Good Data Practices
Session 2
Managing and Documenting Data Workflow

Denise Hynes, PhD, MPH, RN
September 10, 2013
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)
Funding Agency Requirements

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.

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.

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 control towers will be exposed to influence of source effects associated with wind directions. In addition to the CO2 data, we will collect data on additional variables such as temperature, humidity, precipitation, and cloud cover. Site conditions at Mauna Loa will be retained. The final data product will consist of 5-minute, 15-minute, hourly, daily, and monthly data.

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 released to the public as soon as the recalibration of standard gases 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://scrippscjo2.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.
Creating a Data Management Plan

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 procedures: i.e., how will you make data available?

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
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.

Specifications and documentation

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

Server

- Project documents folder
  - HSRFile
  - Programs
    - Cost
    - Survival
    - Adverse Event
  - Organized by Outcome Measures/Manuscripts
    - Programs/Documentation
    - SAS Programs
    - Cost analysis
    - Survival Analysis
    - Adverse Event Analysis

- Program folders
Data management workflow

Program Documents
- Cost
- Survival
- Adverse Event
- Data Walk-through
- Code Book
- Administrative
- Publication and Presentation

Program document folders
Data management workflow

- Project data folder
- HSRData
  - Cost
  - Survival
  - Adverse Event
- HSRXWalk
  - Data with sensitive information
- Datasets
- Data folders
Data management workflow

Identifier crosswalk in secure folder
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
30


**RCT Cohort Flowchart Example**

Source: Open Mesh versus Laparoscopic Mesh Repair of Inguinal Hernia, VA CSP Project # 456; PI: L. Neumayer
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. 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 year or taken into account any baseline differences between study groups that may have affected cost effectiveness. Shorter operation times and greater use of outpatient techniques could also help reduce costs.

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

Citation: Hynes, DM, Stroupe, K, Luo, P, Hurder, AH, Neumayer, L, et al., Cost effectiveness of laparoscopic versus open tension-free hernia repair: Results from the Veterans Affairs Cooperative Study. *Journal of the American College of Surgeons.* 2006; 203: 447-457
Observational Study Flowchart Example

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

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 healthcare 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.5,6,11

Briefly, the sampling frame comprised all elderly veterans who were known to the VA (i.e., 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 VA registry and then validate them with matching to the VA Central Cancer Registry. The resulting analytic cohort comprised 166 VA patients and 7,951 non-VA patients.
Walkthrough document for creation of final cohort for ESA study

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

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

**Data used**

---

### 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_REVEIVED08.SAS7BDAT
Saved in: HSRXWALK\ESA_CancerXwalk

Variables Description in VACCR_RECEIVED data set:

<table>
<thead>
<tr>
<th>Number of Observation</th>
<th>45,225</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Variables</td>
<td>89</td>
</tr>
</tbody>
</table>
### Modification of program:

<table>
<thead>
<tr>
<th>Sequence #</th>
<th>Date Modified</th>
<th>Reason Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>06/24/2009</td>
<td>Create INDEX_SUG_VENUE_REVISE variable</td>
</tr>
<tr>
<td>002</td>
<td>09/09/2009</td>
<td>Create VA_REPT2 variable</td>
</tr>
<tr>
<td>003</td>
<td>09/11/2009</td>
<td>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</td>
</tr>
<tr>
<td>004</td>
<td>09/14/2009</td>
<td>Rename DIED and DX_TO_DEATH_DAY: DIED=DIED9904, DX_TO_DEATH_DAY=DX_TO_DEATH_DAY9904</td>
</tr>
<tr>
<td>005</td>
<td>09/14/2009</td>
<td>Create DIED and DX_TO_DEATH_DT</td>
</tr>
<tr>
<td>006</td>
<td>03/06/2012</td>
<td>Create chemo within 9 month and radiation within 9 month flags</td>
</tr>
</tbody>
</table>

### 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 
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.
**Data management workflow**

### New variables created:

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMO1-HMO96</td>
<td>The flag indicates if the patients enrolled in HMO in each month from year 2002-2009.</td>
</tr>
</tbody>
</table>

---

Clinical Guidelines for ESA use in Cancer
Impacts on Clinical Practice

<table>
<thead>
<tr>
<th></th>
<th>Values: 1 = ‘Yes’ 0 = ‘No’</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARTB1-PARB96</td>
<td>The flag indicates if the patient has Part B coverage only each month from year 2002 to 2009. The flag indicates if the patient has other primary payer in each month from year 2002 to 2009.</td>
</tr>
<tr>
<td>Values: 1 = ‘Yes’ 0 = ‘No’</td>
<td></td>
</tr>
</tbody>
</table>

---

Example: Documenting derived variables
### Clinical Guidelines for ESA Use in Cancer Impacts on Clinical Practice

#### SAS Program Walk-through Summary

<table>
<thead>
<tr>
<th>Seq Num</th>
<th>Date</th>
<th>Program Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>2/14/2011</td>
<td>CHEMO.SAS</td>
<td>Get Chemo records for incident cases</td>
</tr>
<tr>
<td>29</td>
<td>2/14/2011</td>
<td>RADIATION.SAS</td>
<td>Get Radiation records for incident cases</td>
</tr>
<tr>
<td>30</td>
<td>2/15/2011</td>
<td>EPO.SAS</td>
<td>Get ESA records for incident cases</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Program modified on 05/18/2011 and Updated program and document attached</td>
</tr>
<tr>
<td>31</td>
<td>2/15/2011</td>
<td>HEMOGLOBIN_LAB.SAS</td>
<td>Get Hgb records from DSS LAB data for incident cases</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Program modified on 05/18/2011 and Updated program and document attached</td>
</tr>
<tr>
<td>32</td>
<td>2/15/2011</td>
<td>HEMOGLOBIN_LAR.SAS</td>
<td>Get Hgb records from DSS LAR data for incident cases</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Program modified on 05/18/2011 and Updated program and document attached</td>
</tr>
<tr>
<td>33</td>
<td>2/23/2011</td>
<td>ALL_CANCER_COHORT</td>
<td>Create all Cancer Cohort</td>
</tr>
<tr>
<td>33.1</td>
<td>9/7/2011</td>
<td>DOD_FOR_PRIOR_ON_DX_DEATH.SAS</td>
<td>Assign death date for the cases died prior or on the day of diagnosis</td>
</tr>
<tr>
<td>34</td>
<td>2/25/2011</td>
<td>BLOOD_TRANSFUSION_AND_ANEMIA.SAS</td>
<td>Get blood transfusion and anemia records for incident cases</td>
</tr>
<tr>
<td>35</td>
<td>2/25/2011</td>
<td>LUNG_COLON_COHORT</td>
<td>Create Lung and Colon Cohort</td>
</tr>
<tr>
<td>35.1</td>
<td>2/28/2011</td>
<td>COHORT_CHEMO.SAS</td>
<td>Create cohort for pre/post paper</td>
</tr>
<tr>
<td>36</td>
<td>3/1/2011</td>
<td>TRIGGER_HGB_COHORT</td>
<td>Create Trigger Analysis Cohort</td>
</tr>
<tr>
<td>37</td>
<td>3/9/2011</td>
<td>HGB_TRENDS_90DP_ESA</td>
<td>Create Trend data for trigger cohort</td>
</tr>
<tr>
<td>38</td>
<td>3/17/2011</td>
<td>2009 VACCR Data Received</td>
<td>Description of received 2009 VACCR data</td>
</tr>
<tr>
<td>39</td>
<td>3/17/2011</td>
<td>VACCR_RECEIVED2009.SAS</td>
<td>Add study ID to received 2009 VACCR data</td>
</tr>
<tr>
<td>40</td>
<td>3/17/2011</td>
<td>VACCR_KEEP2009.SAS</td>
<td>Selected records from received 2009 VACCR data</td>
</tr>
</tbody>
</table>

**Documentation:** Analysis File Preparation Program List
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

<table>
<thead>
<tr>
<th>Action to be Taken</th>
<th>Completed Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 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</td>
<td>• Program updated</td>
</tr>
</tbody>
</table>

Signature: Elizabeth Tarlov
Signature: Thomas W. Weichle
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

<table>
<thead>
<tr>
<th>Action to be Taken</th>
<th>Completed Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>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</td>
<td>Program updated</td>
</tr>
</tbody>
</table>

Signature: Elizabeth Tarlov

Signature: Thomas W. Weichle
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 Tarlov, Thomas Weichle, Lucy Zhang

<table>
<thead>
<tr>
<th>Action to be Taken</th>
<th>Completed Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 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</td>
<td>• Program updated</td>
</tr>
</tbody>
</table>

Signature: Elizabeth Tarlov

Signature: Thomas W. Weichle
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

<table>
<thead>
<tr>
<th>Action to be Taken</th>
<th>Completed Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 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</td>
<td>• Program updated</td>
</tr>
</tbody>
</table>

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

Signature of reviewers
Welcome to VINCI Workspace
Standard and Development Workspaces

**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 JavaScript 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.
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

Example: Analysis program Hynes, Cancer Survival Analysis
Analysis workflow

Example: Analysis program Hynes, Cancer Survival Analysis

Extended Cox Base Model
Analysis workflow

Kaplan-Meier Survival Estimates

Example: Analysis program Hynes, Cancer Survival Analysis
Analysis workflow

Stata: Time-to-Event Survival Analysis Model
Analysis workflow

Kaplan-Meier Survival Estimates, by Reporting Hospital

Log-rank test: $p = 0.2712$

Number at risk
- VA: 2099, 1739, 1397, 1040, 808, 621, 463, 340, 223, 122, 43, 0
- Non-VA: 16610, 13216, 10894, 9216, 7158, 5343, 3856, 2635, 1574, 803, 192, 0

Kaplan-Meier Survival Estimates
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