

Good Data Practices

- Series Recap
 - Session 1: Early Data Planning for Research
 - Session 2: Managing and Documenting Data Workflow
 - Session 3: Planning for Data Re-use
 - Session 4: Research Application

Session 5: Research Application

- Planning for documentation of study design & measurement
- Data cleaning
- Construction of cohort
- Outcomes construction
- Covariate construction
- Linkage of primary (survey) data & VA secondary data
- Summary: Value of documentation (what worked, what didn't)

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

Observation & Objective Motivating this Lecture

- Observation: Little guidance in graduate school 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 efficient timely manner after begging, borrowing and stealing best practices from other investigators

Bottom Line about the Value of Documentation

- Ultimately, you will write a paper presenting methods and results
- Methods require reporting definition of outcomes, treatment group, control group, statistical analysis, sensitivity analyses
 - Documentation is the only source for the logic of your choices
 - Easy way: Document as you go along
 - Hard way (subject to recall bias): Go back to minutes, programmer, code & scribbled notes
- Documenting as you go along can save tons of trouble
 - What if all of your staff is gone & you haven't done all the programming?

Poll Question

- How do you document the major decisions in your study that relies entirely on claims data?
 - We don't. Once we get IRB approval for my protocol, I just get to work
 - We use minutes to document major decisions
 - We amend protocol for use by team to reflect major decisions
 - Other

Session 5: Outline

- **Planning for documentation of study design & measurement**
- Data cleaning
- Construction of cohort
- Outcomes construction
- Covariate construction
- Linkage of primary (survey) data & VA secondary data
- Summary: Value of documentation (what worked, what didn't)

Planning for documentation of study design & measurement

- 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



Planning for documentation of study design & measurement

- My general process of working with study team
 - Have 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

Planning for documentation of study design & measurement

- My process between calls
 - Keep track of tasks via minutes and protocol
 - Regularly update master protocol to make it a living document
 - Iterative process
 - Not great about making sure prior sections are completely current
 - Programmer generates code from protocol after validating the protocol and identifying errors

Planning for documentation of study design & measurement

- 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



Planning for documentation of study design & measurement

- Two Datasets Ultimately Needed
 - 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

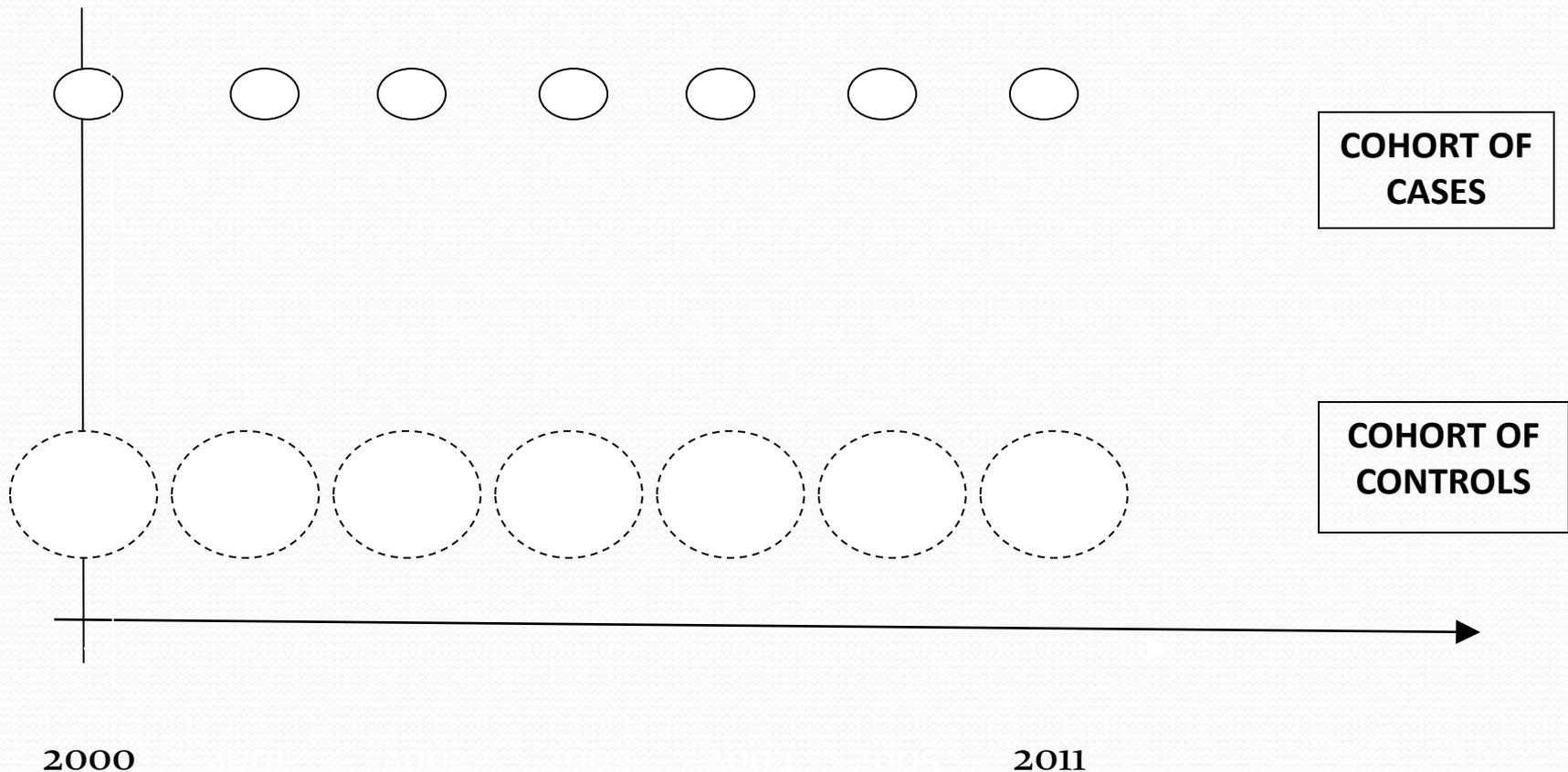
Planning for documentation of study design & measurement

- Outcomes to be Analyzed
 - 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

Planning for documentation of study design & measurement

- Start of Documentation: Overview of Study Design and Major Design Challenges
 - “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.”

Planning for documentation of study design & measurement



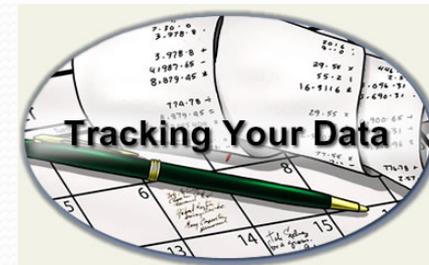
Study Design Visualized

Session 5: Outline

- Planning for documentation of study design & measurement
- **Data cleaning**
- Construction of cohort
- Outcomes construction
- Covariate construction
- Linkage of primary (survey) data & VA secondary data
- Summary: Value of documentation (what worked, what didn't)

Data cleaning

- Data tracking
 - Two levels
 - Study-level
 - What specific dataset?
 - What years?
 - What purpose?
 - In each dataset
 - What variables to be pulled?
 - What variables to be derived?



Data cleaning

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

Data cleaning

- 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?

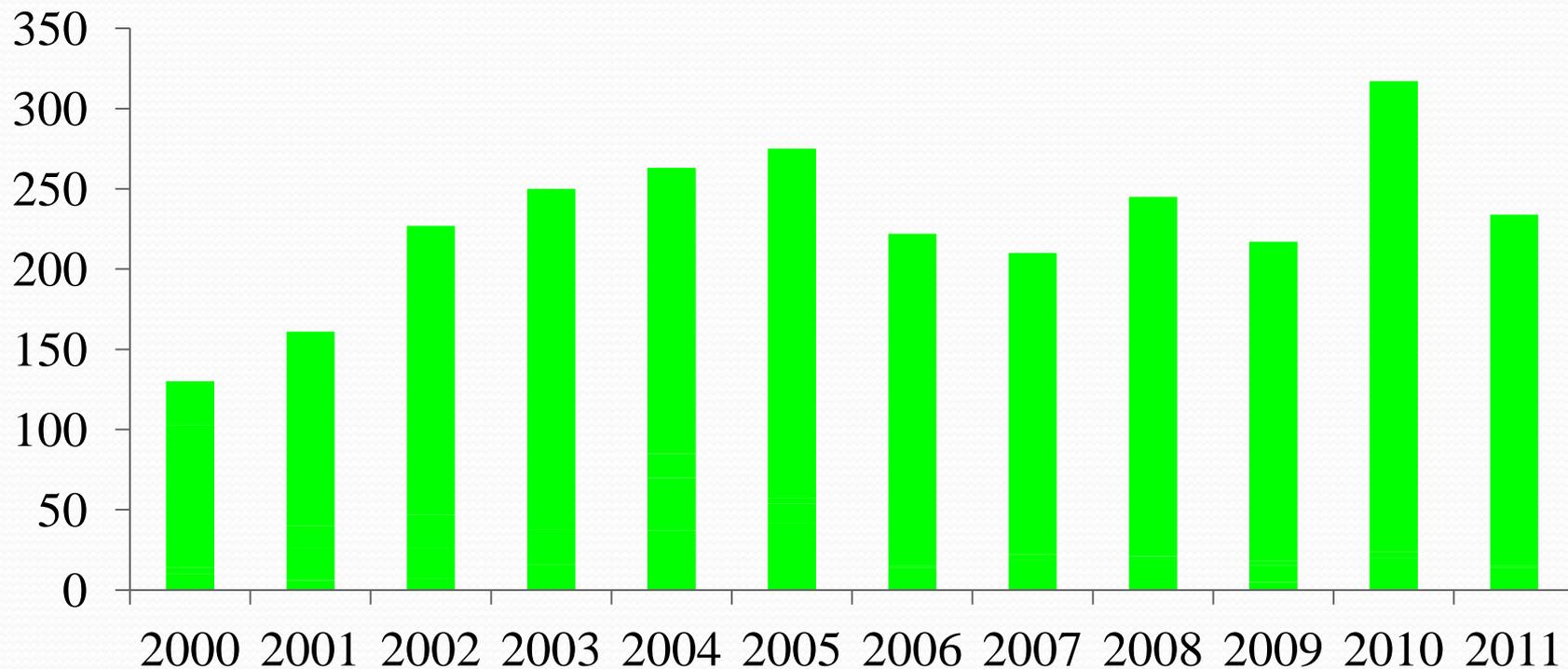
Session 5: Outline

- Planning for documentation of study design & measurement
- Data cleaning
- **Construction of cohort**
- Outcomes construction
- Covariate construction
- Linkage of primary (survey) data & VA secondary data
- Summary: Value of documentation (what worked, what didn't)

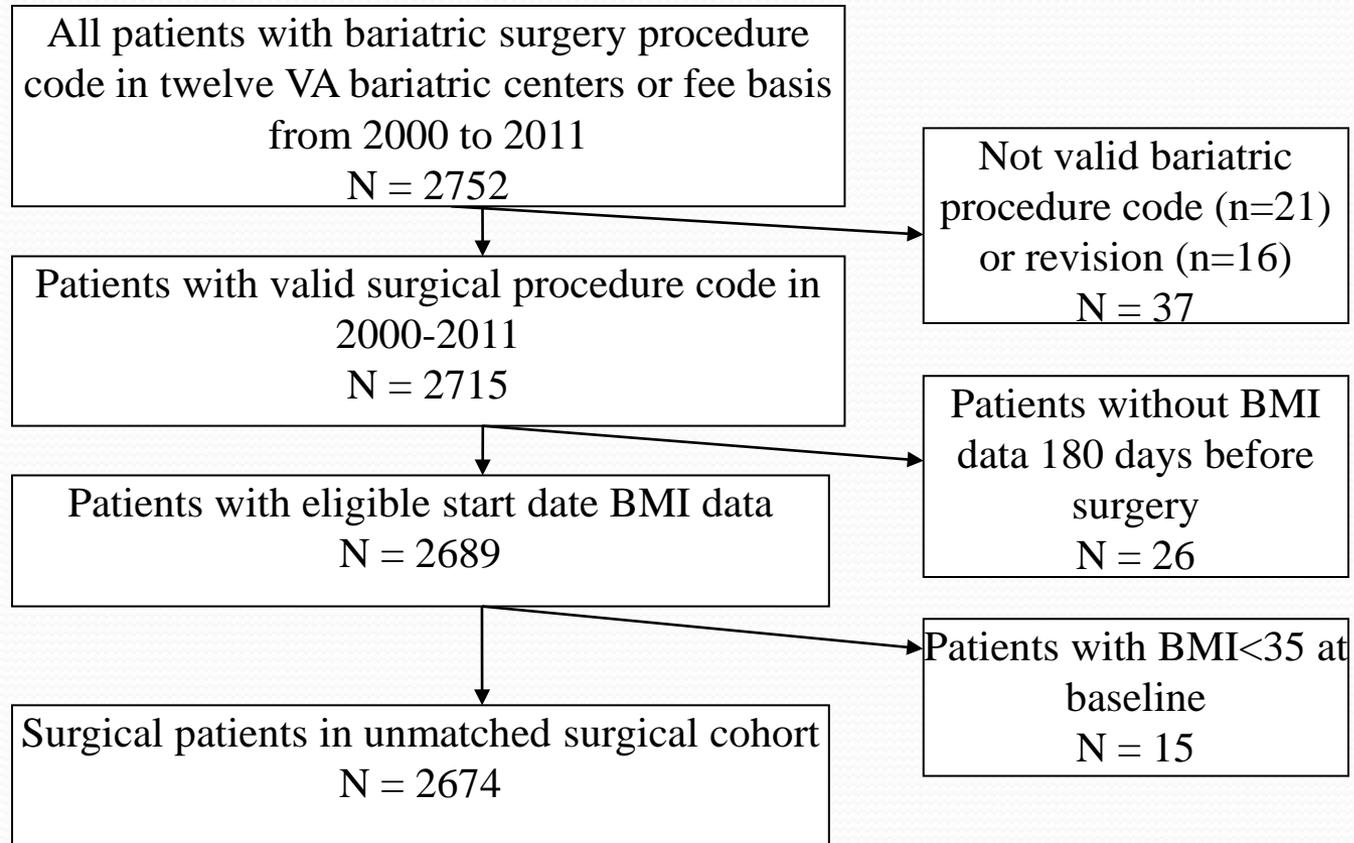
Construction of cohort

<i>Surgical Procedure</i>	<i>ICD-9 Procedure Coding</i>	<i>CPT-4 Procedure Coding</i>
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

Construction of cohort



Construction of cohort



Initial Cohort Figure: I do this as early as possible

Construction of 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

Construction of cohort

<i>Description of BMI Data Available</i>	<i># Cases</i>
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 cohort

Since non-surgical controls will be identified from CDW weights and they are likely to have multiple weights over multiple years, we will need to develop a strategy for choosing which weight will make them eligible. The likely BMI measurement patterns we are likely to observe in our large non-surgical cohort from 2000-2011 CDW data are represented below. Each 'x' represents a measurement of BMI in CDW data. Death is represented by a 'D'.

---x-----x-----	(Sporadic & Alive)
-----x-----	(One-off & Alive)
--x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-x-	(Common & Alive)
--x-x-x-x-----x-x-x-x---	(Prodigal Son & Alive)
-----x-x-x-xx-x-----	(Temporary User)
---x-----x---D	(Sporadic then Die)
-----x--D	(One-off then Die)
--x-x-x-x-x-x-x-x-x-x-x-x-x-x-D	(Common then Die)
--x-x-x-x-----x-x-x-D---	(Prodigal then Die)
-----x-x-x-x-x-----D---	(Temporary User then Die)

Construction of cohort

- Cohort identification: Inclusion & Exclusion Criteria
 - 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 (Arterburn)
 - Reviewed list of criteria as a group and refined
 - Once criteria listed, then developed coding rules

Construction of cohort

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

Session 5: Outline

- Planning for documentation of study design & measurement
- Data cleaning
- Construction of cohort
- **Outcomes construction**
- **Covariate construction**
- Linkage of primary (survey) data & VA secondary data
- Summary: Value of documentation (what worked, what didn't)

Outcomes & covariate construction

	<i>Inclusion & Exclusion</i>	<i>Matching Model</i>	OUTCOMES ANALYSIS			
			<i>Weight Change</i>	<i>Disease Resolution</i>	<i>Survival</i>	<i>Utilization or Cost</i>
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					√	√

Outcomes & covariate construction

- Covariate construction
 - Directed acyclic graph (DAG) development with entire team before we saw any data
 - Informed by prior literature, our prior work

Outcomes & covariate construction

- 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
 - How did we choose between alternative measures of a construct? For example, distance to nearest VA or relative distance
 - Criteria for choosing: What makes sense conceptually, what is used in related studies, what we did in the past

Outcomes & covariate construction

- Covariate construction
 - How did we choose between alternative versions of a specific measure when there are multiple values?
 - For example, marital status
 - Which value to take in a year? First, last, modal?
 - Time-invariant or time-varying?

Outcomes & covariate construction

<i>Name of Covariate</i>	<i>Data Source</i>	<i>Binary or Continuous</i>	<i>Definition for Surgical Pts</i>	<i>Definition for Controls</i>	<i>Time-Varying?</i>
<u>Index date</u> indicating end of pre-period and start of post-period <ul style="list-style-type: none"> 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

Session 5: Outline

- Planning for documentation of study design & measurement
- Data cleaning
- Construction of cohort
 - Special issues with surgical patients & controls
 - Inclusion & exclusion
- Outcomes construction
- Covariate construction
- **Linkage of 1^o & secondary data**
- Summary: Value of documentation (what worked, what didn't)

Linkage of 1° & 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

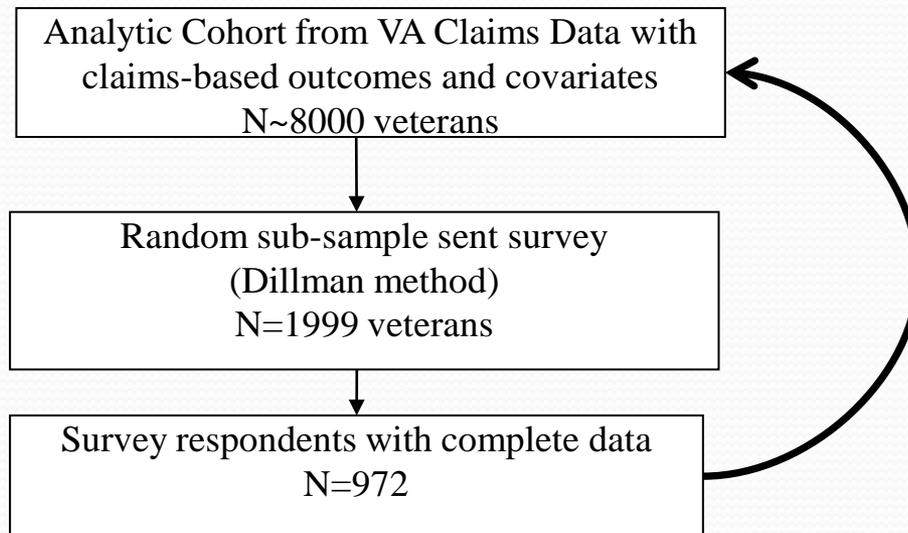
Linkage of 1° & secondary data

- Linking Patient Survey Data with VA Claims Data
 - 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

Linkage of 1^o & secondary data

- 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 1° & secondary data



Data Flow

Linkage of 1° & secondary data

- Linking Patient Survey Data with VA Claims 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

Session 5: Outline

- Planning for documentation of study design & measurement
- Data cleaning
- Construction of cohort
 - Special issues with surgical patients & controls
 - Inclusion & exclusion
- Outcomes construction
- Covariate construction
- Linkage of 1^o & secondary data
- **Summary: Value of documentation (what worked, what didn't)**

Value of documentation

- Conclusion: Things We are Not Doing that We Should
 - Documenting flow of programs in data cleaning, data construction and sample construction
 - Diligently updating master protocol

Value of documentation

- 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

Questions

VIReC@Va.gov

Thank You!