Good Data Practices
VIReC Cyberseminar Series

May 2014

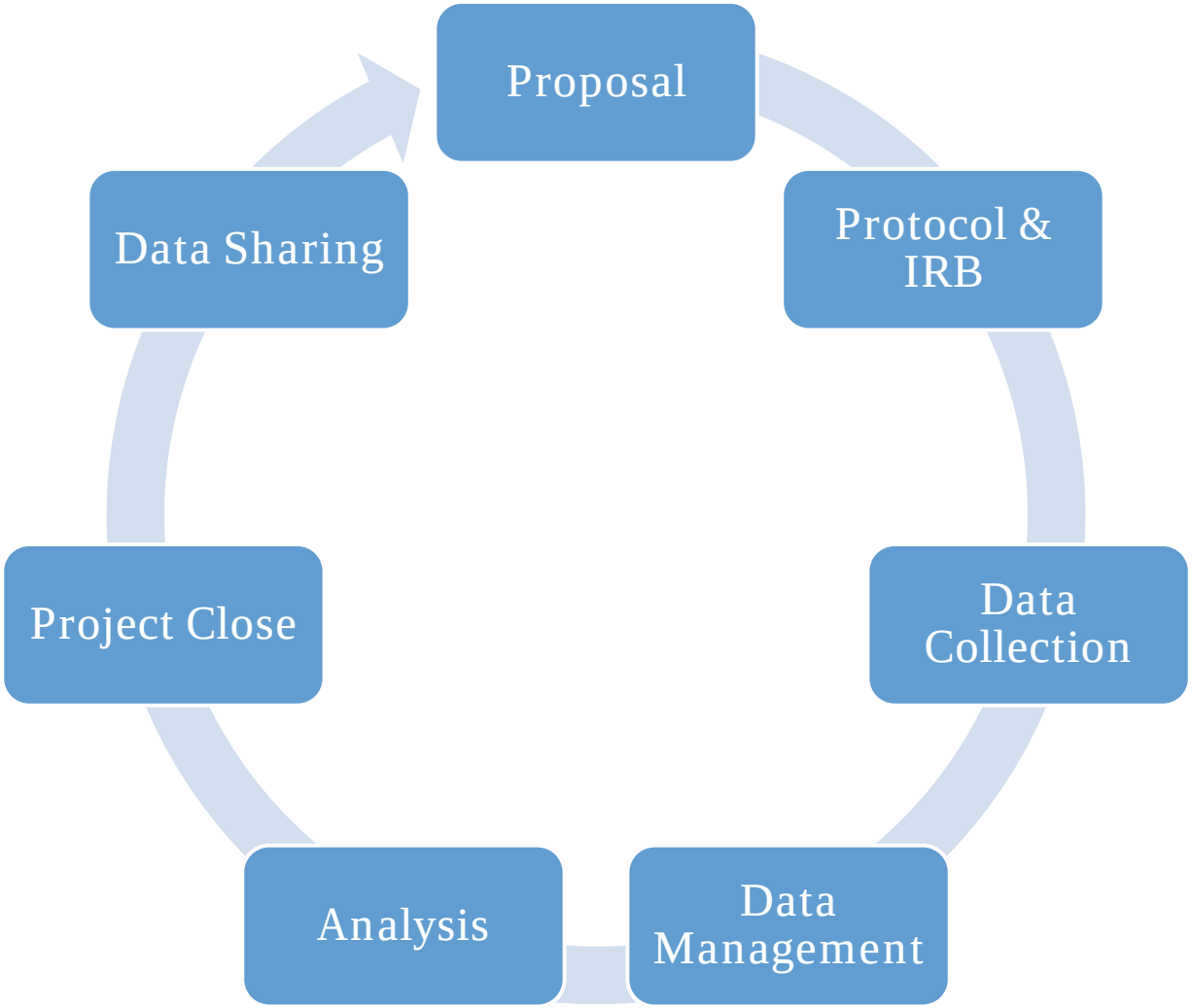
VIReC
RESEARCHERS' GUIDE TO VA DATA

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

VIReC
RESEARCHERS' GUIDE TO VA DATA

Research Life Cycle



VIReC
RESEARCHERS' GUIDE TO VA DATA

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**

VIReC
RESEARCHERS' GUIDE TO VA DATA

Poll #1: 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

Session 2:
"The Living Protocol" – Managing Documentation While Managing Data

Matthew Maciejewski , PhD
Durham VA Medical Center

VIReC
RESEARCHERS' GUIDE TO VA DATA

Good Data Practices 2014

Acronyms & Abbreviations

- BOSS: Bariatric Outcomes Surgery Study
- DSS: Decision Support System
- CDW: Corporate Data Warehouse
- HERC: Health Economics Resource Center
- COMM: Continuity of Medication Management
- WOC: Without Compensation
- VA: Veterans Affairs
- VASQIP: VA Surgical Quality Improvement Program
- OPC: Outpatient Care File
- PTF: Patient Treatment File

VIReC
RESEARCHERS' GUIDE TO VA DATA

Session 2 Outline

- Observation, objective & bottom line
- Living Protocol
- Example: Managing secondary data
- Example: Managing linkage of primary & secondary data
- Conclusions: The value of documentation

Observation, objective & bottom line

Observation:

Little guidance in graduate school, no guidance during research career and no literature about how best to…

- Work with a team to operationalize a protocol
- Prioritize the order of tasks
- Document data…

Objective:

- Share examples of conducting these tasks in timely manner after….
  - Trial and error
  - Begging, borrowing and stealing best practices from other investigators

Bottom Line:

- Clear & complete documentation enables…
  - Easy recall for manuscript writing
  - Application of components to future work, so you don't reinvent the wheel

Easy recall

Ultimately, you will write a paper presenting methods (design, measurement, analytics) and results

- Documentation is the <u>only</u> source for the logic of your choices
  - Easy way: Document as you go along so you have a **"living protocol"**
  - Hard way (subject to recall bias): Go back to minutes, programmer, code & scribbled notes

Documenting as you go along can save tons of trouble

- This project: How do you write the methods when your staff is gone & you haven't done all the programming?
- Next project: What if you want to re-use ways of defining cohort or measuring certain outcomes & covariates, but can't remember why or how?

Poll Question #2: How do you document?

- How do you document the major data decisions in your study?
  - We don't.
  - We use minutes to document major decisions
  - We create email summaries that are retained
  - We amend protocol for use by team to reflect major decisions
  - Other – add your process to Q& A

The Living Protocol

Genesis of "Living protocol"

- Trying to re-establish good project management practices that are also replicable across projects
- Improving depth and breadth of documentation

Serves multiple purposes

- For me: refer back to alternative options and reasoning for specific choices in design, measurement & analysis in future work
- For colleagues: may provide an example for them to improve upon, which they can improve and then give back to me

Study team and data management workflow

- My general process of working with study team
- Composition of bariatric study team
  - Clinical: 2 general internists & 2 bariatric surgeons
  - Non-clinical: 3 methods folks, 1 data analyst, 1 coordinator
- Everyone is free to contribute to the "Living protocol"

Planning for documentation of study design & measurement

- Working with my study team
  - Calls every other week, unless special calls needed to work through data/methods issue
  - Structure of each call
    - Update on data & programming
    - Review recent article to stay up on literature
    - Get into weeds on current issue to make decisions
    - Outline next steps

How the team evolves the "Living protocol"

- With team on bimonthly calls with set structure
  - Update on data & programming
  - Review recent article to stay up on literature
  - Get into weeds on current issue to make decisions for protocol
  - Outline next steps
- Special calls (rarely) to work through particularly tricky issues

Process between bimonthly calls

- Keep track of tasks via minutes and protocol
- Update master protocol to make it a <u>living document</u> to reflect decisions just settled
  - Iterative process, mostly before we get into data
  - Challenge for me: Making sure prior sections are completely current
- Programmer generates code from protocol after validating the protocol and identifying errors

Major sections of “Living protocol”

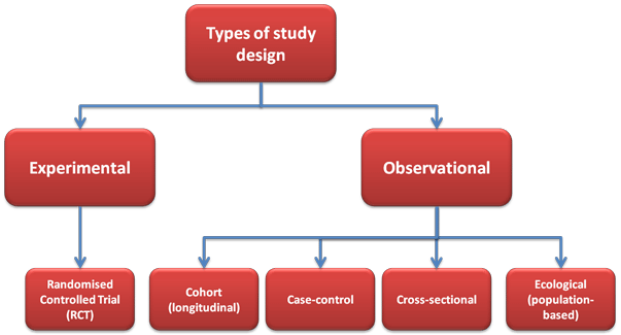
- Design of study
- Measurement
- Analysis

Example: Study Design Summary

- Start of documentation: Overview of study design and major threats to internal validity
  - "We will be conducting a retrospective pre-post cohort study with non-equivalent controls made up of veterans who were *eligible* (as best we can determine) for bariatric surgery in VA but never had surgery.
  - This is a pre-post cohort study because we observe all surgical patients before and after they had bariatric surgery in VA, and all non-surgical controls before they had surgery.
  - The non-equivalence of the non-surgical controls will be reduced via sequential stratification."

Study design issues

- Ideal vs. What we actually are stuck with
- Internal validity threats as a result of study design
- Cohort derivation
  - Inclusion/exclusion criteria for all patients
  - Steps for identifying and refining treatment group, controls
  - Consort figure and justification for exclusions



Measurement issues

- Data sets: variables contained in each
- Outcomes: definition, data cleaning, frequency distributions
- Covariates for all aims
  - Definition (ICD-9, CPT-4, recoding) in detailed tables that can be re-used/shared and don't solely reside in SAS code
  - Specify purpose (inclusion/exclusion, matching, adjustment)

Analysis issues

- Analytics
  - Overall: missing data, descriptive statistics
  - Internal validity threats that we will address via statistics
  - Aim-by-aim methods
- In theory, all of this should be done before we start analysis
  - Ultimate test: Can you complete it, so it could be posted on website (like Oregon health insurance experiment group)?

Parallel documentation of SAS programs

- Critical
- Each program should have
  - Clear statement of purpose
  - Relate to major aims of grant

Example: Managing secondary data

Planning for documentation of study design and measurement

- Bariatric Surgery's Return on Investment for Veterans and VHA (IIR 10-159)

Aims of Bariatric Study

- Compare veterans who did and did not have surgery in 2000-2011…
  - Aim 1: Weight change and resolution of diabetes, hypertension, and hyperlipidemia
  - Aim 2: Long-term survival and major surgical complications
  - Aim 3: Long-term trends in VA health care utilization and VA expenditures

End Goal: Two datasets

Matching dataset

- Cohort of surgical cases and non-surgical controls satisfying inclusion/exclusion criteria
- Covariates to be used for matching

Outcomes dataset

- Cohort of matched surgical cases and non-surgical controls
- All outcomes, covariates used for matching & other covariates needed for adjustment

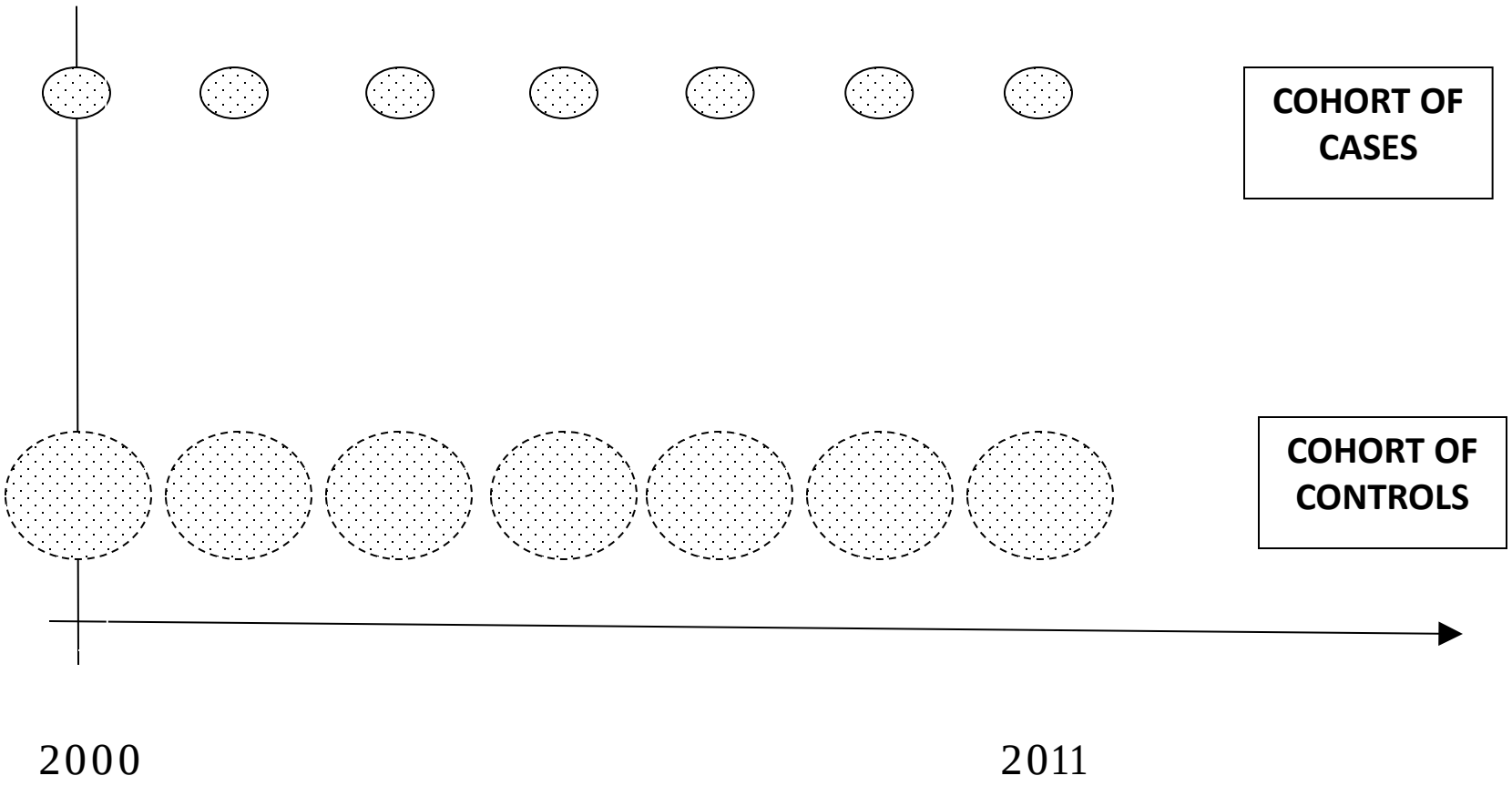
Dataset documentation and data cleaning

- Study-level documentation of dataset
  - What specific dataset?
  - What years?
  - What purpose?
- In each dataset, documentation of variables
  - What variables to be pulled?
  - What variables to be derived?

Plan for Pulling Multiple VA Claims Datasets

Dataset	Aim	Outcome To Create	Covariates To Create (in brief)	Years We Have on Cases	Years We Have on Controls
VASQIP	All	--	Surgery type	2000-2011	Not applicable
Fee Basis	All	--	Surgery type	2000-2011	Not applicable
Mini-Vitals	All, 2	Death	Age, gender	Most current	Most current
HERC	3	Cost		2000-2011	??
DSS LAR	All	Lab results	Baseline values for A1c, LDL	2000-2011	2000-2011
PBM	1	Disease control	Medications at baseline	2000-2011	2000-2011
CDW	All	BP, Weight Δ	Baseline value of BP, BMI	2000-2011	2000-2011
OPC	All	Utilization, complications	Race, marital status, Dx-based covars (comorbidity)	2000-2011	In process (covariate & exclusions)
PTF	All	Utilization, complications	Dx-based covars	2000-2011	In process (covariate & exclusions)
DCG	All	--	DCG risk score	2000-2011	2000-2011
Enrollment	All	--	Copay status (from Priority St)	2000-2011	2000-2011

Visual representation of sampling frame



Study Design Visualized

VIReC
RESEARCHERS' GUIDE TO VA DATA

Cohort identification: Inclusion & exclusion

- Process for coming up with list of inclusion & exclusion criteria after initially identifying surgical patients and controls
  - Reviewed prior RCTs and observational studies
  - Reviewed our prior work
  - Reviewed ongoing trials
- Reviewed list of criteria as a group and refined
- Once criteria listed, then developed coding rules

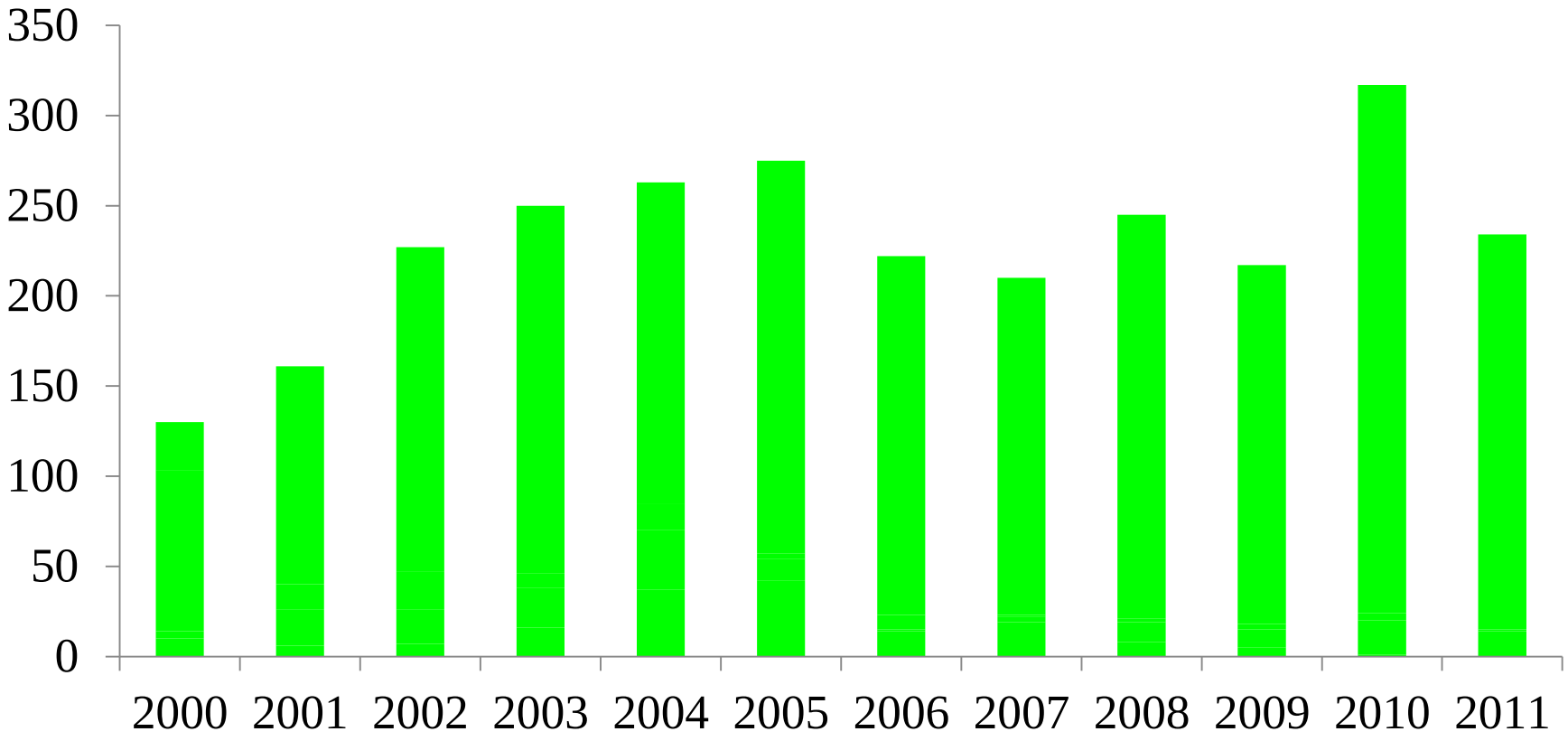
Identification of treatment group

- What datasets to use for identification?
- What codes for identification
  - CPT-4 procedure?
  - ICD-9 procedure?
  - Medication?
- Do identification (coding) rules change over time?
  - If so, how to validate that we've minimized errors of omission and commission?

Construction of surgical cohort: Identifying Surgical Patients

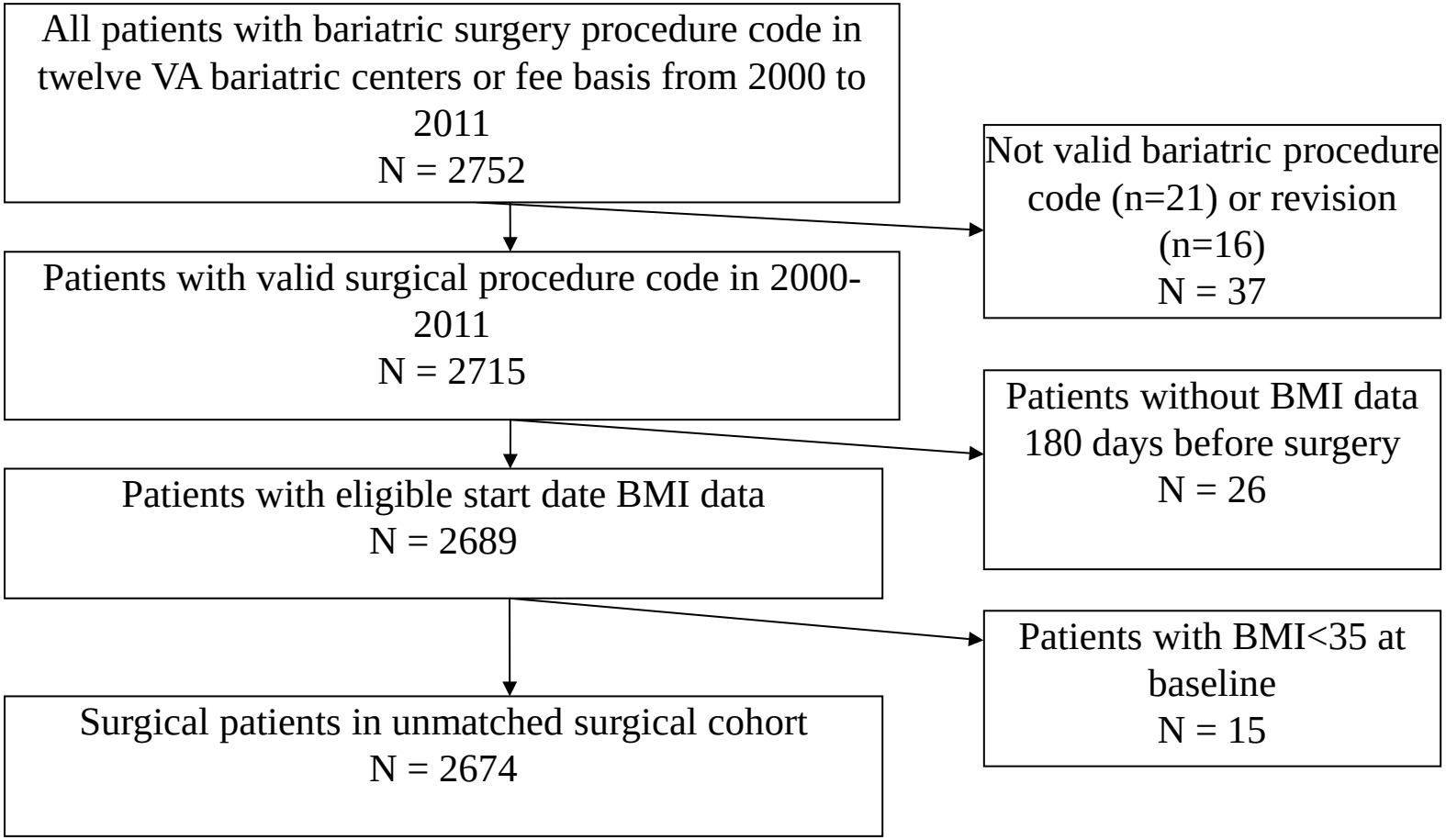
Surgical Procedure	*ICD-9 Procedure Coding*	*CPT-4 Procedure Coding*
RYGB, Open	44.31, 44.39	43621, 43846, 43847
RYGB, Laparoscopic	44.38	43644, 43645
VBG	44.68	43842
AGB, LAGB	44.69, 44.95	43770, 43843
Sleeve		43775
BPD	43.7, 45.91	43633, 43845
Unclear (could be BPD, sleeve)	43.89	43659
Surgical revision (will exclude)		43771, 43772, 43774, 43848
Not bariatric surgery?		43860, 43999, 44180
Revision		43771, 43772, 43773, 43774, 43848

Time trends in surgical volume



VIReC
RESEARCHERS' GUIDE TO VA DATA

Consort figure of surgical cohort



Initial Consort Figure: I do this as early as possible

Construction of surgical cohort

- Defining and documenting alternative index dates
  - Index date serves several purposes
    - Date of treatment (or not)
    - Differentiates pre-period from post-period
    - Many covariates conditional on index date
      - Baseline covariates
  - In observational studies, timing of measurements for baseline covariates likely to vary across patients
    - BMI, BP, A1c, LDL from clinic measurements that happen whenever they happen, not at fixed intervals preferred by researcher as in RCT

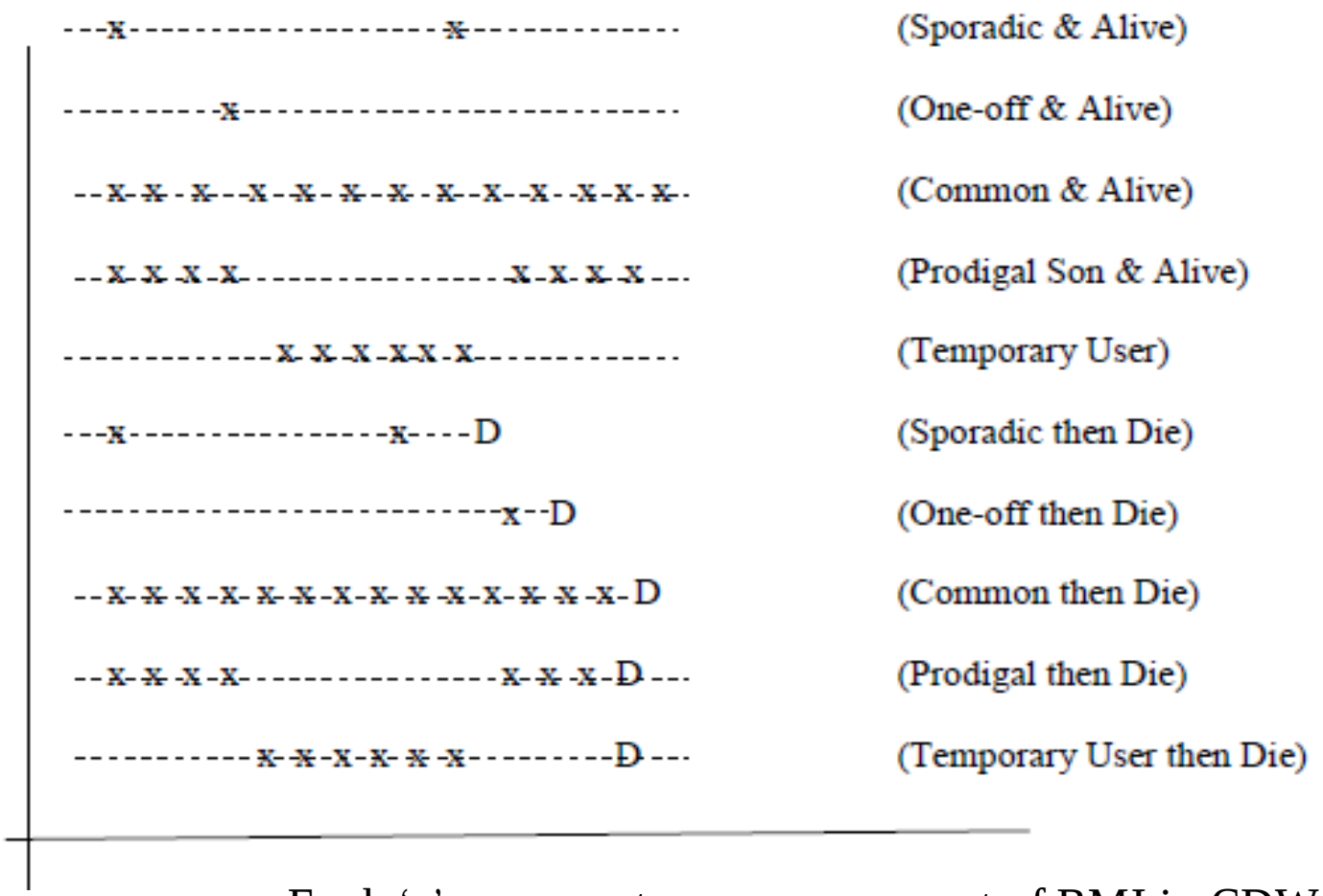
Timing of "Baseline" BMI on Surgical Patients

Description of BMI Data Available	# Cases
BMI data available on day of surgery	1855
BMI data available 1-182 days before day of surgery	425
BMI data available on day of surgery from old study	15
BMI data available 183-672 days before day of surgery and set to missing (n=445) or BMI data unavailable from any source (n=12)	457

Construction of control cohort

- Non-surgical controls will be identified from CDW weights, so are likely to have multiple weights over multiple years
- We will need to develop a strategy for choosing which weight will make them eligible
- BMI measurement patterns we are likely to observe in our large non-surgical cohort from 2000-2001 CDW are represented on the next table.

Identifying non-surgical controls



Each ‘x’ represents a measurement of BMI in CDW data.
Death is represented by a ‘D’.

Explicit statement of purpose for covariates

- Four purposes for variables
  - Inclusion/exclusion criteria
  - Propensity score model of treatment selection
  - Covariates for outcome model
  - Outcome itself

Outcomes and covariate construction

- Covariate Selection informed by theory
  - Directed acyclic graph (DAG) development with entire team before we saw any data
    - Informed by prior literature, our prior work
    - Tells us what confounders we don't observe

Specific outcomes to be documented

- Living Protocol included definitions & specific coding rules
  - Survival: Time from surgery until death
  - Post-surgical complications
  - Weight: Before and after surgery
  - Remission of disease: DM, HTN and dyslipidemia
  - VA health care utilization
  - VA expenditures

Outcomes and covariate construction

	Inclusion & Exclusion	*Matching Model*	OUTCOMES ANALYSIS *Weight Change*	*Disease Resolution*	*Survival*	*Utilization or Cost*
Indicator of surgery or not			√	√	√	√
Fiscal year of start time						
Age	√	√	√	√	√	√
Male		√	√	√	√	√
Caucasian		√	√	√	√	√
Non-Caucasian		√	√	√	√	√
Unknown Race		√	√	√	√	√
Married			√	√	√	√
Not Married			√	√	√	√
Unknown Marital Status			√	√	√	√
Copay status			√	√	√	√
VISN		√				√
BMI at baseline	√	√	√	√	√	√
Distance to closest VAMC					√	√

Inclusion & Exclusion of Bariatric Cases & Controls (Subset of larger table)

VIReC RESEARCHERS' GUIDE TO VA DATA

Covariate construction

- Once DAG created, how did we choose between alternative measures of a construct? For example, comorbidity
  - Criteria for choosing: Clinical interpretability, predictive power, what is used in related studies, what we did in the past

Covariate construction and choices

- How did we choose between alternative measures of a construct?
  - When construct lacks clarity? For example, distance to nearest VA or relative distance
  - When construct is clear but there are alternative versions of a specific measure to choose from? For example, marital status or race
    - Which value to take in a year? First, last, modal?
    - Time-invariant or time-varying?
- Criteria for choosing: What makes sense conceptually, what is used in related studies, what we did in the past

Documentation of provenance of covariates

Name of Covariate	*Data Source*	*Binary or Continuous*	*Definition for Surgical Pts*	*Definition for Controls*	*Time-Varying?*
Index date indicating end of pre-period and start of post-period • This defines the baseline	VASQIP, Fee Basis, OPC, PTF	Continuous	Day of surgery	Day of surgery for surgical patient he/she paired with	No
Diabetes at baseline	OPC, PTF	Binary	ICD-9 250.x, 357.2, 366.41, 362.01-362.07		No
Hypertension Dx at baseline	OPC, PTF	Binary	ICD-9 401.x – 404.x		No
Dyslipidemia Dx at baseline**	OPC, PTF	Binary	ICD-9 272.0, 272.1, 272.2, 272.3, 272.4		No
Sleep apnea Dx at baseline***	OPC, PTF	Binary	ICD-9 327.20, 327.21, 327.23, 327.27, 327.29, 780.51, 780.53, 780.57, 786.03		No

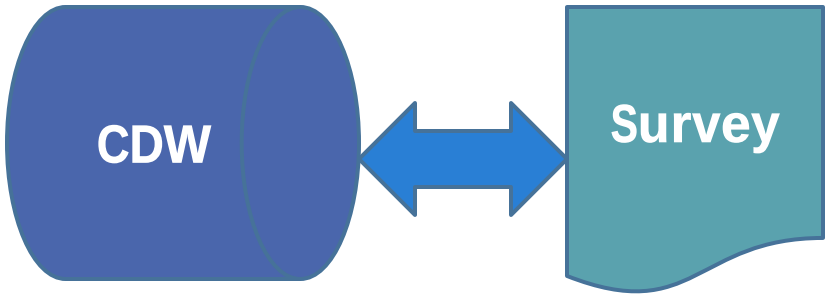
VIReC RESEARCHERS' GUIDE TO VA DATA

Example: Covariate Construction

Example: Managing linkage of primary & secondary data

Linking patient survey data with VA claims data

- Example from AHRQ-funded R21 (COMM)
- Purpose of doing patient survey
  - Examine outcomes not available in VA claims
  - Obtain covariates not available in VA claims



Issues related to surveys

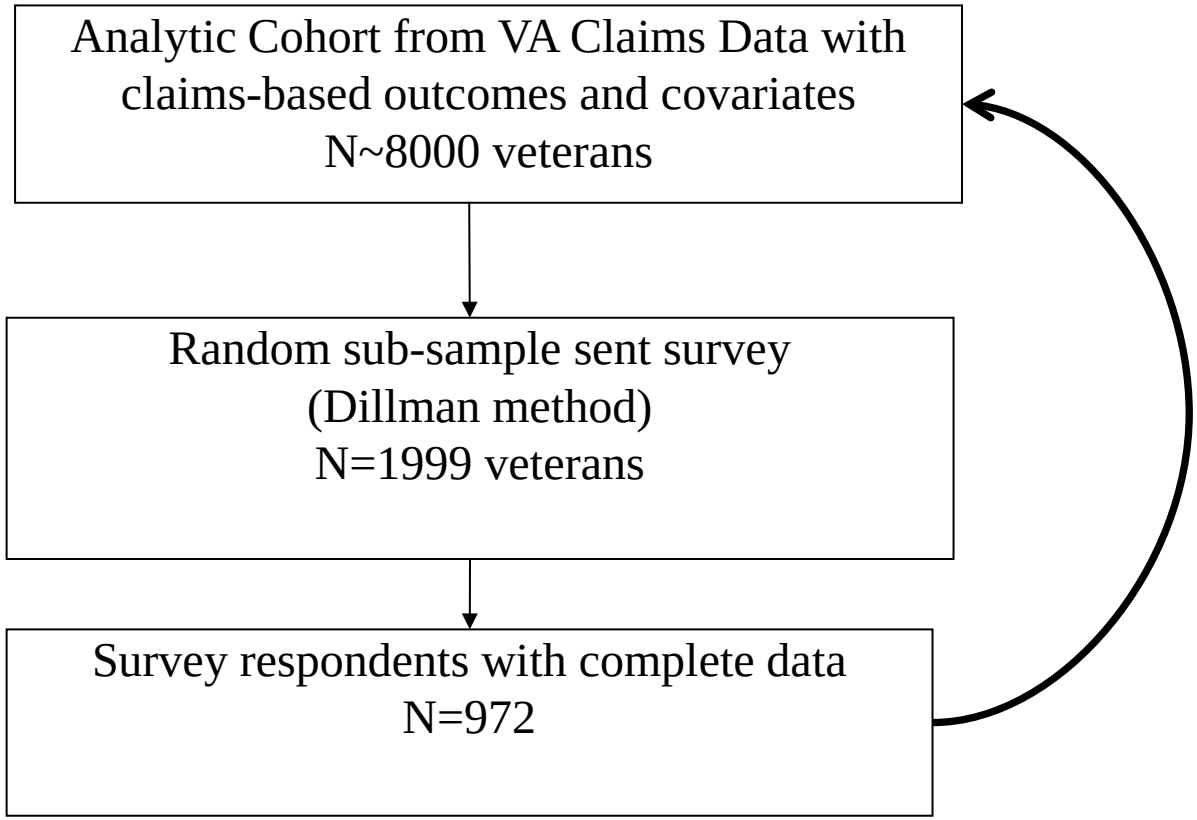
- If doing de novo survey, need to work with Office of Management and Budget (OMB) early because process is VERY slow
  - Required to get OMB approval if 10+ veterans are surveyed due to the Paperwork Reduction Act
  - Need to be aware of existing VA surveys to justify why your survey is not duplicative effort for veterans

Linking patient survey data with VA claims data

- Surveys require a 3rd patient identifier
  - Scrambled SSN (routine)
  - Unique study ID (routine)
  - Unique survey study ID
- Two processes for coding surveys
  - All in-house by VA staff
  - Contract with university-based staff who get without compensation (WOC) appointments (via RedCap)

Linkage of primary and secondary data

Data Flow



VIReC
RESEARCHERS' GUIDE TO VA DATA

Analyses enabled by linkage of survey + claims data

- Association between survey-based covariates and claims-based outcome
  - Do survey-based covariates improve predictive power compared to model comprised only of claims-based covariates?
- Association between claims-based covariates and survey-based outcome
  - Can examine outcomes not possible with claims alone

Conclusions: The value of documentation

Conclusions

- Things we are not doing that we should
  - Documenting flow of programs in data cleaning, data construction and sample construction
  - Updating "living protocol" in a timely manner

Conclusions

- To be useful, data documentation needs to be an iterative process
  - It is time consuming but it is the project's only historical record
- If done well, it can provide a comprehensive guide to your study for people new to project
- May be a useful source for guidance in future projects

Contact Information

Matthew Maciejewski, PhD

Health Economics and Policy Unit in the Center for Health Services Research in Primary Care

Durham VA Medical Center

matthew.maciejewski@va.gov

VIReC
RESEARCHERS' GUIDE TO VA DATA

Questions

Next Week in GDP…

May 22

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

Agenda

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

VIReC
RESEARCHERS' GUIDE TO VA DATA