Research Design and Analysis

Kate Arnow, MS

Senior biostatistician, Stanford-Surgery Policy Improvement Research and Education Center

S-SPIRE Stanford–Surgery Policy Improvement Research & Education

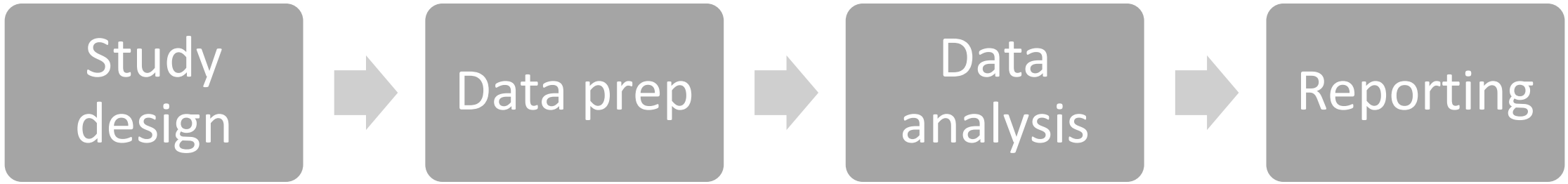
Questions

- Coding experience
- Master's program
- Active research projects--examples

Goals

- Good preparation is key!
- Available resources

Research Phases



Study design

- Define research question
- Align research question with appropriate study design, data source
- Calculate power/sample size
- Define variables and plan for measurement

Research Question (PICOT)

- **Patient population:** Condition / disease, demographics, setting
- **Intervention:** Procedure / policy / test / process intervention
- **Comparison group:** Controls (standard of care, non-exposed, medical management, no treatment)
- **Outcome:** Treatment effects, patient-reported outcomes, healthcare utilization
- **Timeframe:** study years, repeated measures, how long to experience the outcome

Example Research Question

- Do hospitals with 200+ beds perform better than smaller hospitals?

Example Research Question

- Do hospitals with 200+ beds perform better than smaller hospitals?
- More developed question
- Do large California hospitals with 200+ beds have lower surgical site infection rates for adults undergoing inpatient surgical procedures?
  - **P**opulation: California adults undergoing inpatient surgical procedures with general anesthesia
  - **I**ntervention (structural characteristic): 200+ beds
  - **C**omparison: smaller hospitals with <200 beds
  - **O**utcome: surgical site infections within 30 days post-op
  - **T**imeframe: 2017

Other planning questions

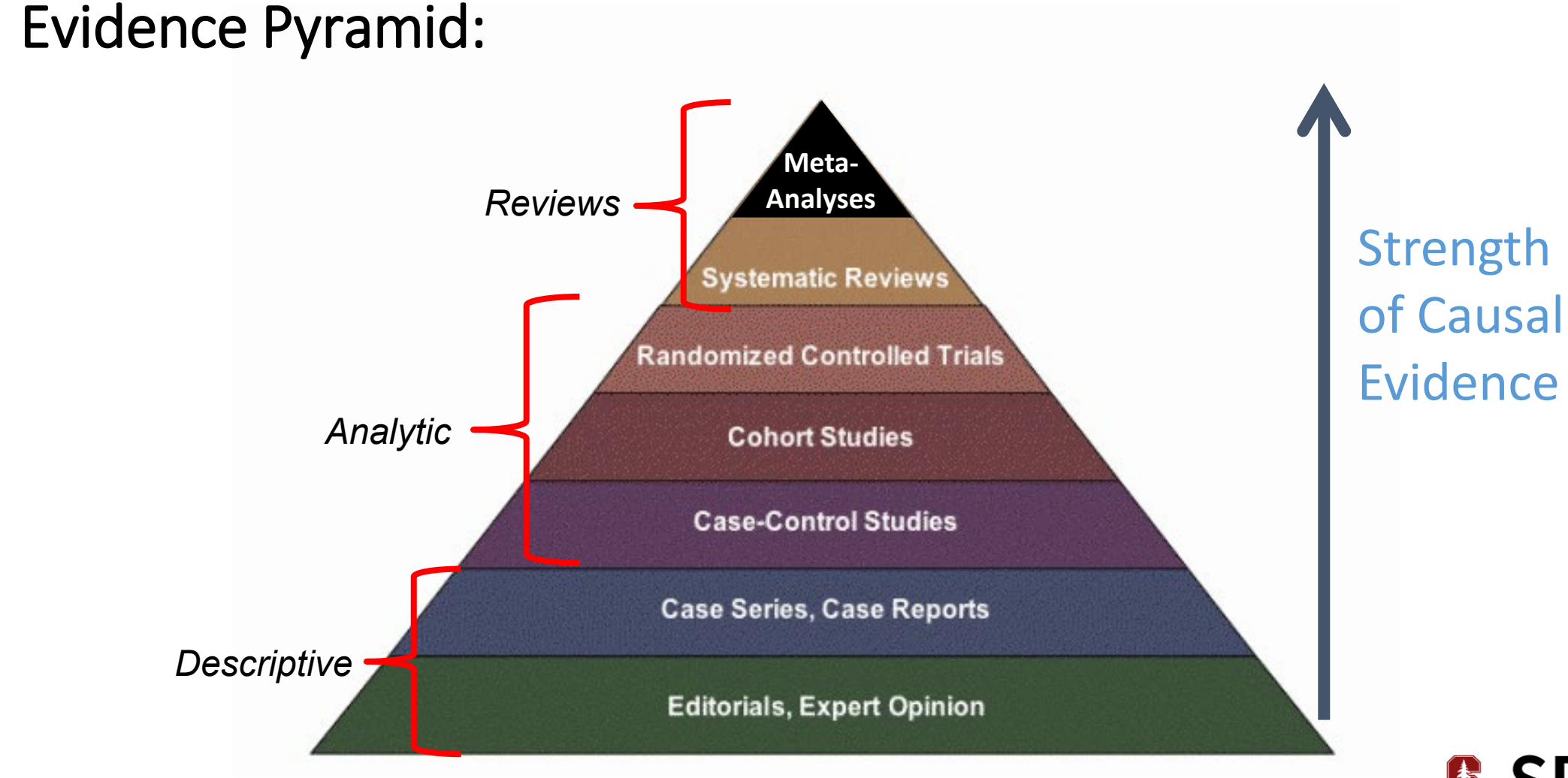
- What effect do you expect to observe?
- What other variables may affect your results?
- How many patients do you anticipate?
- Do you have repeated measures per individual/analysis unit?
- What are your expected consent and follow-up completion rates if collecting data?
- Does your data source help you answer the research question/what are the limitations of your data source?
- Do you have preliminary data?
  - Previous studies / pilot data
  - Published literature

Project plan

- Write up project plan, abstract shell
  - Brief background, project purpose
  - Data source
  - Outcome, comparison groups
  - Time period

Study Designs

Evidence Pyramid:



Study design

Study design	Description	Advantage	Disadvantage
Cross-sectional	One time point measurement of risk factors and outcomes	Inexpensive/quick, often generalizable, valid estimates of prevalence of risk factors/outcomes	Can't determine whether risk factors preceded outcomes; subject to nonresponse or recall bias and confounding
Case-control	Identifies participants based on outcome and asks about past risk factors	Inexpensive/quick, efficient for rare outcomes	Difficult to find comparable control subjects; Sometimes unclear whether risk factors preceded outcomes; can't estimate rates or risks of outcome; recall bias and confounding

Study design

Study design	Description	Advantage	Disadvantage
Prospective cohort	Measures risk factors in outcome-free cohort over time to observe outcome development	Clear temporality	Time consuming and costly; not efficient for rare outcomes; loss to follow-up and confounding
Retrospective cohort	Cohort is studied after outcomes have occurred using stored records	Inexpensive/quick	Data not collected specifically for the study and relevant variables may be unavailable; loss to follow-up and confounding
Randomized controlled trial	Participants randomly assigned to intervention and followed over time	Gold standard for cause-effect; Randomization minimizes confounding	Expensive; not always feasible or ethical

Power and sample size

Type I Error (α): False positive

- Find an effect when it is truly not there
- Solution: Repeat the study

Type II Error β: False negative

- **Do not find an effect when one truly exists**
- Due to: Insufficient power, high variability / measurement error
- Solution: Increase sample size

Statistical Power

A study with low power has a high probability of committing type II error.

- Power = $1 - \beta$
- Sample size planning aims to select a sufficient number of subjects to keep α and β low without making the study too expensive or difficult.
- Translation: How many subjects do I need to find a statistically & meaningful effect?
- Sample size calculation pitfalls:
  - Requires MANY assumptions
  - If power calculation estimated effect size >> observed effect size, sample may be inadequate or observed effect may not be meaningful.

Steps for Data Prep/Analysis

- Write an analysis plan before modeling
- Prepare data for analysis
- Check data
- Assess missingness
- Explore relationships
- Create descriptive table/consort chart
- Build models
- Conduct sensitivity analyses

Data Preparation

- More complicated with retrospective analyses—data capture not designed for your analysis
- Keep the data in one place, name versions
- Merge—does the number of observations line up after merge
  - Person vs. visit level
- Reformat
- Rename/code variables
  - Make the names informative
- Collapse categories with sparse observations
- Remove irrelevant variables/observations (keep track!)
- Document with code
  - Don’t manually fix values or calculate new variables

Data Checking

- Check for errors, inconsistencies
  - Assess range of values: are they all plausible?
  - Logic inconsistencies
- Generate reports of relevant variables
  - Sample sizes, number of missing values, mean, standard deviation, median, interquartile range, minimum, maximum
  - Histogram for continuous variables
- Answer basic data questions
  - Participant age, ethnicity, sex
  - Which variables are normally distributed

Missing data

- Make decision regarding approach before model building
  - Why are the data missing?
  - Compare participants with and without missing data
  - If participants with missing data are different, then this can introduce bias
    - Ex: lab tests
- Options
  - Drop observations with missing data
  - Include missing as a category
  - Imputation

Explore relationships

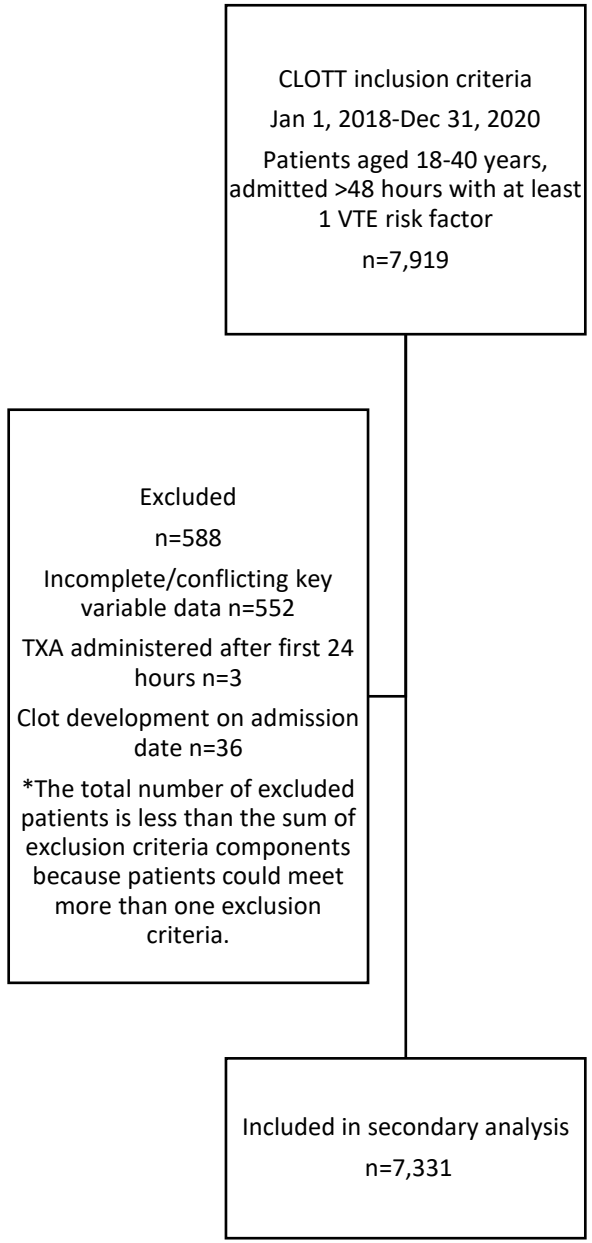
- Correlation matrix for continuous and binary variables
- Scatterplots between continuous variables
- Test model assumptions
  - Proportional hazards for Cox regression
  - Parallel trends for difference-in-differences

Create descriptive table

	No TXA	TXA	p	SMD
	n= 6865	n=466		
Age (mean (SD))	28.6 (6.5)	28.0 (6.5)	0.048	0.094
Male sex	5090 (74.1)	362 (77.7)	0.10	0.083
Race (%)			0.044	0.147
Asian/Pacific Islander/Native American	226 (3.3)	15 (3.2)		
Black	1953 (28.4)	155 (33.3)		
Hispanic/Latino	1165 (17.0)	71 (15.2)		
White	3050 (44.4)	183 (39.3)		
Unknown/Other	471 (6.9)	42 (9.0)		
Head injury	2041 (29.7)	127 (27.3)	0.28	0.055
Spinal injury	285 (4.2)	27 (5.8)	0.11	0.076
Chest injury	2126 (31.0)	219 (47.0)	<0.001	0.333
Pelvic fracture	1233 (18.0)	114 (24.5)	0.001	0.160
Major venous injury	373 (5.4)	66 (14.2)	<0.001	0.297
Long bone fracture	2197 (32.0)	161 (34.5)	0.28	0.054
Major abdominal injury	1425 (20.8)	184 (39.5)	<0.001	0.417
Shock	404 (5.9)	103 (22.1)	<0.001	0.481

S-SPIRE

Consort chart



Common regression models

OUTCOME VARIABLE	APPROPRIATE REGRESSION	MODEL COEFFICIENT
Continuous AND Normal	Linear Regression	Slope (β): How much the outcome increases for every 1-unit increase in the predictor
Binary	Logistic Regression	Odds Ratio (OR): How much the odds for the outcome increases for every 1-unit increase in the predictor
Time-to-Event	Cox Proportional-Hazards Regression	Hazard Ratio (HR): How much the rate of the outcome increases for every 1-unit increase in the predictor

Data Reporting

- Clean up code
  - 3 files:
    - Data cleaning/preparation
    - Tables/figures
    - Additional analyses
  - Comment your code
  - If this is submitted as a paper, you will need to be able to revisit months later!

Resources

- S-SPIRE
  - Work-in-progress sessions
- Stanford statistics department:
  - https://statistics.stanford.edu/resources/consulting-services
- Stanford Department of Biomedical Data Science
  - https://dbds.stanford.edu/data-studio/
- Data analysis examples: UCLA
  - https://stats.oarc.ucla.edu/
- Consort
  - https://www.equator-network.org/reporting-guidelines/consort/
- Observational research reporting: STROBE
  - https://www.strobe-statement.org/

S-SPIRE

Supplementary file 1 STROBE Statement—checklist of items that should be included in reports of observational studies

	Item No.	Recommendation	Page No.	Relevant text from manuscript
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	2	Abstract
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2	Abstract
Introduction				
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4	Introduction: paragraphs 1, 2,3
Objectives	3	State specific objectives, including any prespecified hypotheses	4	Introduction: paragraph 4
Methods				
Study design	4	Present key elements of study design early in the paper	5	Methods: section 1
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	1	Methods: section 1
Participants	6	(*a*) *Cohort study*—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study*—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants	1	Cohort study: Methods: section 1
		(*b*) *Cohort study*—For matched studies, give matching criteria and number of exposed and unexposed *Case-control study*—For matched studies, give matching criteria and the number of controls per case		
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	5,6	Methods: section 2-4
Data sources/	8	For each variable of interest, give sources of data and details of methods of	5,6	Methods: section 2-4

Resources: PHS https://med.stanford.edu/phs/data.html

DATASET	DATASET TYPE	POPULATION	SMALLEST GEO UNIT	SAMPLE SIZE	DATE RANGE	TIME TO ACCESS	STRENGTH
American Family Cohort (AFC)	EMR - Primary Care	US	Census Block	8 million	2010 - 2024	1 month	linkable by individual
MarketScan	Claims - Commercially Insured	US	Metropolitan Area	149 million	2006 - 2022	7 days	prices, variability in insurance type
Medicare 20% sample	Claims - Medicare	US	9 digit zip	11 million	2006 - 2020	6 - 9 months	representative of Americans over 65; rich, longitudinal
Medicaid 100%	Claims - Medicaid	US	5 digit zip	Over 100 million	2011 - 2019	6 - 9 months	representative of Americans enrolled in Medicaid
SEER and CA Cancer Registry - CMS linked data	SEER and CA Cancer Registry will do linkages w/CMS	US	5 digit zip	Varies	Varies	3 - 6 months	Linked dx/treatment data
Aarhus Danish Registers	National cohort, Surveys Administrative data, Biologic samples	Denmark	Census Block	5 million	1968 - 2020	No direct access. 3 - 6 months	Rich, longitudinal, Individual linkages

Common errors

- Not leaving enough time for data cleaning, checking
- Wanting to do something too complex with small dataset
- Asking for last minute help, feedback

Recent resident projects

- **Annals of Surgery:** Chart review analysis using 2011-2023 Stanford Hospital data comparing two surgical techniques in management of pancreatic necrosis, outcomes: 30-day readmission, mortality, etc.
- **Jama Surgery:** Veterans Affairs electronic health record data of patients newly diagnosed with primary hyperparathyroidism 2000-2019, comparing surgically to non-surgically managed patients in development of new diagnosis of depression
- **Annals of Surgery:** commercial claims database patients with traumatic injury 2008-2018, comparing out of pocket costs between groups of increasing injury severity.

P-value Definition

The p-value represents the probability

of finding the observed,
or a more extreme, test statistic

- if the null hypothesis is true.

- Measures evidence against H_0
- Smaller p-value, larger evidence against H_0
- Reject H_0 if p-value $\leq \alpha$

P-Value Pitfalls

- P is highly dependent on sample size
- The *statistical* significance …
  - ➢ does not equal *clinical* significance
  - ➢ does not equal *effect size*
  - ★ Report descriptive statistics with p: n1, n2, %'s, means, SD…
- P is not dichotomous yes/no, but a continuum, <0.001 to >0.99

S-SPIRE