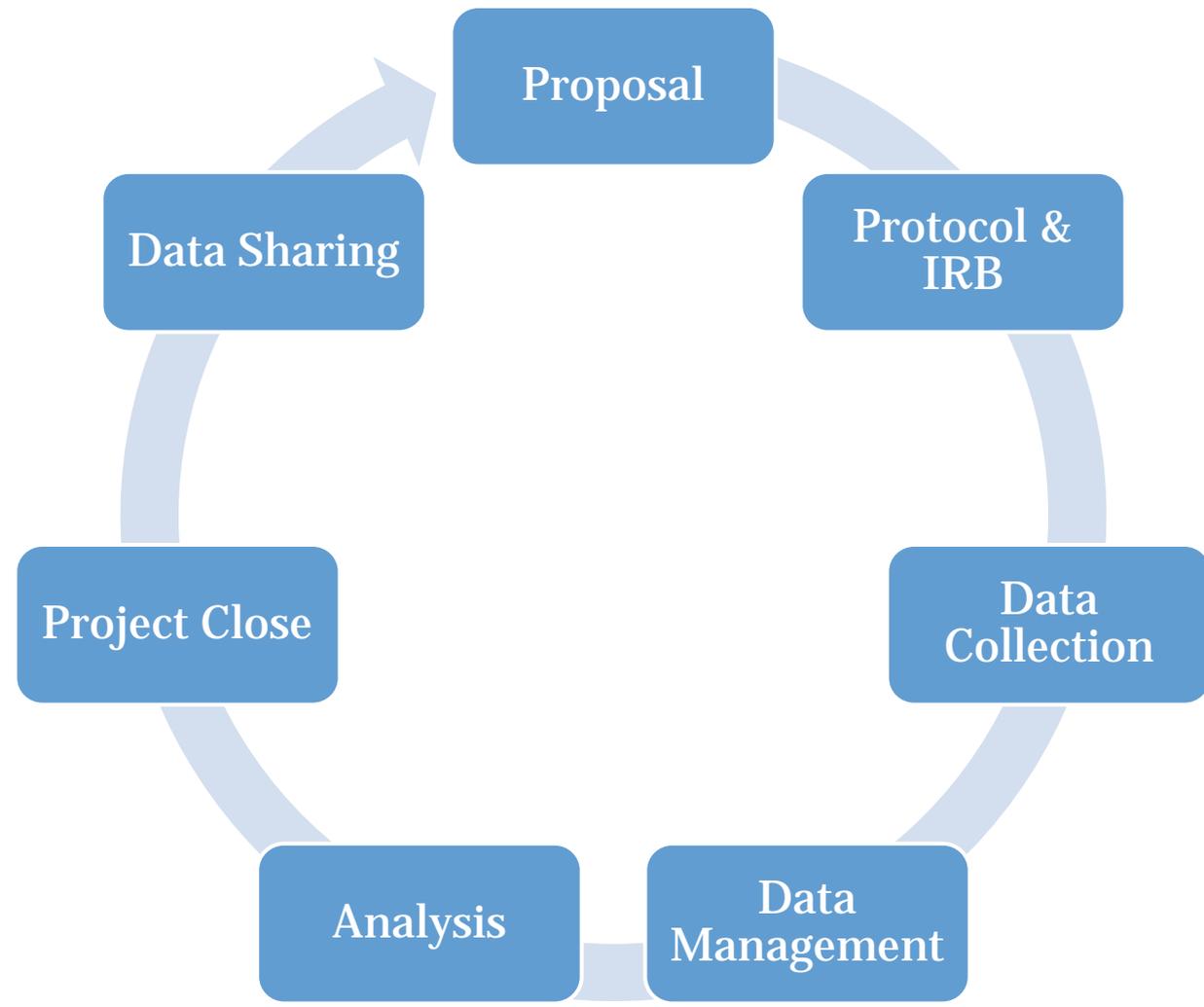


GDP 2014 Acknowledgements

- **Laurel Copeland, San Antonio VA**
- **Brian C. Sauer, Salt Lake City VA**
- **Kevin Stroupe, Hines VA**
- **Linda Williams, Indianapolis VA**
- **Denise Hynes, VIREC**
- **Arika Owens, VIREC**
- **Maria Souden, VIREC**

Research Life Cycle



Good Data Practices Series Overview

May 8

- *The Best Laid Plans: Plan Well, Plan Early* - Jennifer Garvin

May 15

- *"The Living Protocol:" Managing Documentation While Managing Data* - Matt Maciejewski

May 22

- *Controlled Chaos: Tracking Decisions During an Evolving Analysis* - Pete Groeneveld

May 29

- *Reduce, Reuse, Recycle: Planning for Data Sharing* - Linda Kok

About me ..

- **Core Investigator, VA Center for Health Equity Research & Promotion**
- **Health Services investigator with a particular interest in health economics**
- **PI of a Merit Award investigating cardiovascular technology utilization & outcomes in heart failure**
- **10 years of experience using VA data for research**
- **Vice-Chair of the Research & Development Committee at Philadelphia VAMC**

About you...

- What is your role in research and level of experience?
 - Research investigator
 - New? Experienced?
 - Data manager/analyst
 - New? Experienced?
 - Project coordinator
 - New? Experienced?
 - Other – please describe via the Q&A function

Agenda for *Controlled Chaos*: *Tracking Decisions During an Evolving Analysis*

- Challenges of documentation
- Why good documentation?
- A schematic for document organization
- “Good practices:” Documents, communications & presentations
- Why organize?

Challenges of documentation

The challenge of documenting observational data research

- Cohort selection and analysis details evolve during project
- Dataset formation and analytics are evolving processes

Additional challenges ...

- Unanticipated data problems can require creative “one-off” solutions
- Multiple people are creating and modifying study documents simultaneously
- Changes in the study team, protocol, or research environment

Why good documentation?

Adverse effects of poor documentation

- It's time to write the manuscript, but nobody can remember how the cohort was formed!
- A key staff member (or PI!) leaves VA, and nobody can duplicate her/his memory or intricate knowledge of the project
- Lessons learned in the project have to be re-learned (the hard way) in the next project

Essential for good project management

- **Lack of clarity about milestones can**
 - bog down a project
 - result in misunderstandings &
 - needless duplication of tasks
- **Can be difficult for PI or project manager to monitor progress without milestone accounting**

Helps identify potential problems

- Problems impeding research progress can fester if documentation doesn't identify them
- Bad documentation increases the risk of
 - analyzing the wrong dataset
 - using the wrong model
 - producing erroneous results



It's essential for regulatory compliance

- **Good documentation demonstrates to auditors that**
 - **Only data approved by IRB was obtained & used.**
 - **The number of patients in research cohorts is within the explicit limits of the research protocol**

Also, auditors like to see that ...

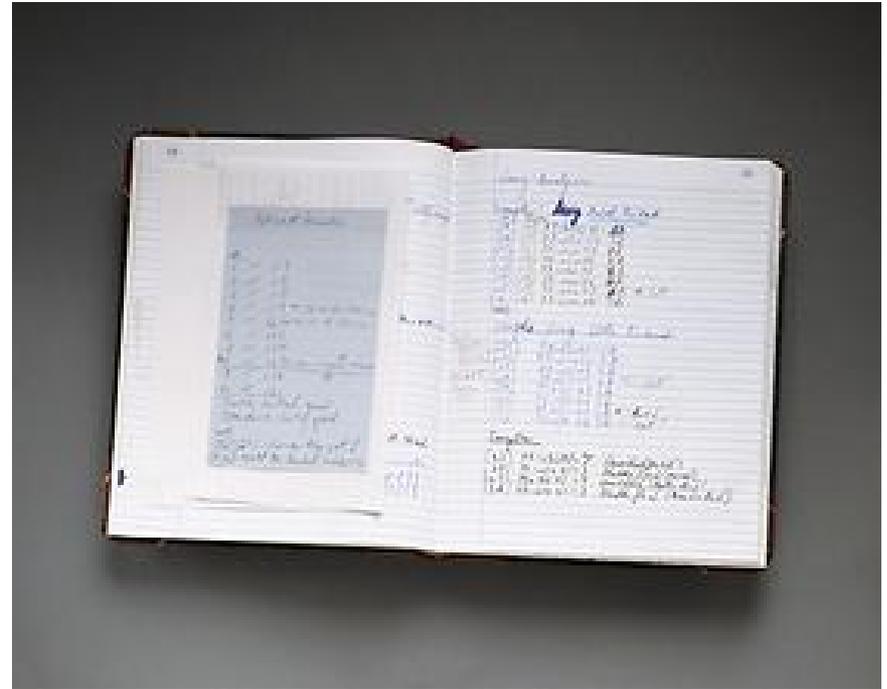
- Analyses of the data conform to the IRB-approved protocol (i.e., no “side-bar” research)
- Only authorized project staff are touching the data.



A schematic for
document organization

Good documentation in the 20th century

- The Laboratory Notebook
- Single volume in which experimental ideas, research plans, and results were recorded
- Primary source document for scientific manuscripts
- A non-shareable, irreplaceable document made of highly flammable material!
- Loss or destruction of a PI's Lab Notebook was disastrous



Source: National Library of Medicine
(<https://www.nlm.nih.gov/visibleproofs/galleries/cases/saldivar.html>)

Typical dataset formation milestones requiring careful documentation

- Search strategy for the initial data “pull” and establishment of the rough-cut study datasets
- Cleaning of key selection variables
- Exclusions based on cleaned selection variables
- Cleaning/recoding analytic variables
- Merging datasets and finalizing cohort

Typical dataset formation milestones requiring careful documentation

Search strategy for the initial data “pull” and rough-cut study datasets

Cleaning of key selection variables

Exclusions based on cleaned selection variables

Cleaning/recoding analytic variables

Merging datasets and finalizing cohort

Other milestones needing documentation

- Producing “Table 1” summary statistics.
- Design of primary analytic models, including imputation schemes
- Tests of model robustness and underlying assumptions (e.g., proportional hazards, heteroskedasticity)
- Sub-group and sensitivity analyses
- Exploratory analyses within the scope of the project

Other milestones needing documentation

Producing
“Table 1”
summary
statistics

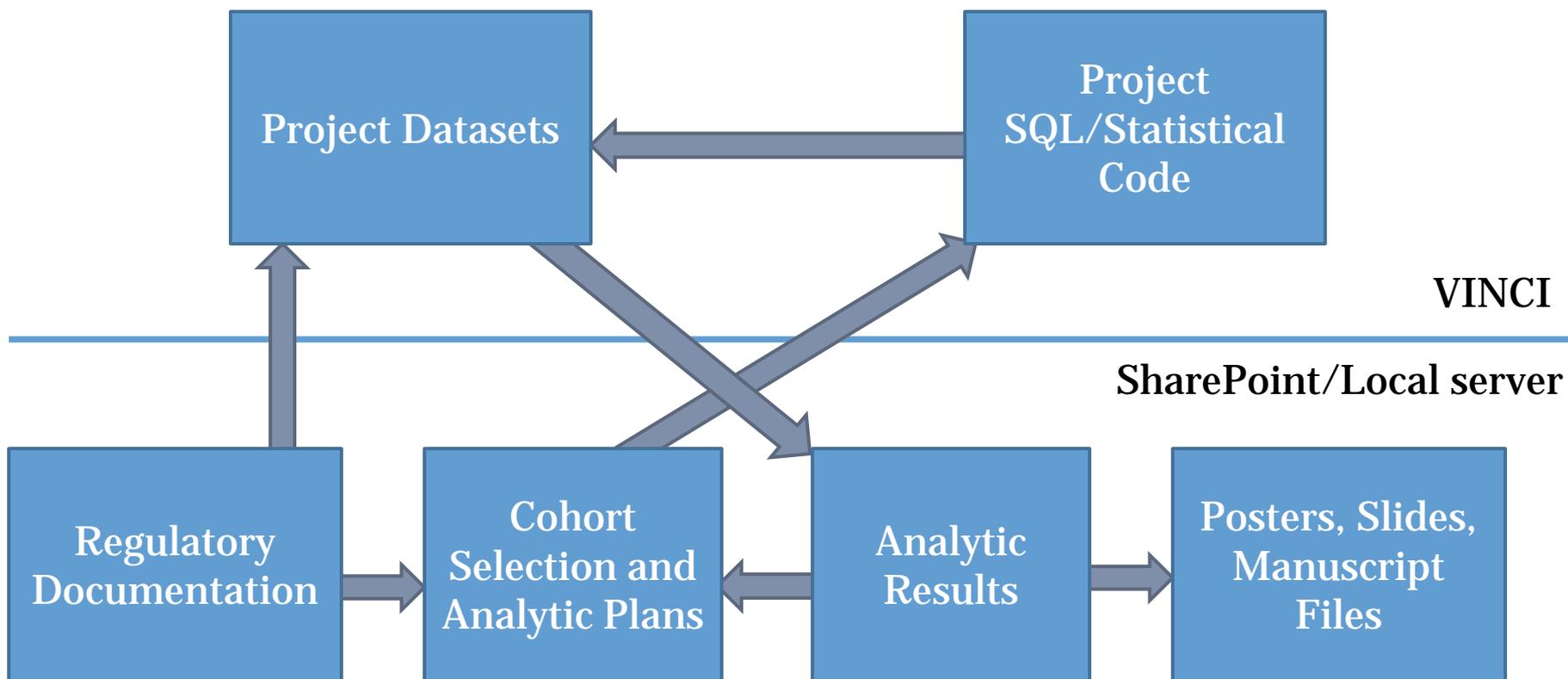
Design of
primary
analytic
models,
including
imputation
schemes

Tests of
model
robustness
and
underlying
assumptions

Sub-group
and
sensitivity
analyses

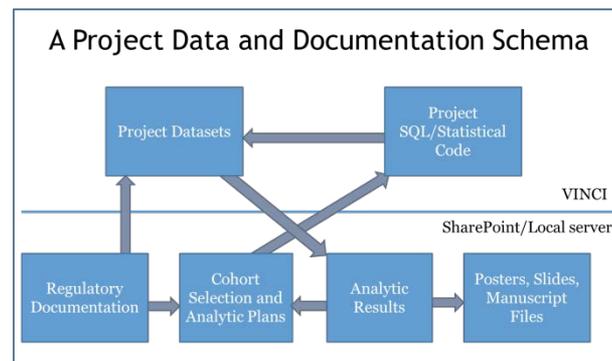
Exploratory
analyses
within the
scope of the
project

A project data and documentation schema



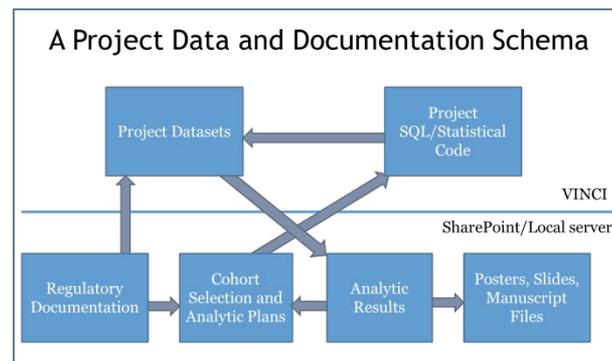
Overarching goals of schema

- Documentation of all scientific process decisions in “real time”
- All 6 filetypes should be clearly linkable
- Unambiguous version control
- An entirely new PI & team should be able to take over the project and understand what was done!



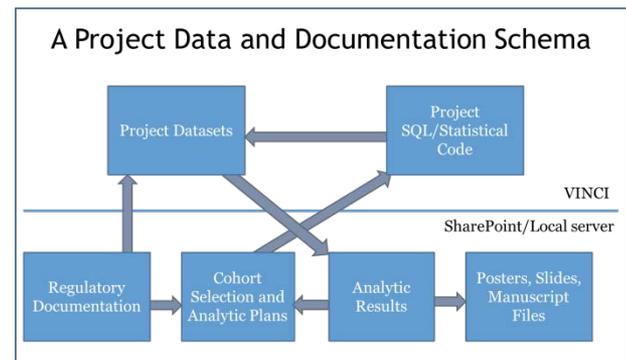
Folder setup and maintenance

- All study personnel should have shared access to the documents they need
- No study documents should be stored outside the system
- Everyone must follow the “rules of the road” for naming, updating, sharing, and archiving documents



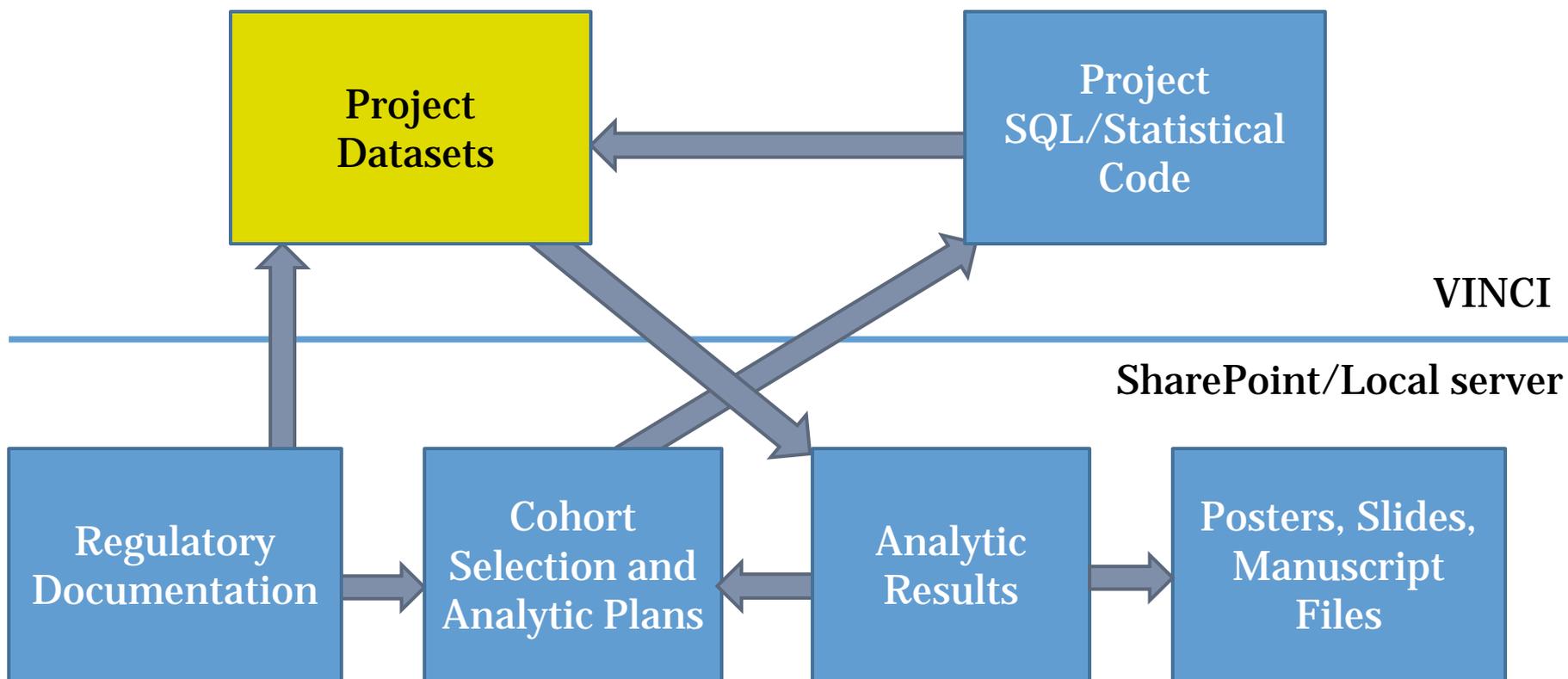
Critical to use archives

- All 6 folder types (datasets, code, regulatory, analytics, output, presentation) should have an “Archive” subfolder where “old” and/or “draft” versions of documents are stored
- Good practice to move old versions out of the active portion of the folder **as soon as** they become “old”



“Good practices” for documents,
communications & presentations

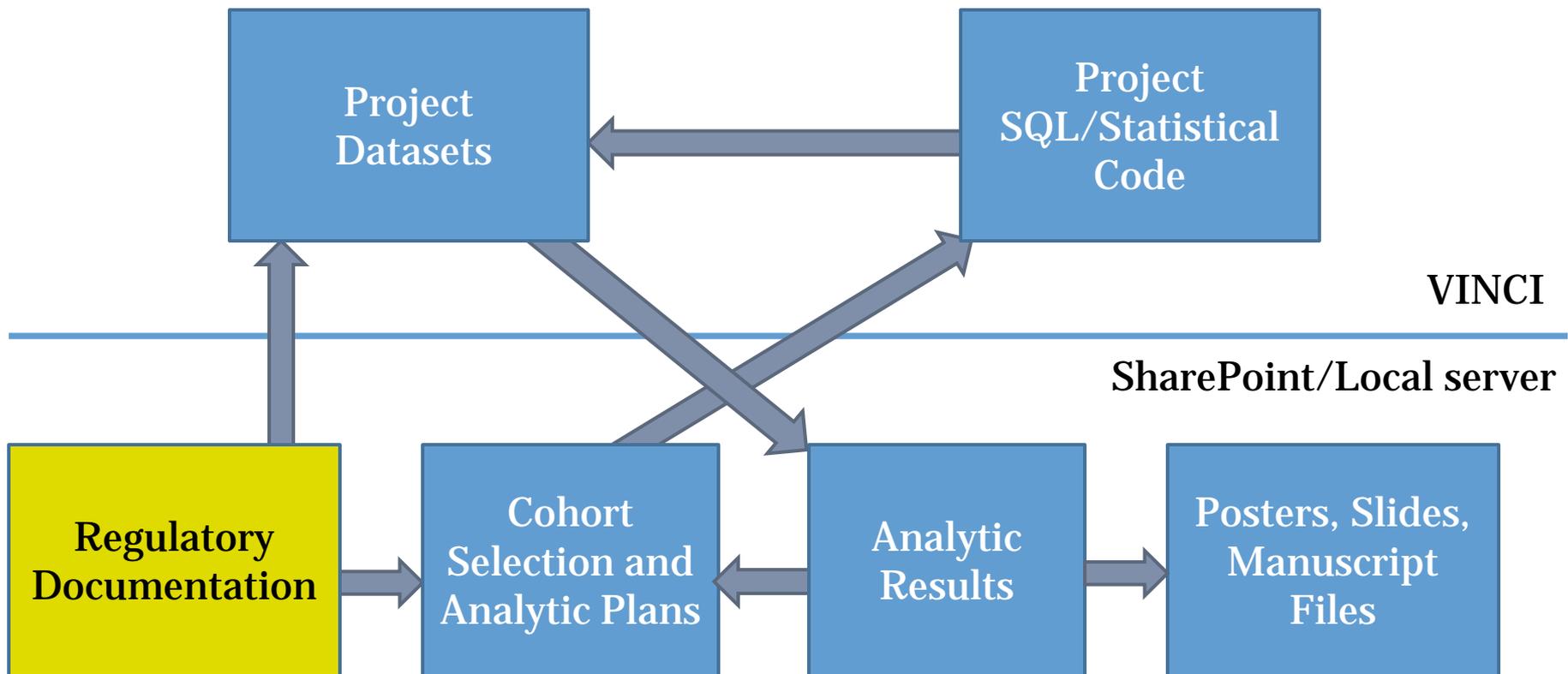
Project data and documentation schema



Dataset folder

- **'Raw' datasets should be**
 - clearly labeled
 - stored in their own subfolders
- **Don't mingle processed and unprocessed data!**
- **Derived datasets should be linked -by dataset name- to the code that produced them:**
 - CohSel_04142014.sas produces
 - Cohort_04142014.sas7bdat

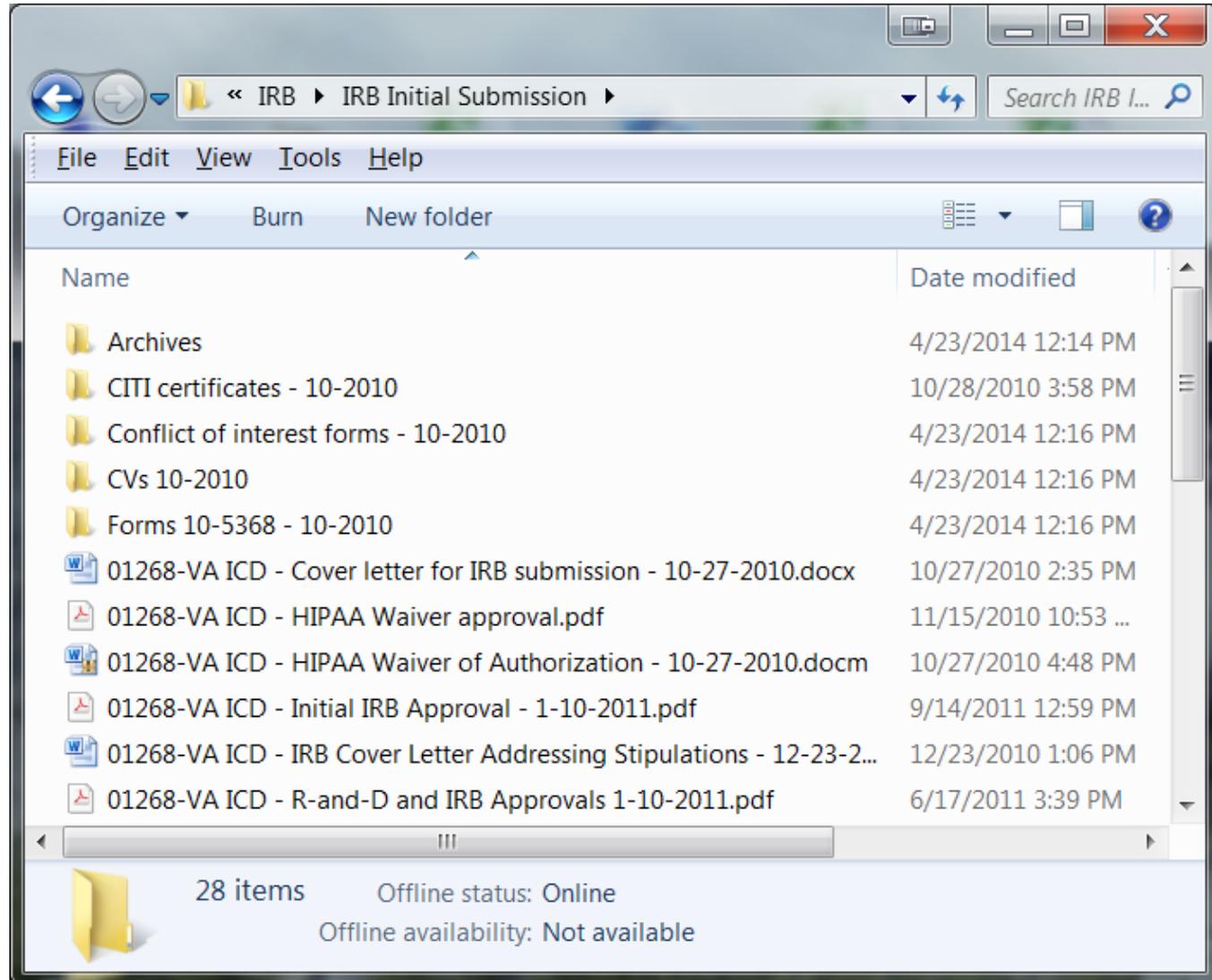
Project data and documentation schema



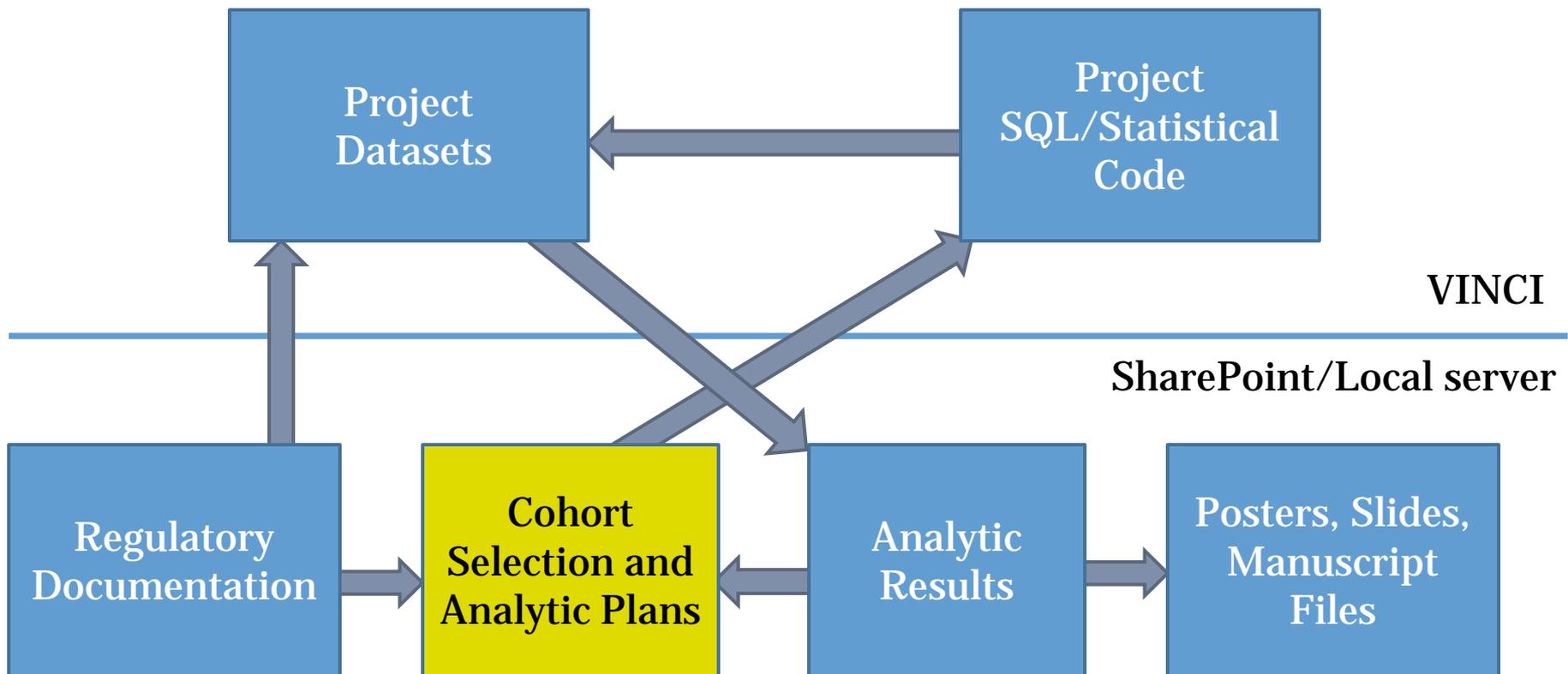
Regulatory folder setup

- Essentially the electronic version of the “Regulatory Binder”
- All IRB/R&D forms, correspondence, approvals, with dates in the file names
- Data use agreements and other external regulatory documents
- Archiving of defunct/draft versions

Regulatory folder



Project data and documentation schema

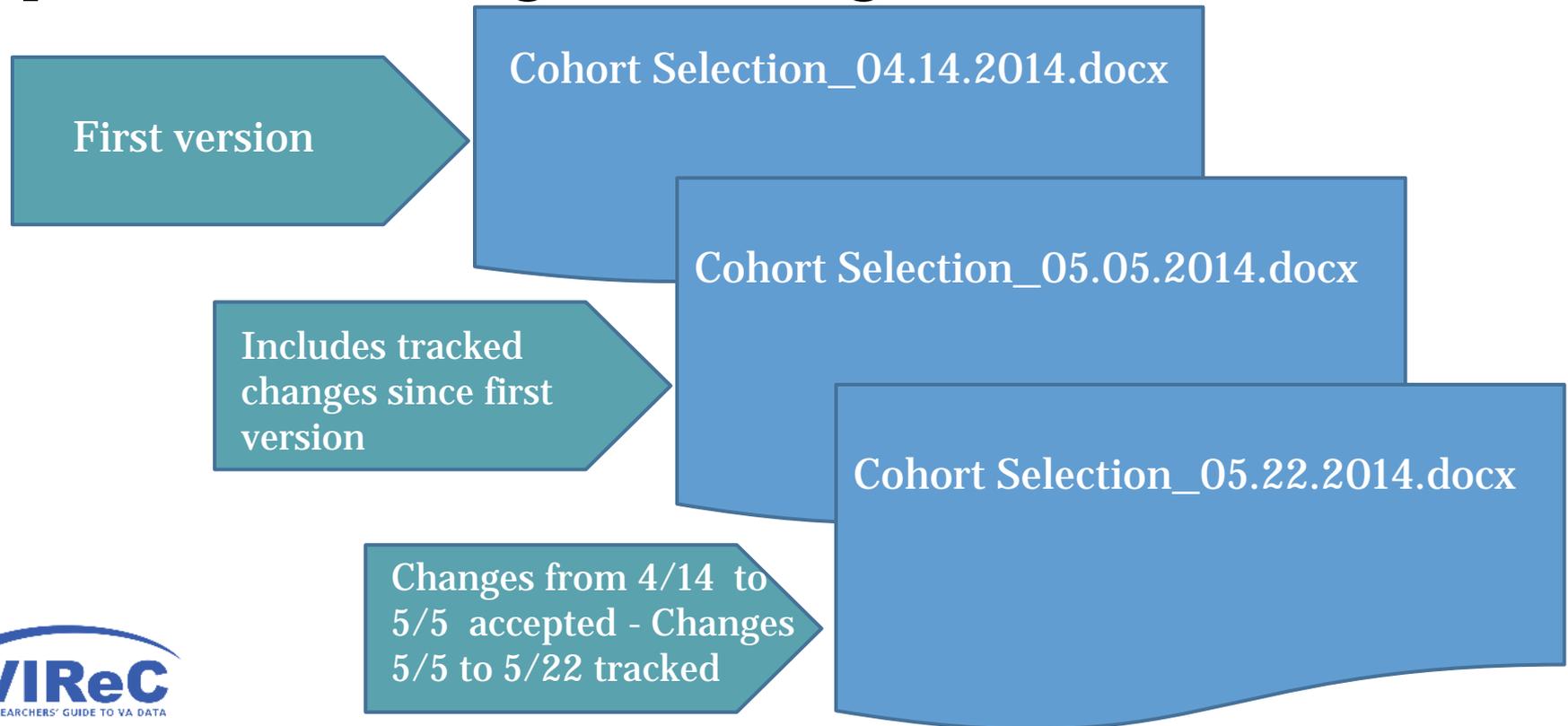


Good practices for analytic plans: Filenames

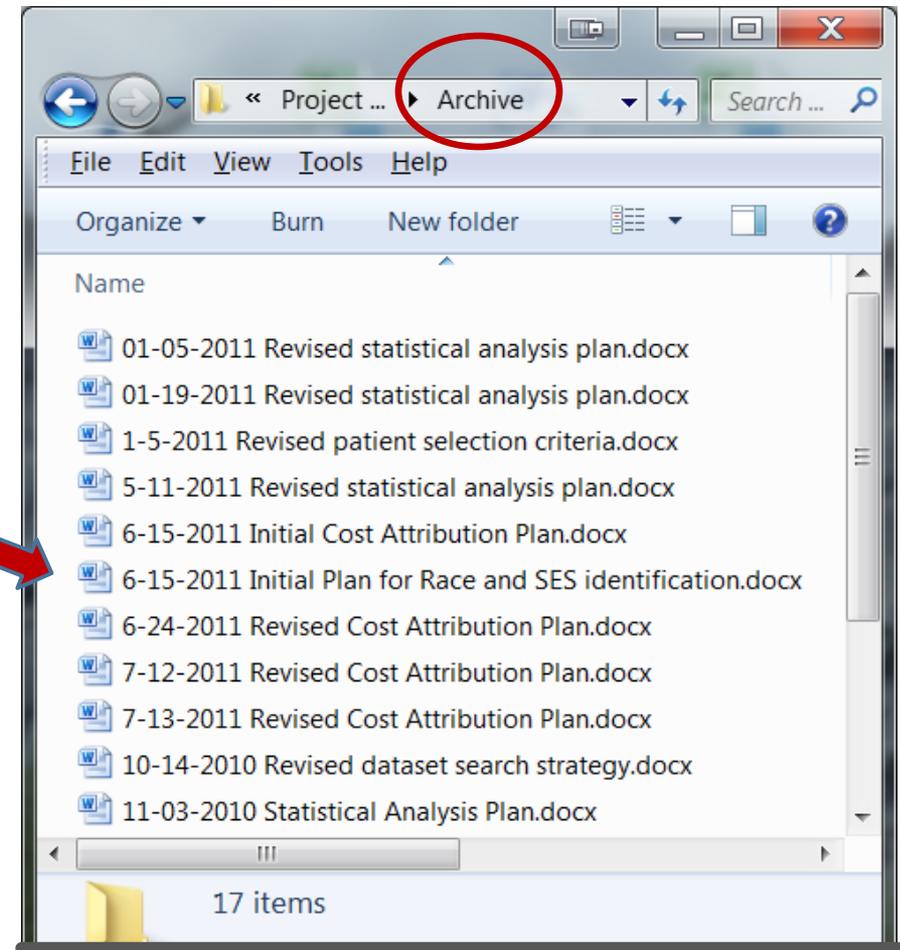
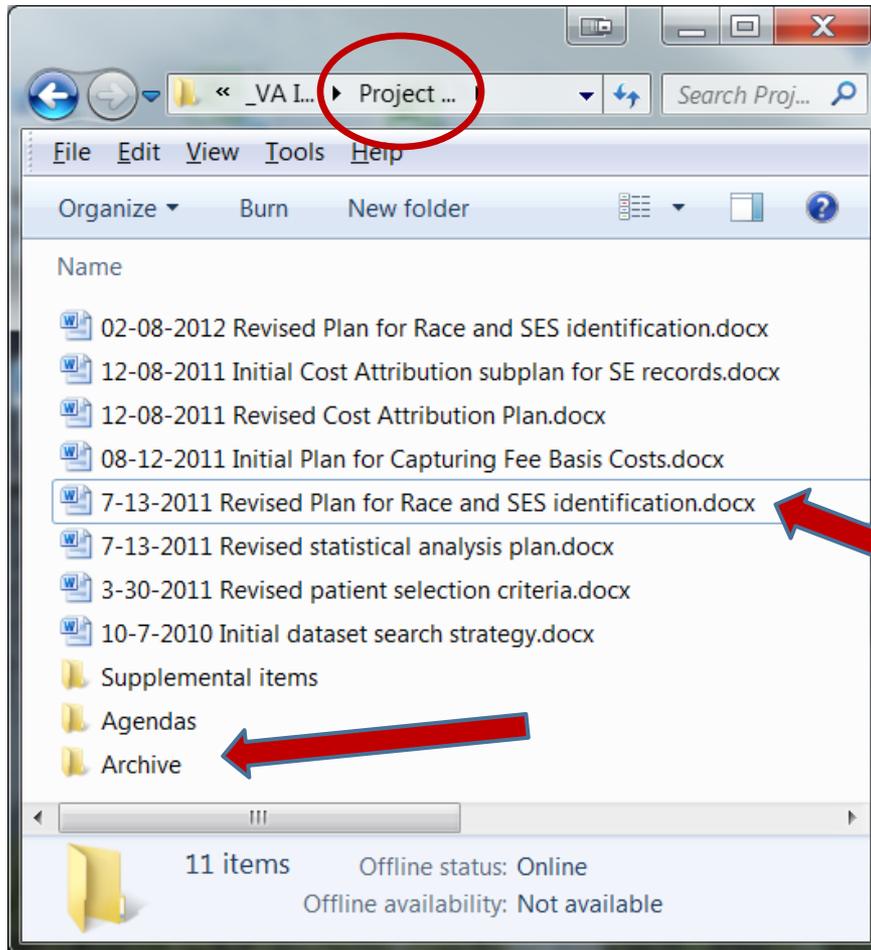
- All documents should have a date in their filename indicating the date on which the document is considered valid:
 - Cohort Selection_04.14.2014.docx
 - CohSel_04142014.sas
- Do not rely on system “save dates” to track date of file creation

Good practices for analytic plans: version control

Word documents should be linked to the immediate prior version using track changes



Versions and archives



Bad Practices for version control ...

- **Do NOT use number or letter schemes, which can easily result in sequence errors**
- **Example:**
 - Cohort_selection_A (save date 1/3/2014)
 - Cohort_selection_A1 (save date 1/5/2014)
 - Cohort_selection_B (save date 1/9/2014)
 - Cohort_selection_A2 (save date 1/15/2014)
 - Cohort_selection_A1-fixed (save date 1/17/2014)

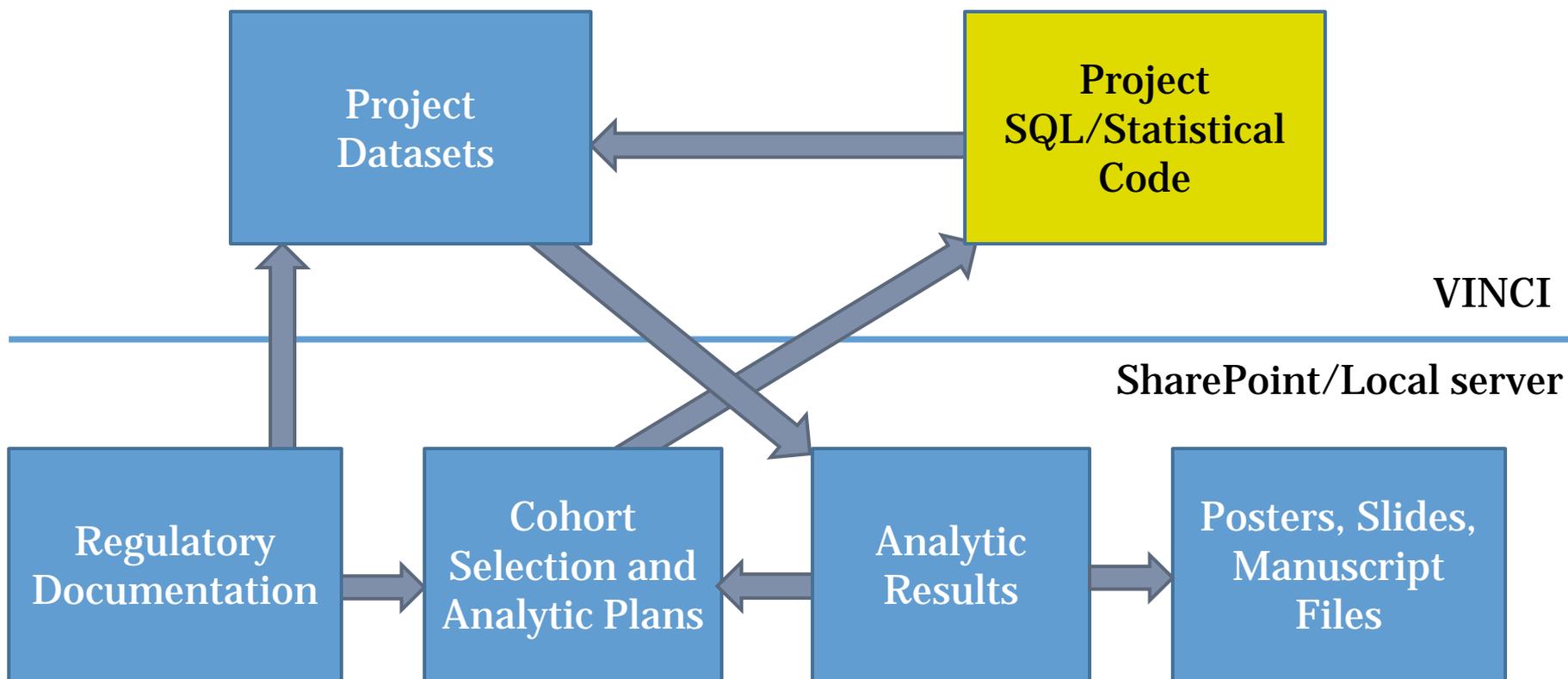
Good practices for analytic plans: recording decisions

- Changes in cohort selection or analytic plans should be recorded in a new document ideally within 24 hours of a research team meeting, teleconference, or email discussion
- All participants in the decision should be asked to review the new version of the document for accuracy, clarity, completeness

Good practices for analytic plans: adequate detail

- Documents should be sufficiently detailed that another investigator could replicate the efforts without much help
- Explain everything clearly to your “future self” who will have to decipher these documents while writing a scientific manuscript in the distant future

Project Data and Documentation Schema



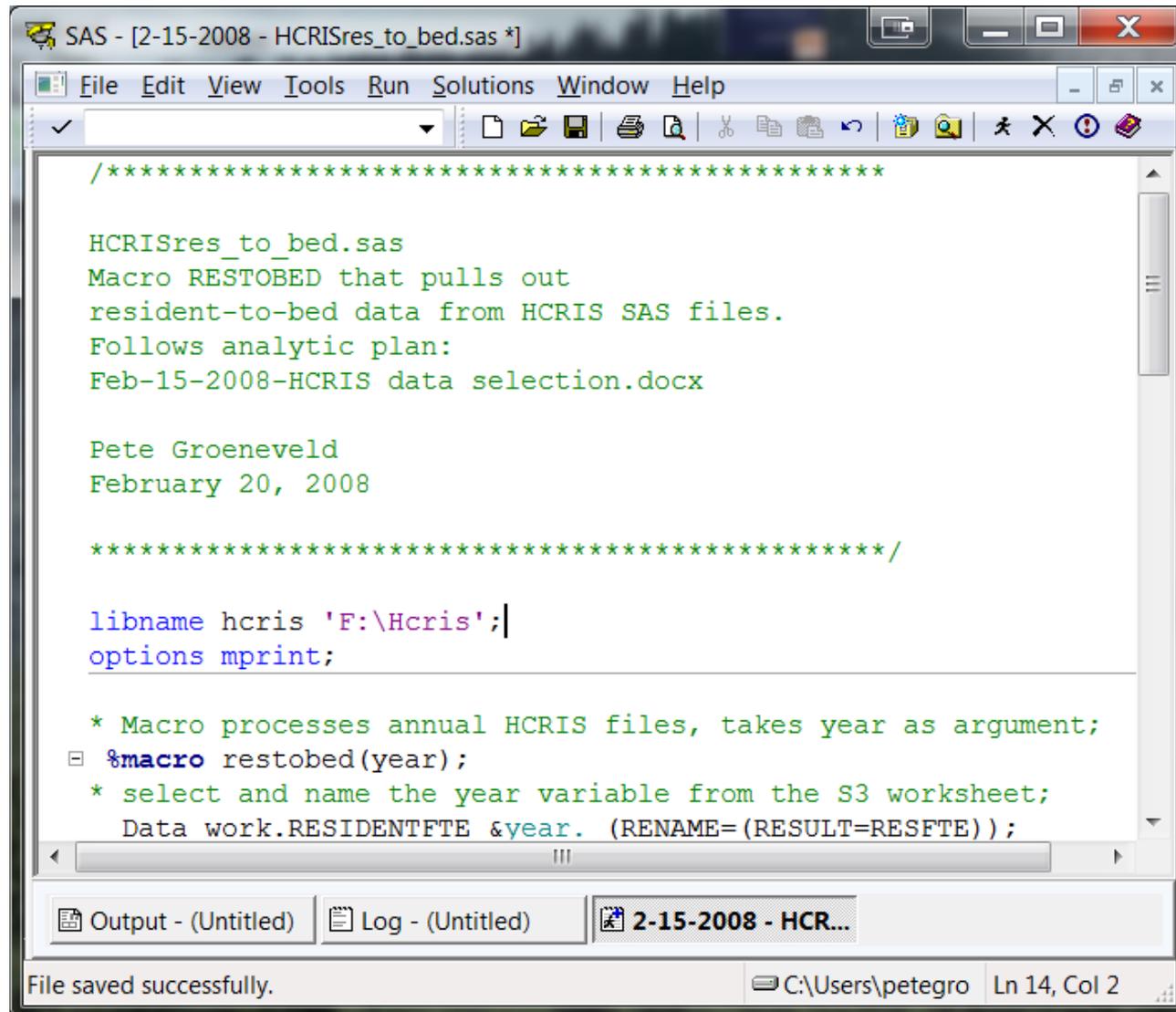
Good practices for analytic plans: SQL/statistical code file names

- A revised analytic plan should result in a newly named code file, rather than a “save/replace” of “analysis.sas”
- Names of the analytic plans, statistical code, and any output should be easily linkable
 - Cohort Selection_04.14.2014.docx
 - CohSel_04142014.sas
 - CohSel_04142014.lst

Good practices for analytic plans: SQL/statistical code

- As with all computer programming code, should be well commented, with any changes in code dated.
- Code should always have a header describing date, project, author, linking information to analytic plan document, and a general description of what the code does
- Code should be liberally interspersed with comments

Use of comments



The screenshot shows a SAS editor window titled "SAS - [2-15-2008 - HCRISres_to_bed.sas *]". The window contains the following text:

```

/*****
HCRISres_to_bed.sas
Macro RESTOBED that pulls out
resident-to-bed data from HCRIS SAS files.
Follows analytic plan:
Feb-15-2008-HCRIS data selection.docx

Pete Groeneveld
February 20, 2008

*****/

libname hcris 'F:\Hcris';
options mprint;

* Macro processes annual HCRIS files, takes year as argument;
%macro restobed(year);
* select and name the year variable from the S3 worksheet;
Data work.RESIDENTFTE &year. (RENAME=(RESULT=RESFTE));

```

The status bar at the bottom of the window displays "File saved successfully." and "C:\Users\petegro Ln 14, Col 2".

Coding an Analytic Plan: Risk!

- A key risk point: miscommunication or inadvertent coding errors by which the statistical code does not follow the analytic plan exactly
- Typical errors include duplicated observations, inadvertent keeps/drops, variable miscoding, model misspecification, etc.



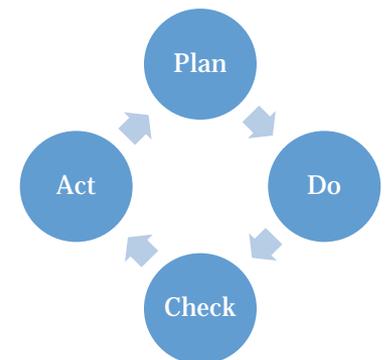
A Complex Process Requiring QI Tools

- Unfortunately SAS, STATA, and R won't produce an error message that the plan isn't being followed!
- This is a critical production process in research that requires similar QI techniques used elsewhere in VA to ensure quality of complex processes

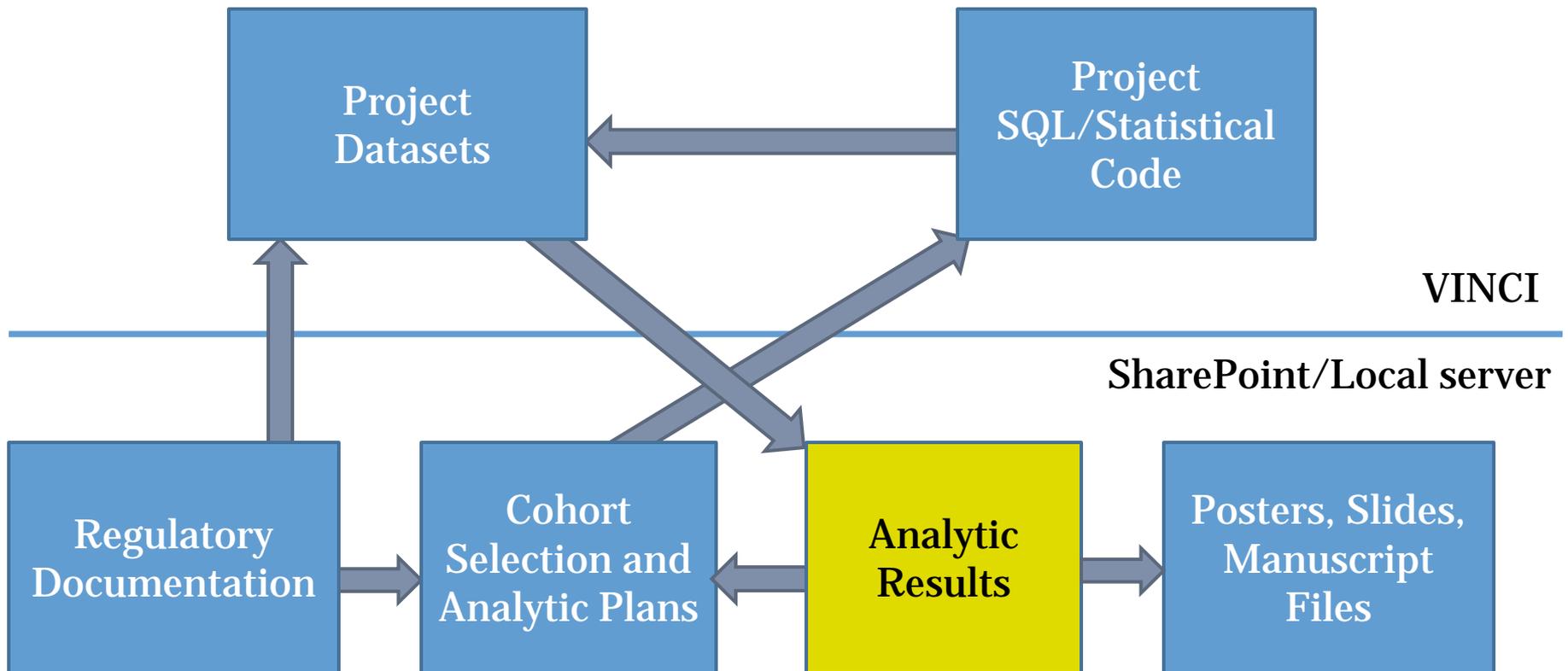


Solution: Data Walk-through

- Statistical programmer prints out all statistical code and output (e.g. in SAS, the .sas, .lst, and .log files) and “walks” through the code with a knowledgeable “second set of eyes”
- To avoid “ruffled feathers,” this should be explained by the PI as a routine quality control practice and applied to even your most experienced and meticulous programmers



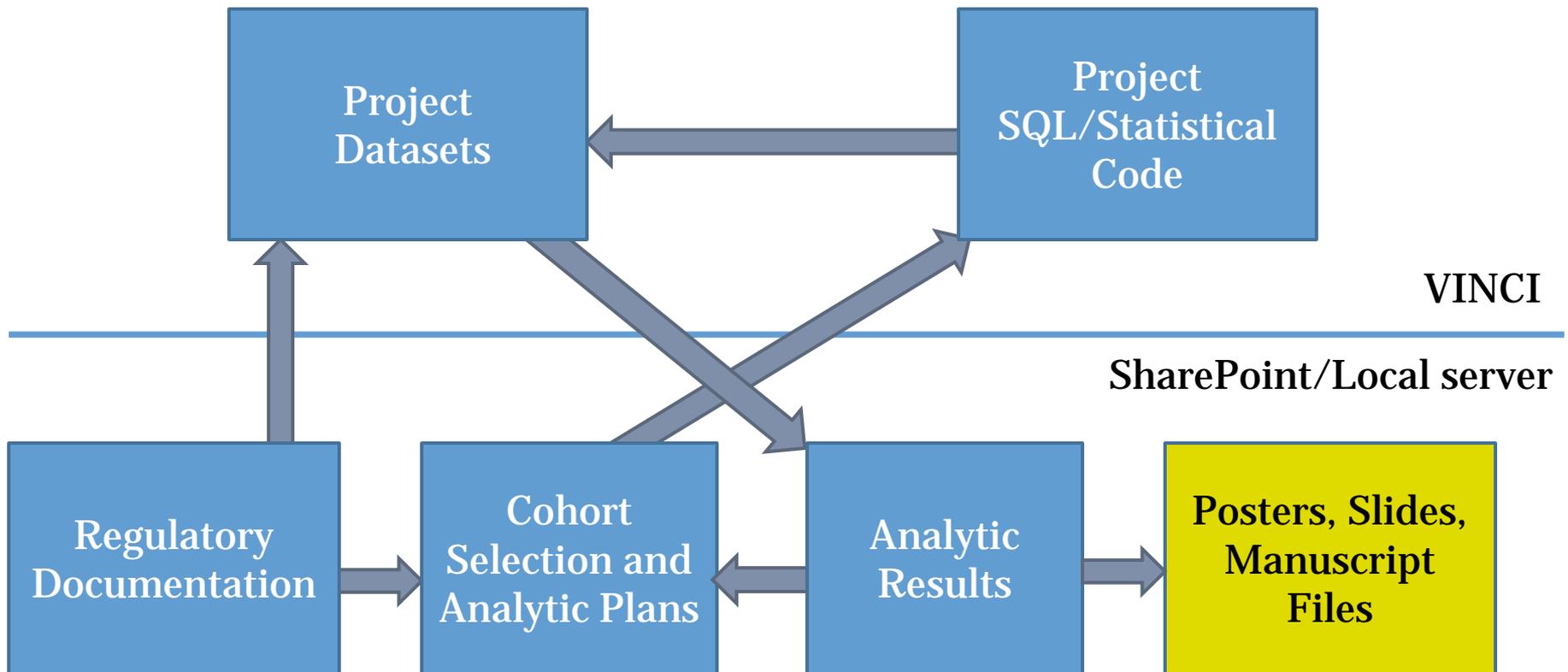
Project Data and Documentation Schema



Output files

- It should be clear what code file produced the output, and on from what versions of the data the output was created (i.e., good linkages)
- Annotated output (e.g., raw SAS output that has been modified with comments or edits) should be labeled with the date of annotation:
 - CohSel_04142014_Annot_04192014.lst

Project Data and Documentation Schema



Presentation files

- Each Table and Figure should be linked to the specific statistical output documents that are their source.
- Ideally do this via filenames
- Second choice is to include this information embedded in the file.
- In Powerpoint this info can go into “Speaker notes”

1-19-2014_CohSel.lst

comorbidity	count	percent
ANEMDEF	76045	15.0%
CAD	261393	51.6%
CANCER	50309	9.9%
CHRN LUNG	138558	27.3%
DM	148125	29.2%
DMCX	59336	11.7%
HTN_C	363345	71.7%
HYPOTHY	37464	7.3%
LIVER	12343	2.4%
METS	5151	1.0%
NEURO	26146	5.1%
PARA	8876	1.7%
PERIVASC	61780	12.2%
PULMCIRC	18047	3.5%
RENLFAIL	57439	11.3%
VALVE	48854	9.6%

Characteristic	N (%)
Female	10,595 (2)
Age, mean(std)	70 (11)
HTN	363,345 (72)
CAD	261,393 (52)
DM	207,461 (41)
Chronic pulm dz	138,558 (27)
Anemia	76,045 (15)
Periph vasc dz	61,780 (12)
Chronic kidney dz	57,439 (11)
Cancer (non-metastatic)	50,309 (10)
Valvular heart Dz	49,854 (10)
Hypothyroidism	37,464 (7)
Chronic neurological dz	26,146 (5)

2-14-2014-TABLE 1-FROM 1-19-2014_CohSel.lst

With a fully organized
document scheme in place ...

Use it wisely!



Referring to Documents in Email Communications

- Avoid email attachments, which can proliferate and cause version confusion
- Instead, refer to documents in email by their (possibly abbreviated) filepath:
 - “Team, please examine the updated data selection algorithm in
/analytics/data_selection_04.14.2014.docx”

Be specific about document names!

- **Wrong:**

- “Follow the most current algorithm ...”

- **Right:**

- “Follow the data selection algorithm in
/analytics/data_selection_04.14.2014.docx
- “Team, please examine the updated data selection
algorithm in
/analytics/data_selection_04.14.2014.docx”

Why organize?

Benefits!

- Writing and revising manuscripts should be much easier
- Workflow should be much clearer and easier to manage
- Hopefully the research team will need less direction from the PI—it will be more obvious to the entire team which steps are next

... and more benefits!

- Audits are ~~not~~ less terrifying
- Staff transitions will be smoother
- At the end, your project work may be “recyclable” for additional projects, rather than an indecipherable mess

Summary

- High quality study documentation is critical to good science, project management, and regulatory compliance
- Good practices are not terribly time consuming and should produce clear communication, organized files, and smooth workflow
- Productivity gains from these practices will typically be greater than any time costs

Questions?

Contact Information

Peter W. Groeneveld, MD, MS

VA Center for Health Equity Research and Promotion

Philadelphia VA Medical Center

peter.groeneveld@va.gov

Session 4: Reduce, Reuse, Recycle: Planning for Data Sharing

- Typical project close activities
- Why re-use research data?
- Policy and requirements for re-use
- Planning for re-use
- Documentation to make re-use possible