logic
 - Coding instrument

Importance of documentation

- **Where to document?**
 - Master document
 - “Living protocol”
 - Codebook
 - Supplemental documents in organized folders
 - Program code
 - Output & log files
 - Many good systems - select one
 - Be consistent

Importance of documentation

- **When to document?**
 - While defining your cohort
 - During every team meeting
 - During data management & analysis
 - Inside every program
 - In your master document

- **Document as you go...**

Session 2: Outline

- Getting started
- Importance of documentation
- **Data management workflow**
- Analysis workflow

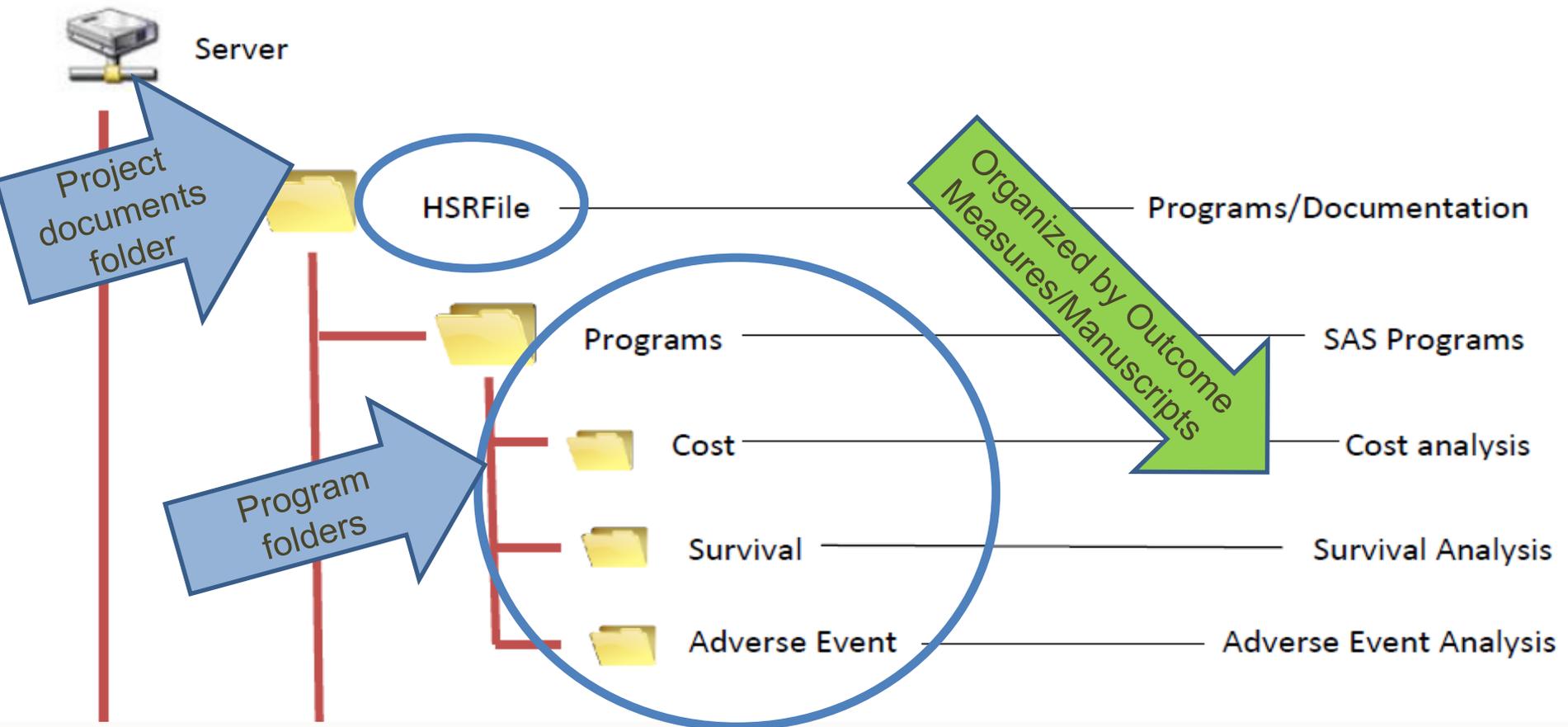
Data management workflow

- File naming standards and organization
- Documenting cohort derivation
- Primary data collection process
- Accessing secondary data in the VA
- Linking data from multiple sources
- Cleaning data and documenting data issues
- Documenting the analytic dataset
- Programing walkthroughs

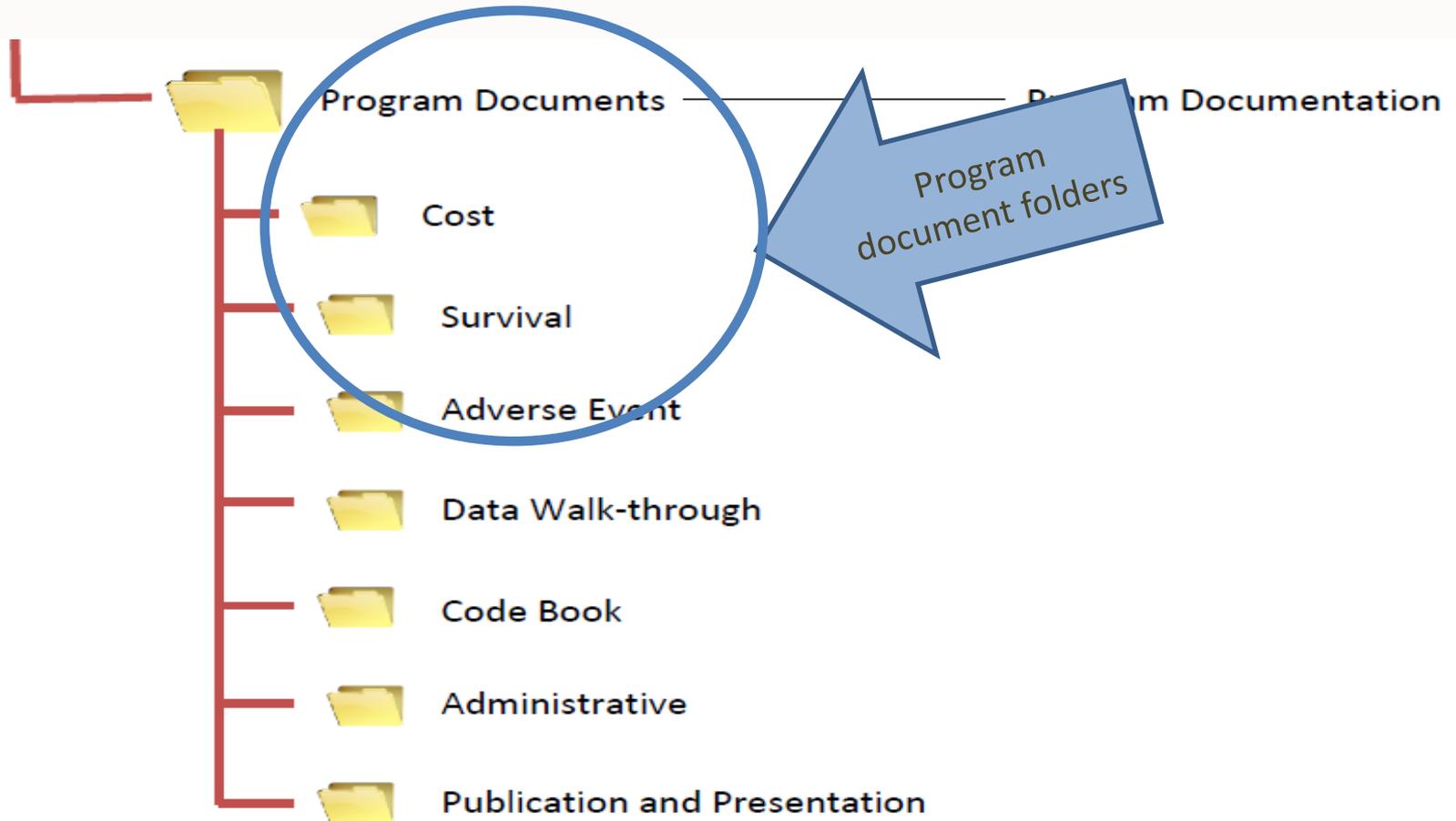
Data management workflow

- File naming standards and organization
 - Concepts for naming files
 - Easily accessible
 - Accurate & clear
 - Meaningful to others
 - Easily distinguishable
 - Recognizable in different environments
 - Consistent

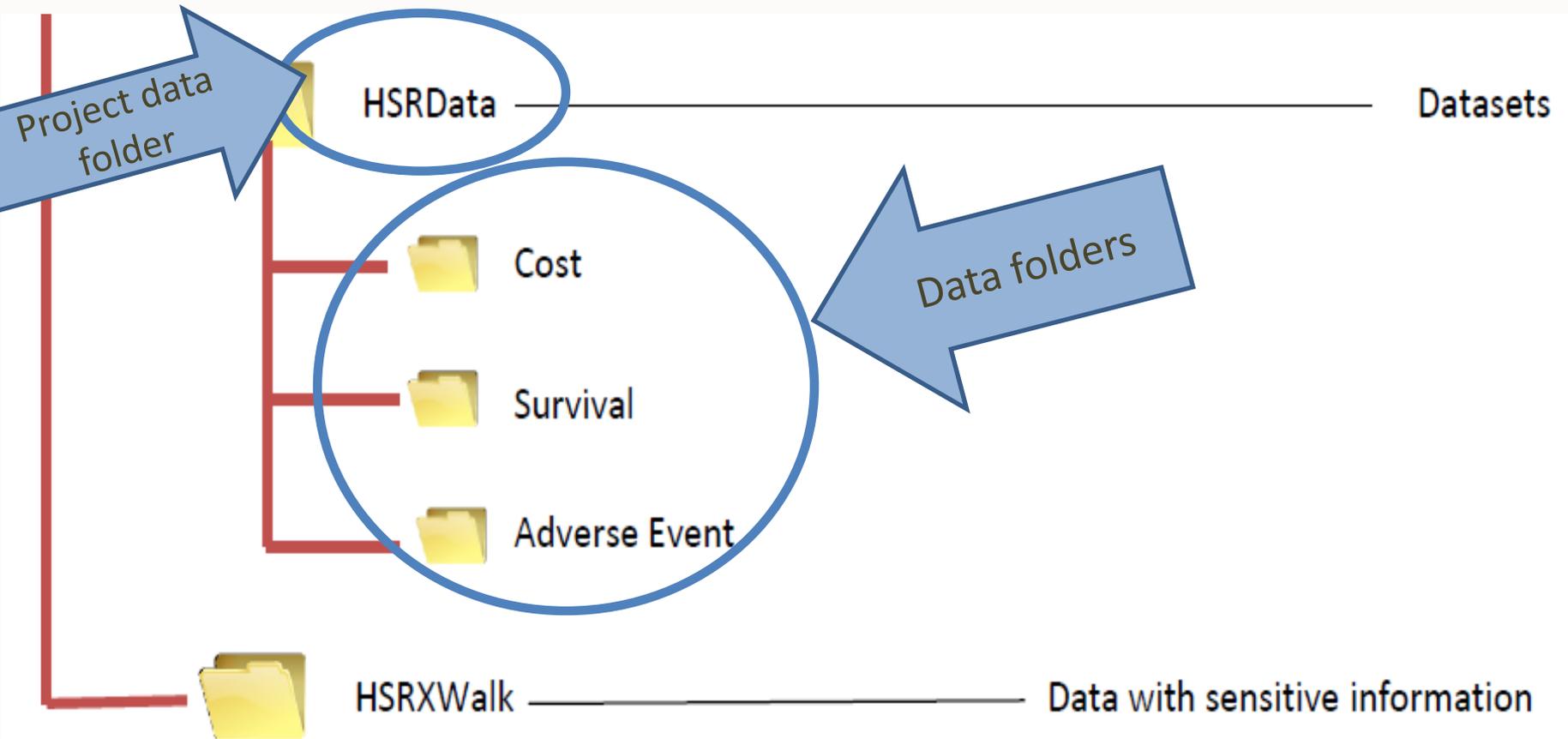
Data management workflow



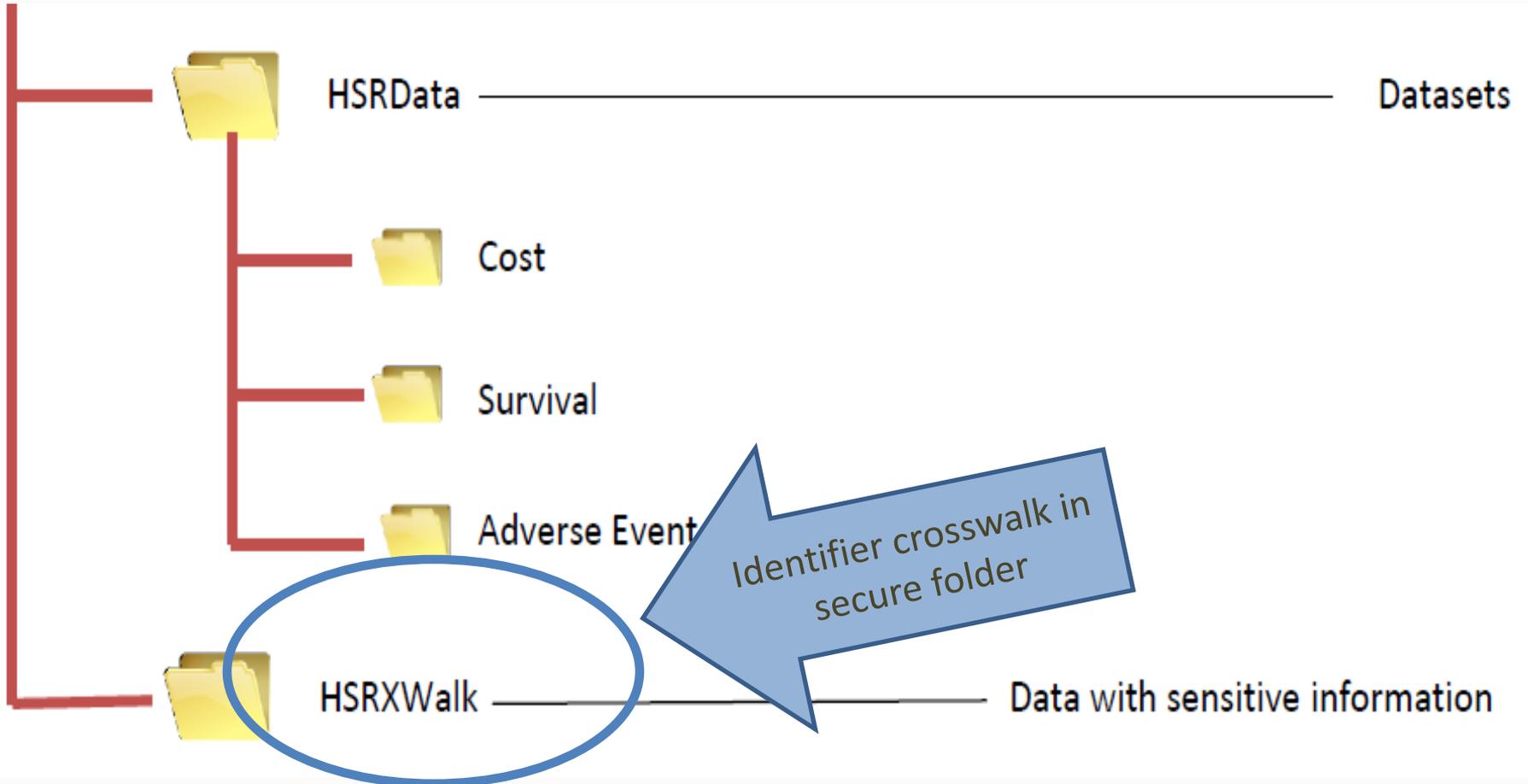
Data management workflow



Data management workflow



Data management workflow



Data management workflow

- Documenting cohort derivation
 - Document the final cohort definition decisions
 - Inclusion & exclusion criteria
 - All sources used
 - Rationale
 - Document as you go
 - Diagrams are critical and generally required for publications

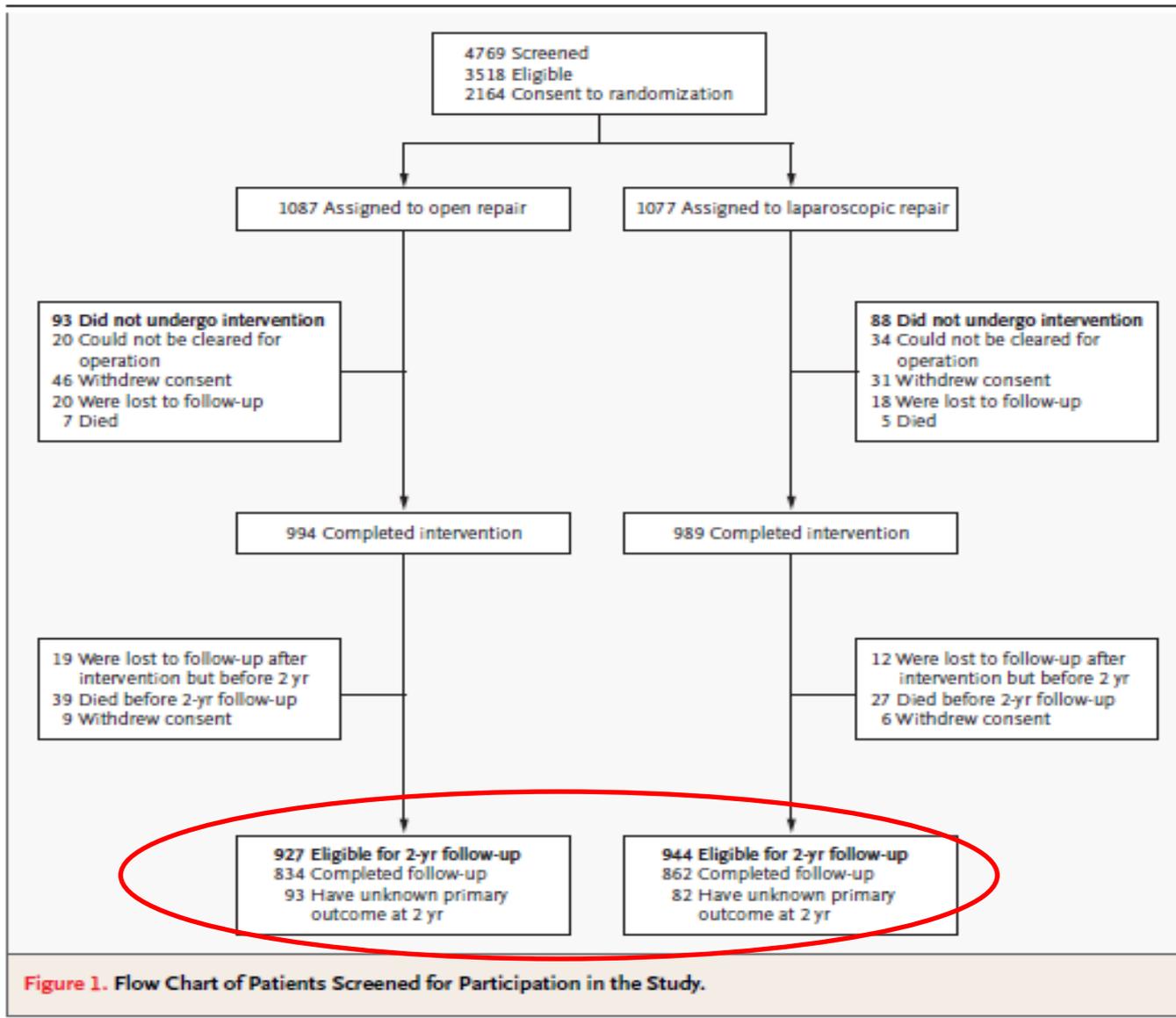


Figure 1. Flow Chart of Patients Screened for Participation in the Study.

RCT Cohort Flowchart Example

Source: Open Mesh versus Laparoscopic Mesh Repair of Inguinal Hernia, VA CSP Project # 456; PI: L. Neumayer

Abbreviations and Acronyms

- HUI2 = Health Utilities Index 2
- ICER = incremental cost effectiveness ratio
- LAP = laparoscopic hernia repair
- OPEN = open hernia repair
- QALY = quality adjusted life year
- VA = Department of Veterans Affairs

Cost effectiveness is an important consideration when evidence for predominance of one surgical technique is lacking. Earlier randomized clinical trials studies reported higher operating room costs for laparoscopic compared with open repairs.⁴⁻¹¹ Some of these studies, however, lacked specific cost data or cost effectiveness measures needed to evaluate relative benefits and costs, and none has followed patients beyond 1 year¹² or taken into account any baseline differences between study groups that may have affected cost effectiveness.¹³ Shorter operation times and greater use of outpatient

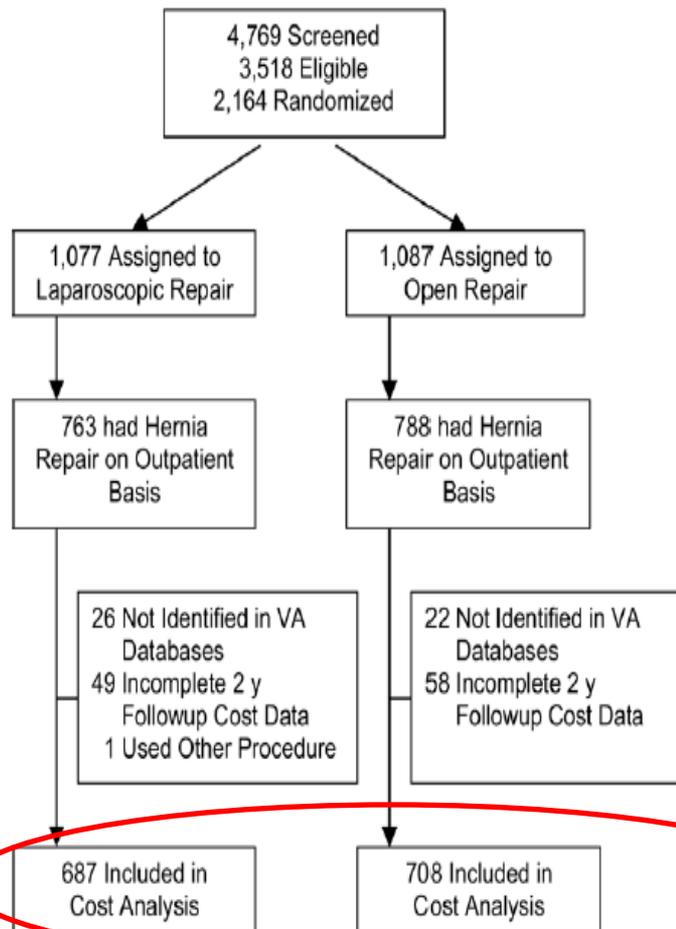
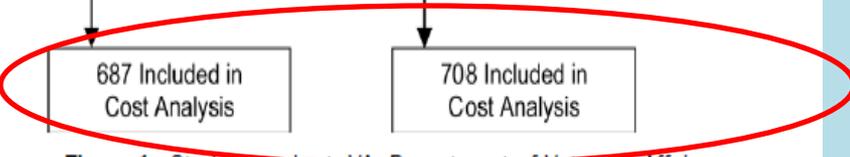


Figure 1. Study flow chart. VA, Department of Veterans Affairs.

RCT Cohort
Flowchart
Example

Source: Open Mesh
versus Laparoscopic
Mesh Repair of
Inguinal Hernia, VA
CSP Project # 456;
PI: L. Neumayer



care and chronic disease management performance measures.^{9,10} Although there has been increased focus on the quality of colon cancer screening and recent efforts to improve treatment,¹⁰ there has been limited work reported about the quality of colon cancer care provided to veterans.

To address these information gaps, we conducted a retrospective cohort study of elderly veterans in California who were diagnosed with colon cancer within and outside of the VA system. We focused on the 6 months after diagnosis to describe colon cancer treatment patterns and to identify factors associated with surgery and adjuvant chemotherapy.

METHODS

Study Design

We assembled a retrospective cohort of elderly veterans diagnosed with stage I to IV colon cancer between 1999 and 2001 and examined their health care use and outcomes over the subsequent 3-year period using VA, Medicare, and National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) California cancer registry data. Our protocol was approved by the Edward Hines, Jr. VA Hospital Institutional Review Board; informed consent was waived. Our study cohort (Fig 1) was identified from a sampling frame of all VA eligible patients in 1999 to 2001 and described elsewhere.^{5,11}

Briefly, the sampling frame comprised all elderly veterans who were known to the VA (ie, they had either used VA health care services, had enrolled for VA health care, or had received compensation or pension benefits from the VA)^{5,11} and were enrolled in Medicare from 1999 to 2001. We matched the California cancer registry data to our sampling frame using a deterministic matching procedure and applied specific criteria for stage I to IV colon cancer diagnosed between July 1, 1999 and December 31, 2001 and age ≥ 66 years at the time of diagnosis, yielding 1,179 veterans. (Details of the matching procedure are available on request from the authors.)

Furthermore, we excluded individuals for whom we had no or incomplete health care utilization data as a result of Medicare health maintenance organization enrollment, non-Medicare primary payer coverage, Part B Medicare coverage only, VA eligibility starting after the beginning of the accrual period, autopsy-only diagnosis, or other unknown reasons. Because the California VA hospitals reported all their new cancer cases to the California registry as well as the VA Central Cancer Registry, we were able to identify patients who were diagnosed and initially treated in VA and non-VA hospitals using the California registry and then validate them with matching to the VA Central Cancer Registry. The resulting analytic cohort comprised 166 VA patients and

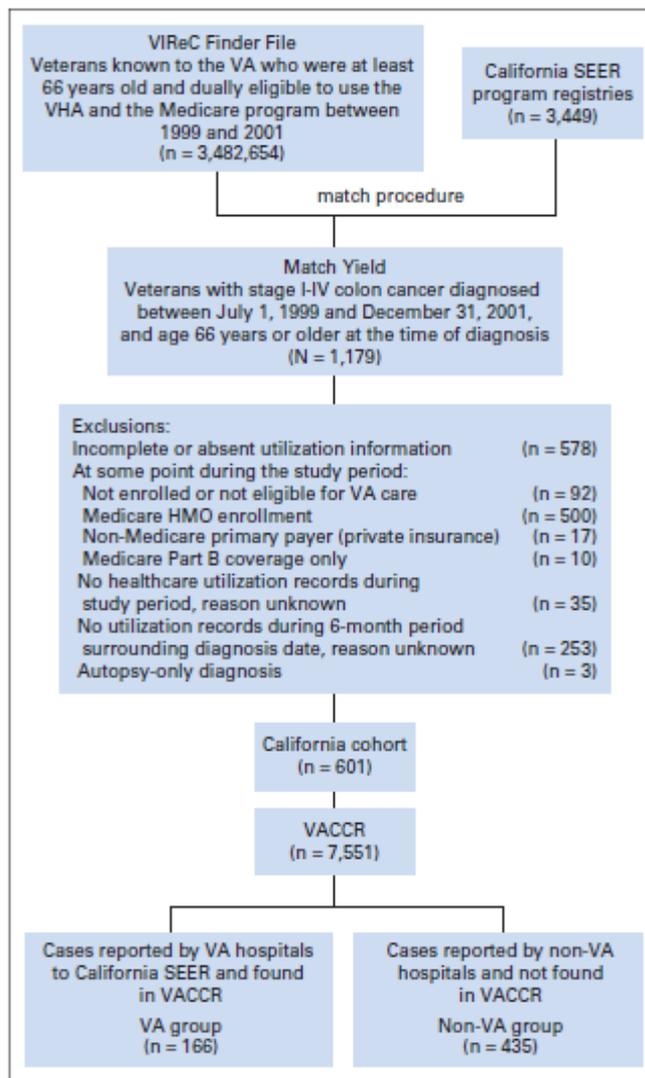


Fig 1. California colon cancer study cohort Department of Veterans Affairs (VA) Central Cancer Registry (VACCRC). VIREC, Veterans Affairs Information Resource Center; VHA, Veterans Health Administration; SEER, Surveillance, Epidemiology, and End Results; HMO, health maintenance organization.

Observational Study Flowchart Example

Source: Quality and Costs
of Colon Cancer in VA and
Medicare; VA HSR&D
Service Project # IIR 03-196

Clinical Guidelines for ESA use in Cancer Impacts on Clinical

Walkthrough
document for creation
of final cohort for ESA
study

**Hynes study,
Clinical Guidelines
for ESA use in
Cancer
Impacts on
Clinical Practice**

Create All Cancer Cohort

Program name: ALL_CANCER_COHORT.SAS

Analyst: Lucy Zhang

Created date: 10/25/2010

Program description: The program will create cohort for all incident cancer cases.

File Location: HSRFILES:\ESA_CANCER\DATA ANALYSIS\PGMS
ESA_USE\VA\ALL_CANCERS\SAS

Date Modified: 04/07/2011

Reason for Modification:

- Lung and Colon cancer group were updated based on the [ICO-3 code book](#).
- Some cancer diagnosis sites were grouped. [ICO-3 code book](#).

Input data set(s) / source(s):

SAS Data Set Name	Description	Total Records	Total Variables
PREV_INCIT_ ACT_VA_FLAGS	Flags of indicating prevalent case, incident cases and active VA user in study month -12 to 0 or -24 to -12	374,605	9
VACCR_SSN_FOUND_DRO P_DIED	Individuals died prior Jan. 1, 2002 were dropped from selected 2000-2009 VACCR cases.	383,450	138
CHEMO	Chemo records for incident cases	7,591,014	13
RADIATION	Any record contains radiation therapy codes	10,882,922	8

Data management workflow

- Primary data collection
 - Finalize forms
 - Authorization language
 - Manage incoming data
 - Check the cohort

Data management workflow

- Finalizing forms
 - Data collection form finalized
 - Final form & language
 - Administration method
- Patient HIPAA authorization language
 - Include re-use for further research?
 - If needed, changes go to IRB for approval
 - Get authorizations signed after revision

- Document as you go

Data management workflow

- Managing incoming data
 - How will the data be received?
 - How will data quality be maintained
 - What data need to be integrated?
 - How will you integrate/link the data?
- Software tools to capture, link primary & secondary data?
 - Access
 - SAS
 - SQL Server Management Studio
 - REDCap
- Establish routines to document as you go

Data management workflow

- Accessing secondary data in the VA
 - What do you need to prepare?
 - Research document packet
 - Data source-specific request forms
 - If real SSNs requested, justification showing why they are essential

Data management workflow

- Factors determining data access process
 - Authorization
 - Source (what data)
 - Level of access (national, VISN, local)
 - Identifier (real or scrambled SSN, IENs, etc.)
 - Data source location (Mainframe, CDW, PBM, PCS, VIREC etc.)
 - Data format (SAS, SQL)

- **Note:** VIREC Database and Methods series, Research Access to Data November 4, 2013

Data management workflow

- Data management activities documented
 - Decisions made in team meetings (minutes, action items, etc.)
 - Variables pulled from each data source
 - Data linkage methods
 - Variables derived - definition & process
 - Programs developed & revised
 - Add to master document as you go

Data management workflow

- Data quality
 - Examine the dataset
 - Identify common data quality issues
 - Missing data
 - Differences in data type
 - Undocumented and out-of-range code values
 - Identify other/systematic data issues
 - Critical data elements
 - Document the process, findings and response as you go...

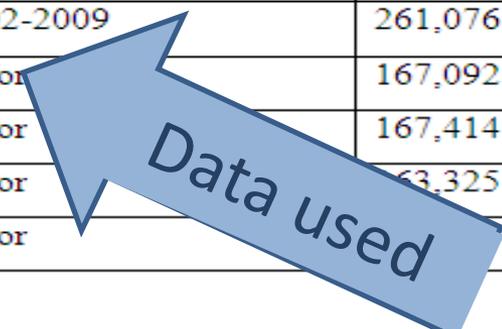
Data management workflow

- Checklist for documenting analysis file preparation
 - Data sources
 - Source variables
 - Linkage process
 - Derived variables
 - Program code – including algorithms
 - Algorithms described in standard English
 - A codebook for every dataset

Data management workflow

Input data set(s) / source(s):

SAS Data Set Name	Description	Total Records	Total Variables
COHORT	Cancer cases diagnosed in 2002-2009	261,076	121
DENOM02	CY2002 Medicare Denominator	167,092	46
DENOM03	CY2003 Medicare Denominator	167,414	46
DENOM04	CY2004 Medicare Denominator	163,325	46
DENOM05	CY2005 Medicare Denominator		45



Clinical Guidelines for ESA use in Cancer Impacts on Clinical Practice

DENOM06	CY2006 Medicare Denominator	156,224	45
DENOM07	CY2007 Medicare Denominator	169,099	49

Clinical Guidelines for ESA use in Cancer Impacts on Clinical Practice

2008 VACCR Data Received

Date Received: 04/28/2010

File Name: HYNES08.DAT

File Saved in: HSRXWALK:\ESA_CancerXwalk

Data Document Saved in: HSRFILES:\ESA_Cancer\DATA ANALYSIS\PGMS
DOCUMENT\VACCR DATA

Data Description: In addition to the cancer cases reported to VA Central Cancer Registry who diagnosed from 2000-2007, we received cancer cases reported to VACCR in 2008. The received 2008 data contained 45,225 records and 89 variables.

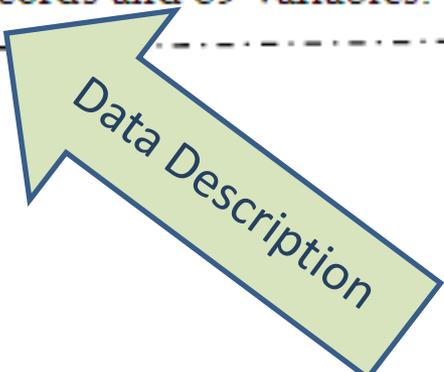
Received ASCII file was transfer to SAS by using DBMSCopy.

SAS Data Set Name: VACCR_RECEIVED08.SAS7BDAT

Saved in: HSRXWALK:\ESA_CancerXwalk

Variables Description in VACCR_RECEIVED data set:

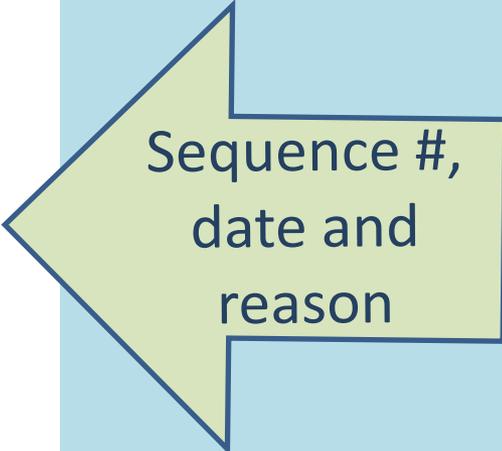
Number of Observation: 45,225
Number of Variables: 89



Data Description

Modification of program:

Sequence #	Date Modified:	Reason Modified:
001	06/24/2009	Create INDEX_SUG_VENUE_REVERSE variable
002	09/09/2009	Create VA_REPT2 variable
003	09/11/2009	Create new variables for the mortality analysis DX_TO_DEATH_DAY_1YEAR DIED_IN_2YEAR_DX DX_TO_DEATH_DAY_2YEAR DIED_IN_3YEAR_DX DX_TO_DEATH_DAY_3YEAR
004	09/14/2009	Rename DIED and DX_TO_DEATH_DAY DIED=DIED9904 DX_TO_DEATH_DAY=DX_TO_DEATH_DAY9904
005	09/14/2009	Create DIED and DX_TO_DEATH_DT
006	03/06/2012	Create chemo within 9 month and radiation within 9 month flags.



Sequence #,
date and
reason

Details of Modification:

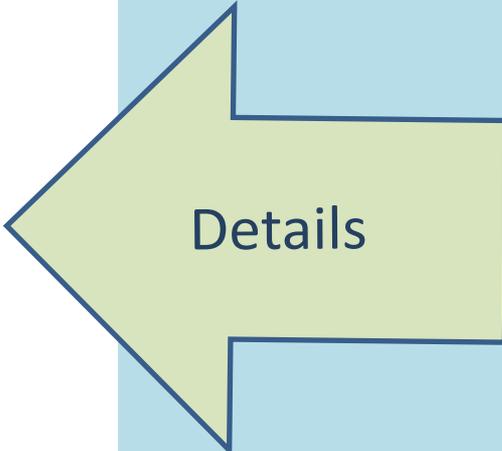
Sequence # 001:

For surgery within in month 0-6 SEER only case, use venue of inpatient stay as surgery

venue if:

1. There are cost within month -1 to 1 of index surgery date
and
2. There are inpatient stay within month -1 to 1 of index surgery date and had colon cancer diagnosis code.

6 SEER surgery only cases qualified.



Details

Data management workflow

New variables created:

Variable Name	Description
HMO1-HMO96	The flag indicate if the patients enrolled in HMO in each month from year 2002-2009.

Page 4 of 5

Created on: 11/15/2011

Clinical Guidelines for ESA use in Cancer Impacts on Clinical Practice

	Values: 1 = 'Yes' 0='No'
PARTB1-PARB96	The flag indicate if the <u>patients has</u> Part B coverage only in each month from year 2002 to 2009. Values: 1 = 'Yes' 0='No'
PAYER1-PAYER96	The flag indicate if the had other primary payer in each month from year 2002 to 2009.

Clinical Guidelines for ESA Use in Cancer Impacts on Clinical Practice

SAS Program Walk-through Summary

Seq Num	Date	Program Name	Description	Note	
Volume II:					
5	28	2/14/2011	CHEMO.SAS	Get Chemo records for incident cases	✓
6	29	2/14/2011	RADIATION.SAS	Get Radiation records for incident cases	✓
7	30	2/15/2011	EPO.SAS	Get ESA records for incident cases Program modified on 05/18/2011 and Updated program and document attached	✓
8	31	2/15/2011	HEMOGLOBIN_LAB.SAS	Get Hgb records from DSS LAB data for incident cases Program modified on 05/18/2011 and Updated program and document attached	✓
9	32	2/15/2011	HEMOGLOBIN_LAR.SAS	Get Hgb records from DSS LAR data for incident cases Program modified on 05/18/2011 and Updated program and document attached	✓
10	33	2/23/2011	ALL_CANCER_COHORT	Create all Cancer Cohort	✓
11	33.1	9/7/2011	DOD_FOR_PRIOR_ON_DX_DEATH.SAS	Assign death date for the cases died prior or on the day of diagnosis	✓
12	34	2/25/2011	BLOOD_TRANSFUSION_AND_ANEMIA.SAS	Get blood transfusion and anemia records for incident cases	
13	35	2/25/2011	LUNG_COLON_COHORT	Create Lung and Colon Cohort	
14	35.1	2/28/2011	COHORT_CHEMO.SAS	Create cohort for pre/post paper	
15	36	3/1/2011	TRIGGER_HGB_COHORT	Create Trigger Analysis Cohort	
16	37	3/9/2011	HGB_TRENDS_90DP_ESA	Create Trend data for trigger cohort	
17	38	3/17/2011	2009 VACCR Data Received	Description of received 2009 VACCR data	✓
18	39	3/17/2011	VACCR_RECEIVED09.SAS	Add study ID to received 2009 VACCR data	✓
19	40	3/17/2011	VACCR_KEEP09.SAS	Selected records from received 2009 VACCR data	

Data management workflow

- Documents for a program walkthrough
 - Prepare
 - Task, definition, instructions, decision
 - Program list
 - Program code
 - Log of program execution
 - Output from program
 - Processing narrative
 - Purpose
 - Decisions
 - Program step by step
 - Outcome



Clinical Guidelines for ESA use in Cancer: Impacts on Clinical Practice
Data Walk-through Form

Program Name: COHORT.SAS

Description: The program will create cohort for all cause mortality analysis

Analyst: Lucy Zhang

Walk-through date: 11/23/2011

Attendee: Elizabeth Tarlov, Thomas Weichle, Lucy Zhang

Action to be Taken	Completed Action
<ul style="list-style-type: none">For inpatient chemo/radiation/blood transfusion/anemia, if admission date prior Dx but discharge date after the Dx then use INDEX_DX_Dt+1 as the chemo/radiation/blood transfusion/anemia date	<ul style="list-style-type: none">Program updated

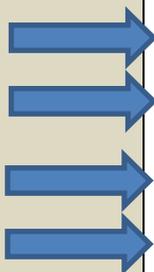
Signature: Elizabeth Tarlov

Signature: Thomas W. Weichle

Example: Cohort
revision
documentation



**Clinical Guidelines for ESA use in Cancer: Impacts on Clinical Practice
Data Walk-through Form**



Program Name: COHORT.SAS

Description: The program will create cohort for all cause mortality analysis

Analyst: Lucy Zhang

Walk-through date: 11/23/2011

Attendee: Elizabeth Tarlov, Thomas Weichle, Lucy Zhang

Action to be Taken	Completed Action
<ul style="list-style-type: none">For inpatient chemo/radiation/blood transfusion/anemia, if admission date prior Dx but discharge date after the Dx then use INDEX_DX_Dt+1 as the chemo/radiation/blood transfusion/anemia date	<ul style="list-style-type: none">Program updated

Signature: Elizabeth Tarlov

Signature: Thomas W. Weichle

Example: Cohort
revision
documentation

**Clinical Guidelines for ESA use in Cancer: Impacts on Clinical Practice
Data Walk-through Form**

Program Name: COHORT.SAS

Description: The program will create cohort for all cause mortality analysis

Analyst: Lucy Zhang

Walk-through date: 11/23/2011

Attendees: Elizabeth Tariov, Thomas Weichle, Lucy Zhang

Action to be Taken	Completed Action
<ul style="list-style-type: none">For inpatient chemo/radiation/blood transfusion/anemia, if admission date prior Dx but discharge date after the Dx then use INDEX_DX_Dt+1 as the chemo/radiation/blood transfusion/anemia date	<ul style="list-style-type: none">Program updated

Signature: Elizabeth Tariov

Signature: Thomas W. Weichle

Example: Cohort
revision
documentation

**Clinical Guidelines for ESA use in Cancer: Impacts on Clinical Practice
Data Walk-through Form**

Program Name: COHORT.SAS

Description: The program will create cohort for all cause mortality analysis

Analyst: Lucy Zhang

Walk-through date: 11/23/2011

Attendee: Elizabeth Tarlov, Thomas Weichle, Lucy Zhang

Action to be Taken	Completed Action
<ul style="list-style-type: none">For inpatient chemo/radiation/blood transfusion/anemia, if admission date prior Dx but discharge date after the Dx then use INDEX_DX_Dt+1 as the chemo/radiation/blood transfusion/anemia date	<ul style="list-style-type: none">Program updated



Signature: _____

Elizabeth Tarlov

Signature: _____

Thomas W. Weichle

Data management workflow

- Elements of analytic file preparation walkthrough
 - Summary of purpose of each program
 - Processing decisions and rationale related to program
 - Differences from previous similar programs for this project
 - Program name, main function, data source name & year
 - Step by step description of program
 - Variables added, dropped, changed
 - Description of how new variables were created
 - Sort & index of output data
 - Links to log, list, freqs & stat files



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					

Session 2: Outline

- Getting started
- Importance of documentation
- Data management workflow
- **Analysis workflow**

Analysis workflow

- Organizing data for analysis
 - Master datasets and work files
 - Data and document versioning
 - Raw data files and statistical analysis files
 - File structures
 - Flat
 - Hierarchical
 - Relational
 - Longitudinal
 - Data back ups

Analysis workflow

- Documenting statistical modelling & analysis
 - Narrative
 - Programming code with annotation
 - Description of model
 - Rationale
 - Statistical package
 - Results
 - Analysis final walkthrough

Analysis workflow

Program description

Cohort handling

```
OS_MODELS_BY_STAGE_v4.do | Untitled.do
1 |***** Open Log File *****
2 |cap log close
3 |log using "I:\SEER_MEDICARE\SURVIVAL\MORTALITY ANALYSIS\STATA\OS_MODELS_BY_STAGE_v4.log", replace
4 |
5 |*****
6 |* PROJECT:    QUALITY & COSTS OF COLON CANCER CARE IN VA AND MEDICARE
7 |* PROTOCOL:   091605 REV VA IIR CCA
8 |*
9 |* PROGRAM:    OS_MODELS_BY_STAGE_v4.DO
10 |*
11 |* AUTHOR:     Tom Weichle
12 |* DATE:       05JUN2013
13 |*
14 |* STATA VER:  12.1
15 |*
16 |* ABSTRACT:   Cohort excludes cases dying within same month of diagnosis.
17 |*
18 |*****
19 |**              MODIFICATION LOG              **
20 |*****
21 |**   Date       Who           Description       **
22 |**   -----   -
23 |*****
24 |
25 |***** Change Directory *****
26 |cd "I:\SEER_MEDICARE\SURVIVAL\MORTALITY ANALYSIS\STATA"
27 |set more off
28 |
29 |***** Open Dataset *****
30 |use "\vhaintmsres4\HSRData2$\ColonData\SEER_MEDICARE\cohort.dta", clear
31 |unique study_id
32 |
33 |***** Assign Formats to Variables *****
34 |run "I:\SEER_MEDICARE\SURVIVAL\MORTALITY ANALYSIS\STATA\FORMAT.do"
35 |
36 |
37 |*****
38 |** Survival Analysis -- Time From Diagnosis Date to Death Date      **
39 |*****
40 |tabulate died
41 |tabstat dx_to_death_month, by(died) stats(n mean sd min p25 median p75 max)
42 |
43 |* SEER-Medicare individuals are diagnosed from 1999-2006.
44 |* --> Cases not dying after the diagnosis date are censored at the later of 12/31/2009 or most recent Medicare claim date.
45 |* VACCR individuals are diagnosed from 2000-2009.
46 |* --> Cases not dying after the diagnosis date are censored at the most recent VA service date.
47 |replace dx_to_death_month = ((year(most_recent_claim_dt_15)-dx_year)*12+(month(most_recent_claim_dt_15)-dx_month))+1 if died == 0
48 |
49 |tabstat dx_to_death_month, by(died) stats(n mean sd min p25 median p75 max)
50 |
51 |***** Identifying records as survival time data *****
52 |* Note: scale measured in years
53 |stset dx_to_death_month, id(study_id) failure(died == 1) scale(12)
```

Example: Analysis program Hynes, Cancer Survival Analysis

Analysis workflow

```
OS_MODELS_BY_STAGE_v4.do  Untitled.do
75  * Centering high school education or more for each Stage at Diagnosis
76  bysort stage: center hs_edu_or_more10, meansave
77
78  * Centering diagnosis age for each Stage at Diagnosis
79  bysort stage: center dx_age, meansave
80
81
82  * Association with chemo and surgery
83  bysort va_rept: tabulate chemo_dx_9ma stage, col expected chi2
84  bysort va_rept: tabulate sug_imp_6ma stage, col expected chi2
85
86  browse study_id sug_imp_6ma dx_to_sug_month chemo_dx_9ma dx_to_chemo_month_dx_9ma dx_to_death_month died _st _d _t _t0
87
88  * Splitting records at surgery and chemo events to treat them as time-varying covariates
89  * Records are split when there isn't already a claim record on the day in which surgery/chemo occurred (i.e. dx_to_sug_month ne dx_to_chemo_month_dx_
90  stsplit surgery_tvc, at(0) after(time = dx_to_sug_month)
91  stsplit chemo_tvc, at(0) after(time = dx_to_chemo_month_dx_9ma)
92  replace surgery_tvc = surgery_tvc + 1 if sug_imp_6ma == 1
93  replace chemo_tvc = chemo_tvc + 1 if chemo_dx_9ma == 1
94  sort study_id _t0
95  by study_id: generate last = (_n == _N)
96  unique study_id
97
98  /* Interpretation of time-varying covariate:
99  At any given time t, individuals already receiving treatment by that time are X times as likely
100  to experience event compared to an individual who has not received treatment (but may receive one later).*/
101
102
103  *** Stage I ***
104  tabulate died if stage6 == "I" & last == 1
105
106  ** VA Reporting Hospital **
107  tabulate died va_rept if stage6 == "I" & last == 1, col row chi2 expected exact
108
109  * Calculate Death Rate (in person-months)
110  stset dx_to_death_month, id(study_id) failure(died == 1)
111  stptime if stage6 == "I", by(va_rept) dd(2) per(100)
112
113  * Compare Death Rates (VA vs. Non-VA)
114  stir va_rept if stage6 == "I"
115
116  streg ib0.va_rept if stage6 == "I", distribution(exponential) nolog
117  margins, dydx(va_rept) predict(hazard)
118
119  test 0.va_rept = 1.va_rept
120
121  * Extended Cox Base Model (including time-varying covariates)
122  * Including colonoscopy indicator
123  stcox ib0.va_rept c_dx_age male black hispanic i.married i.chrlson_grp_6mp_dx other_cancer_6mp_9ma colonoscopy_imp_6ma ///
124  surgery_tvc chemo_tvc op_event_q4 c_hs_edu_or_more10 ib9.division i.grade2 i.xnodetu_cat i.dx_year_grp if stage6 == "I", nolog
125  vif, uncentered
126  estimates store stageI
```

Extended
Cox Base
Model

Analysis workflow

Kaplan-Meier
Survival
Estimates

```
OS_SUMMARY_STATS_BY_STAGE... | Untitled.do
292 * With risktable
293 * Insert minimum survival estimate by Reporting Hospital
294 sts generate surv_stageIII_va_rept = s if stage6 == "III", by(va_rept)
295 tabstat surv_stageIII_va_rept, by(va_rept) stats(n min)
296 sts graph if stage6 == "III", by(va_rept) survival ///
297 title("") ///
298 title("Kaplan-Meier Survival Estimates, by Reporting Hospital", color(black) size(medlarge)) ///
299 subtitle("Stage III", color(black)) ///
300 xtitle("Analysis Time (Years)") ///
301 xlabel(0(1)11) ///
302 ytitle("Probability") ///
303 yline(0 .0939043, lstyle(refline) lpattern(dash) lcolor(gs10) lwidth(vthin)) ///
304 legend(title("Stage III", color(black) size(medlarge)) position(6) ring(0) rows(1) ///
305 label(1 "Non-VA") label(2 "VA") order(2 1)) ///
306 risktable(, size(small) order(2 "VA" 1 "Non-VA") rowtitle(, justification(left) title(, size(small) at(rowtitle))) ///
307 caption("Note: Kaplan-Meier estimates are unadjusted.") ///
308 plotlopts(lpattern(dash) lwidth(medthick)) ///
309 plot2opts(lpattern(solid) lwidth(medthick)) ///
310 text(0.95 9.0 "Log-rank test: p = 0.2712", box fcolor(white) margin(small)) ///
311 scheme(slmono) plotregion(style(none))
312
313 sts graph if stage6 == "IV", by(va_rept) survival ///
314 title("Kaplan-Meier Survival Estimates, by Reporting Hospital", color(black) size(medlarge)) ///
315 subtitle("Stage IV", color(black)) ///
316 xtitle("Analysis Time (Years)") ///
317 xlabel(0(1)11) ///
318 ytitle("Probability") ///
319 legend(label(1 "Non-VA") label(2 "VA") rows(1)) ///
320 caption("Note: Kaplan-Meier estimates are unadjusted.") ///
321 plotlopts(lpattern(dash) lwidth(medthick)) ///
322 plot2opts(lpattern(solid) lwidth(medthick)) ///
323 text(0.95 9.0 "Log-rank test: p = 0.0375", box margin(small))
324
325 * With risktable
326 * Insert minimum survival estimate by Reporting Hospital in yline
327 sts generate surv_stageIV_va_rept = s if stage6 == "IV", by(va_rept)
328 tabstat surv_stageIV_va_rept, by(va_rept) stats(n min)
329 sts graph if stage6 == "IV", by(va_rept) survival ///
330 title("") ///
331 /*title("Kaplan-Meier Survival Estimates, by Reporting Hospital", color(black) size(medlarge)) ///
332 subtitle("Stage IV", color(black))*///
333 xtitle("Analysis Time (Years)") ///
334 xlabel(0(1)11) ///
335 ytitle("Probability") ///
336 yline(0 .0167081, lstyle(refline) lpattern(dash) lcolor(gs10) lwidth(vthin)) ///
337 legend(title("Stage IV", color(black) size(medlarge)) position(3) ring(0) rows(1) ///
338 label(1 "Non-VA") label(2 "VA") order(2 1)) ///
339 risktable(, size(small) order(2 "VA" 1 "Non-VA") rowtitle(, justification(left) title(, size(small) at(rowtitle))) ///
340 /*caption("Note: Kaplan-Meier estimates are unadjusted.")*///
341 plotlopts(lpattern(dash) lwidth(medthick)) ///
342 plot2opts(lpattern(solid) lwidth(medthick)) ///
343 text(0.95 9.0 "Log-rank test: p = 0.0375", box fcolor(white) margin(small)) ///
344 scheme(slmono) plotregion(style(none))
```

Example: Analysis program Hynes, Cancer Survival Analysis

Analysis workflow

The screenshot displays the Stata command window and the results of a Cox regression analysis. The command window shows the following code:

```

* Extended Cox Base Model (including time-varying covariates)
* Including colonoscopy indicator
. stcox ib0.va_rept c_dx_age male black hispanic i.married i.chrlson_grp_6mp_dx other_cancer_6mp_9ma colono
scopy_1mp_6ma ///
      surgery_tvc chemo_tvc op_event_q4 c_hs_edu_or_more10 i.division i.grade2 i.xnodetu_cat i.dx_y
> es, grp if stage6 == "I", nolog

      failure _t: died == 1
      analysis time _t: dx_to_death_month/12
      id: study_id

Cox regression -- Breslow method for ties

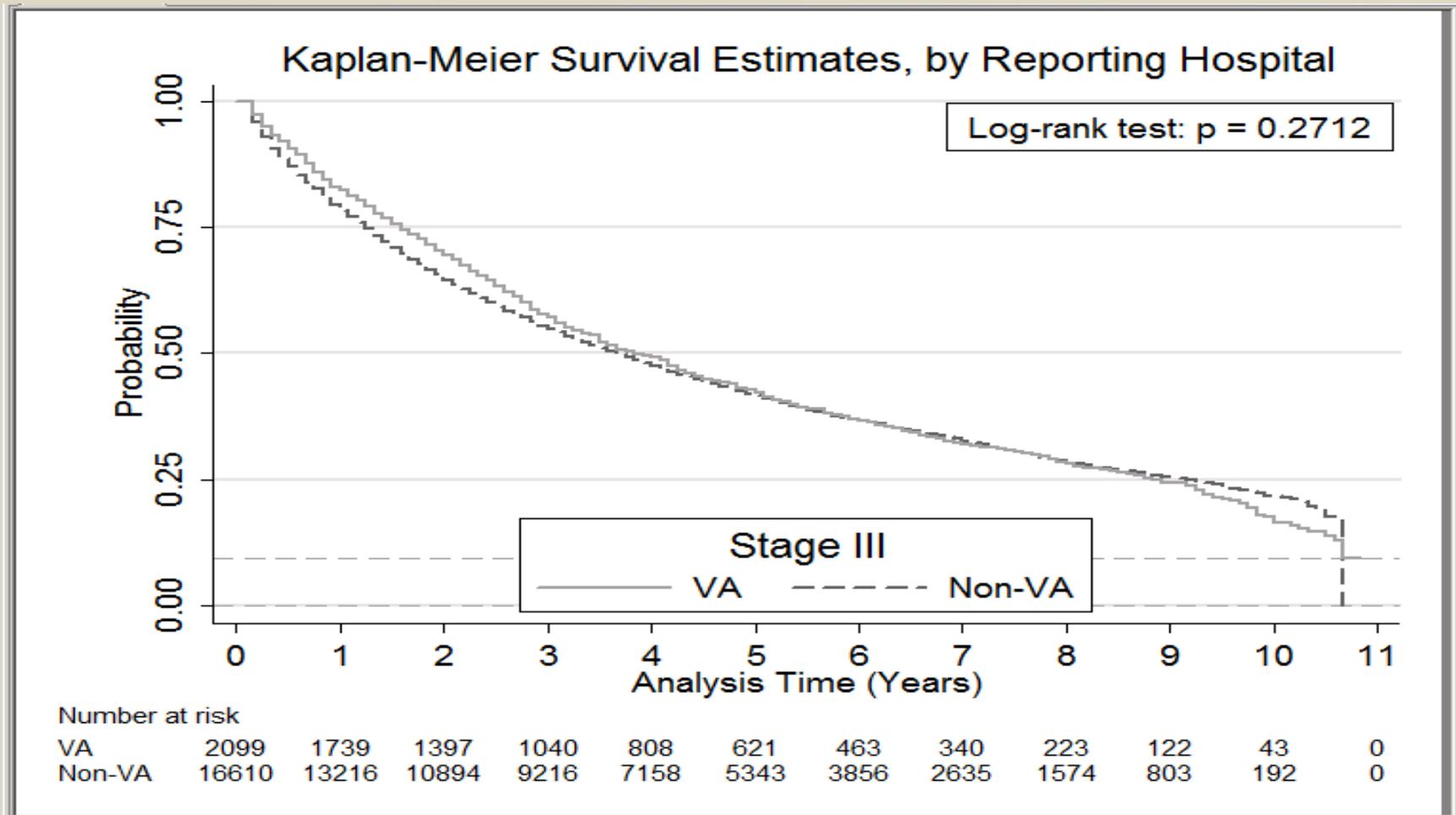
No. of subjects =          19335          Number of obs =          36624
No. of failures =           9060
Time at risk   =  93600.41667

Log likelihood = -81760.651          LR chi2(36) =          5089.76
                                      Prob > chi2  =           0.0000
  
```

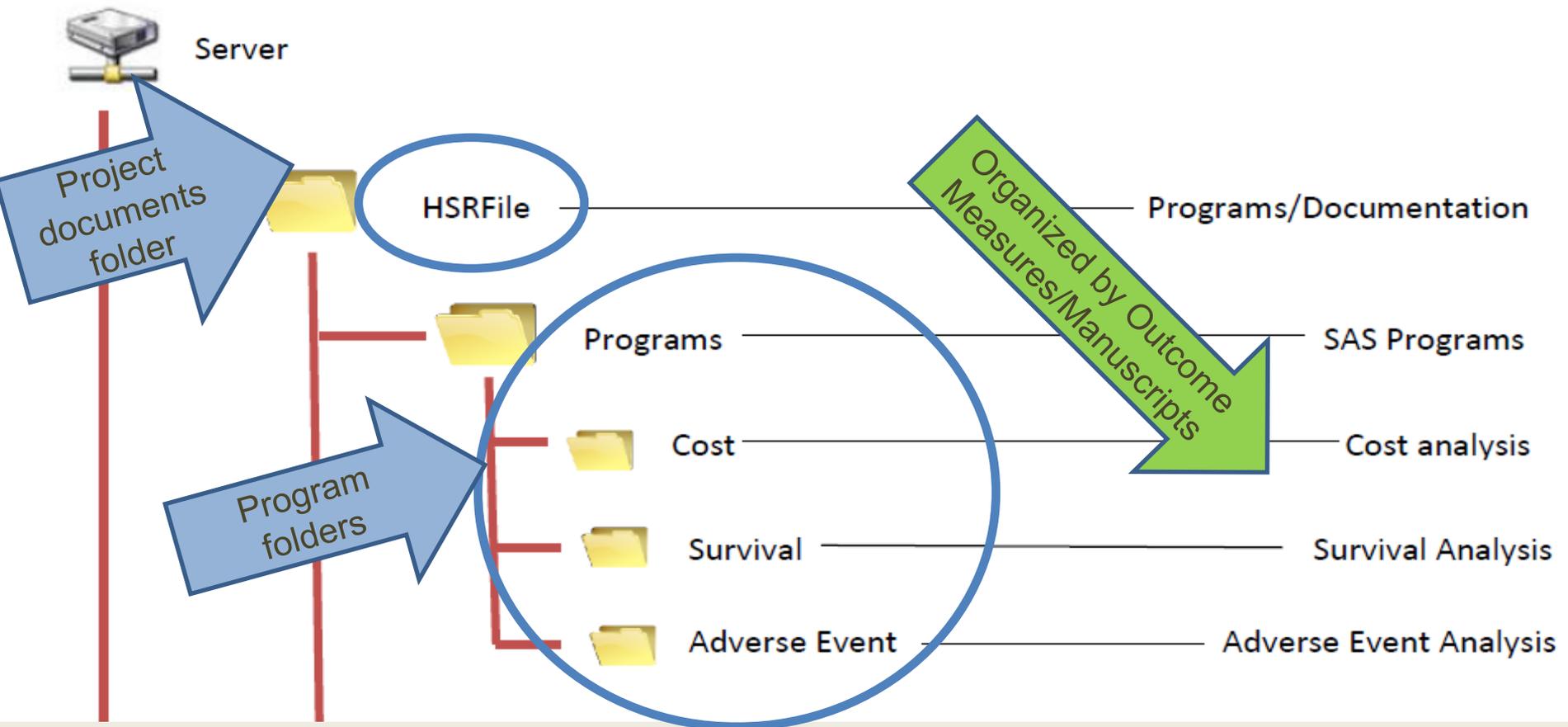
The results table shows the following data:

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
1.va_rept	.8801917	.03028	-3.71	0.000	.8228006 .9415859
c_dx_age	1.074572	.0018315	42.20	0.000	1.070989 1.078168
male	1.32388	.0328975	11.29	0.000	1.260948 1.389954
black	1.052382	.040068	1.34	0.180	.9767088 1.133918
hispanic	.8999876	.0506103	-1.87	0.061	.8060643 1.004855
married					
1	.8502622	.0202387	-6.81	0.000	.8115062 .8908691
9	.8835227	.0497325	-2.20	0.028	.7912332 .9865768
chrlson_grp_6mp_dx					
1	1.554609	.0428932	15.99	0.000	1.472772 1.640992
2	2.421159	.0667374	32.08	0.000	2.293827 2.55556
3	4.024083	.1623643	34.51	0.000	3.718113 4.355233
other_cancer_6mp_9ma	1.403541	.0371008	12.82	0.000	1.332676 1.478174
colonoscopy_1mp_6ma	.8997215	.0203237	-4.68	0.000	.8607567 .9404502
surgery_tvc	.6525589	.0208939	-13.33	0.000	.612866 .6948224
chemo_tvc	1.59902	.074369	10.09	0.000	1.459705 1.75163
op_event_q4	1.152826	.0280828	5.84	0.000	1.099078 1.209202
c_hs_edu_or_more10	.9702623	.0087543	-3.35	0.001	.9532549 .9875731

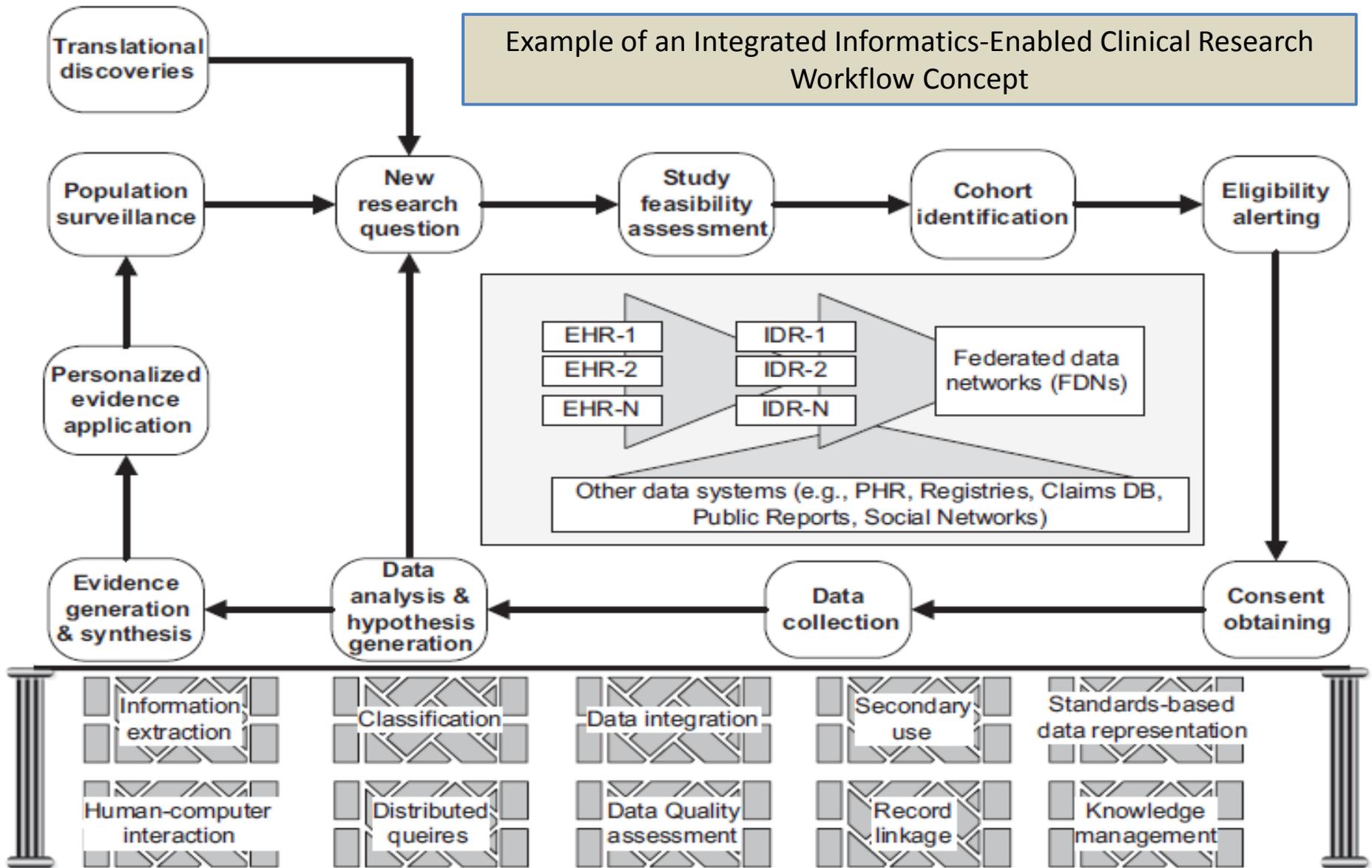
Analysis workflow



Data analysis workflow



Example of an Integrated Informatics-Enabled Clinical Research Workflow Concept



Citation: Kahn, MG, Weng, C. Clinical research informatics: a conceptual perspective. *J Am Med Inform Assoc.* 2012;19:e36-e42.

QUESTIONS

Contact Information

Denise Hynes, PhD, MPH, RN

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 2: Managing and Documenting Data Workflow

- Recap
 - Getting started
 - Importance of documentation
 - Data management workflow
 - Analysis workflow

Session 3: Planning for Data Re-use

- Preview
 - 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