Data Analysis Considerations in Global Tobacco Research

Dongmei Li

Research methodologies in the context of global tobacco research: Challenges and the importance of harmonization Workshop

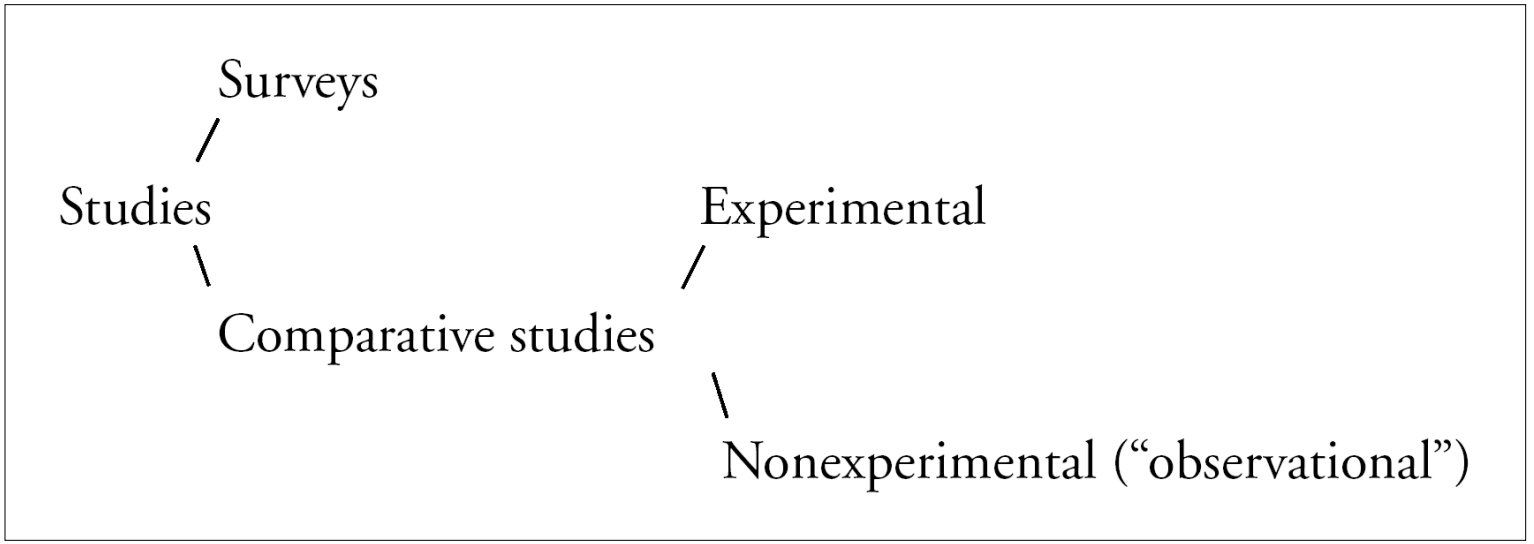
March 11, 2020

Outline

- Type of Studies
  - Surveys
  - Comparative Studies
- Data Collections
  - Sample size consideration
  - Randomization
  - Data harmonization
- Data Analysis
  - Survey data analysis
  - Comparative study data analysis
- Data Interpretations

Type of Studies

- **Surveys**: describe population characteristics (e.g., a study of the prevalence of e-cigarette use in US youth)
- **Comparative studies:** determine relationships between variables (e.g., a study to address whether e-cigarette use causes lung disease)



Surveys

- Goal: to describe population characteristics
- Studies a subset (**sample**) of the population
- Uses sample to make inferences about population
- Sampling :
  - Saves time
  - Saves money
  - Allows resources to be devoted to greater scope and accuracy
- Example of National Surveys
  - Population Assessment of Tobacco and Health (PATH)
  - National Youth Tobacco Survey (NYTS)

Survey Sampling

Simple random samples

Stratified random samples

- Draws independent SRSs from within relatively **homogeneous** groups or ”strata”.

Cluster samples

- Randomly select **large units** (clusters) consisting of smaller subunits.

Multistage sampling

- Large-scale units are selected at random.
- Subunits are sampled in successive stages.

Cautions When Sampling for Surveys

- **Undercoverage**: groups in the source population are left out or underrepresented in the population list used to select the sample.
  - EX: Choose SRS from phone list.

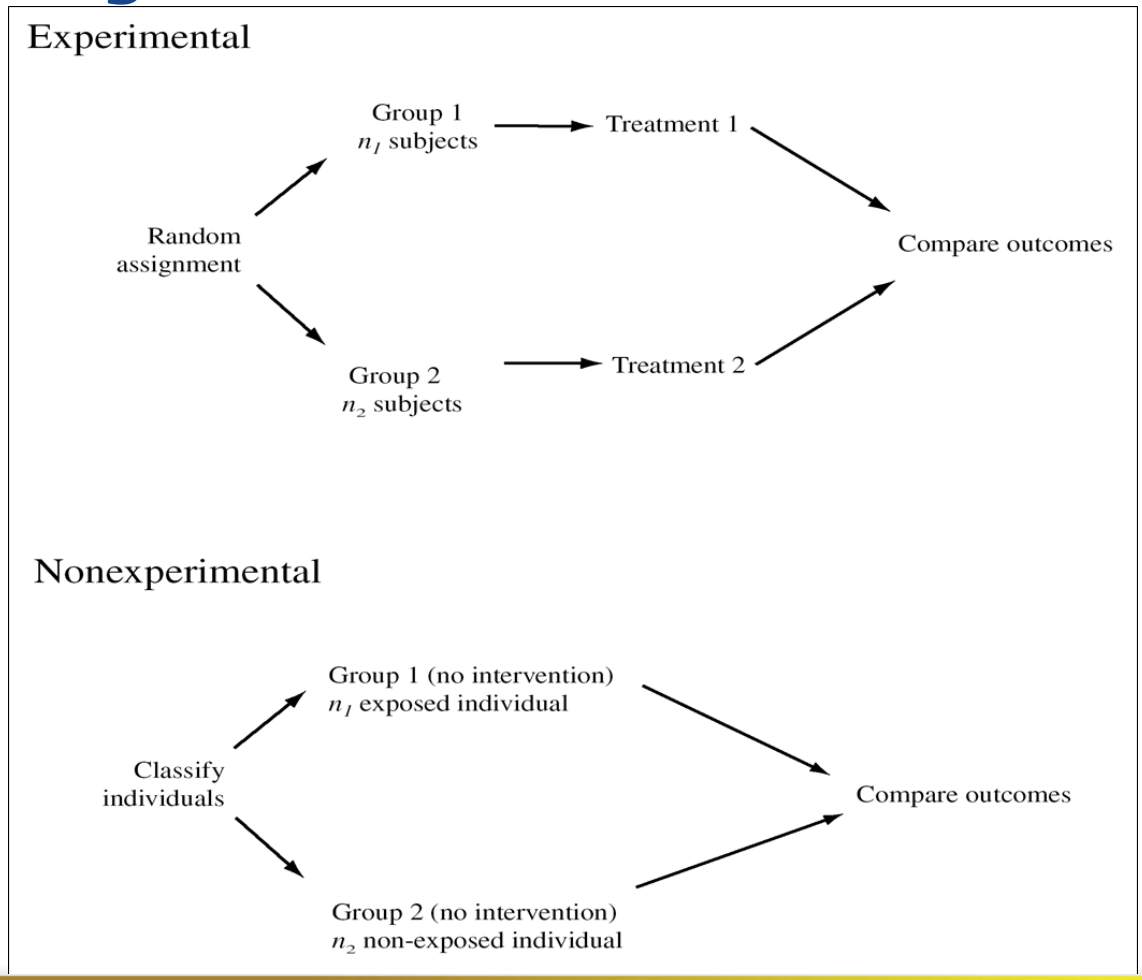
- **Volunteer bias**: occurs when self-selected participants are atypical of the source population.
  - EX: Web survey.

- **Nonresponse bias**: occurs when a large percentage of selected individuals refuse to participate or cannot be contacted.
  - EX: Sensitive topics.

Comparative Studies

- Comparative designs study the relationship between an **explanatory variable** and **response variable**.
- Comparative studies may be experimental or non-experimental.
- In **experimental designs,** the investigator assign the subjects to groups according to the explanatory variable (e.g., exposed and unexposed groups).
- Three important **experimentation principles**:
  - Controlled comparison
  - Randomization
  - Blinding
- In **nonexperimental designs**, the investigator does not assign subjects into groups; individuals are merely classified as "exposed" or "non-exposed."

Study Design Outlines



Data Collection: Sample Size Consideration

- A fun video in YouTube about sample size consideration: https://www.youtube.com/watch?v=Hz1fyhVOjr4

- A tobacco researcher calls and says, "We would like to study the effect of e-cigarettes on risk of stroke. How many patients do I need for the study?"

- A very common statistical question:

- "What is the sample size we need for our study?"

Sample size n = (Total Budget / Cost per patient)? Hopefully not**!**

Data Collection: Sample Size Consideration

- Formulate a PRIMARY research question
  - When you have multiple research questions, choose the most important one as your primary research question
- Identify:
  - A primary hypothesis to test (write down H_0 and H_A), or
  - A quantity to estimate (e.g., using confidence intervals)
- Determine the endpoint or outcome measure associated with the hypothesis test or quantity to be estimated
  - How do we "measure" or "quantify" the responses?
  - Is the measure continuous, binary, or a time-to-event?
  - Is this a one-sample or two-sample problem?
- You can take a conservative approach to use the maximum sample size when you have multiple primary outcomes

Data Collection: Randomization

Randomization refers to the use of chance mechanisms to assign treatments

Randomization balances lurking variables among treatments groups, mitigating their potentially confounding effects

Simple random samples

Data Collection: Data Harmonization

- The PhenX Toolkit (consensus measures for **Phen**otypes and e**X**posures) provides recommended standard data collection protocols for conducting biomedical research.
- The protocols are selected by Working Groups of domain experts using a consensus process, which includes the scientific community.
- The Toolkit provides detailed protocols for collecting data and tools to help investigators incorporate these protocols into their studies.
- Using protocols from the PhenX Toolkit facilitates cross-study analysis, potentially increasing the scientific impact of individual studies.

Data Collection: Data Harmonization

- The PhenX Toolkit (consensus measures for **Phen**otypes and e**X**posures) provides recommended standard data collection protocols for conducting biomedical research.
- The protocols are selected by Working Groups of domain experts using a consensus process, which includes the scientific community.
- The Toolkit provides detailed protocols for collecting data and tools to help investigators incorporate these protocols into their studies.
- Using protocols from the PhenX Toolkit facilitates cross-study analysis, potentially increasing the scientific impact of individual studies.

Data Collection: Data Harmonization

- The Food and Drug Administration's Center for Tobacco Products (CTP) and the National Institutes of Health Tobacco Regulatory Science Program (TRSP) are seeking to expand the depth and breadth of tobacco-related measures that can enhance cross-study analysis in large-scale research.
- TRSP and CTP set a goal to establish consensus measures in tobacco regulatory research for investigators that would enable collaboration, allow for future data comparison, validation, replication, and harmonization.

Data Collection: Data Harmonization

- The Tobacco Regulatory Research Working Groups selected measures for the PhenX TRR Toolkit using criteria suggested by the PhenX Steering Committee.
- The criteria for selecting PhenX measures stipulate that they are: Clearly defined; Well established and have demonstrated utility; Broadly applicable and generally accepted; Low burden to participants and investigators; Reproducible; Specific; Reliable; Available and have existing standard measurement protocols
- Additional criteria for selecting PhenX measures include:Crosscutting relevance across population groups as well as diseases and conditions
- Use in major reference study (e.g., the National Health and Nutrition Examination Survey); Open-source software and nonproprietary instruments preferred; Brevity; Expectation of acceptance by the research community

Data analysis: Survey Data Analysis

- Population Assessment of Tobacco and Health (PATH) Survey
  - Survey Study Design
  - Variance Estimation
  - Variance Estimation for longitudinal samples
- National Youth Tobacco Survey (NYTS)
  - Survey Study Design
  - Variance Estimation
  - Variance Estimation for pooled multiple years samples

Data analysis: PATH Survey

- Survey Study Design
  - PATH is a longitudinal national survey with multiple waves and its target population is the civilian household population 9 years of age or older in the United States (all 50 States and the District of Columbia).
  - PATH Wave 1 use a four-stage, stratified probability sample design
- Variance Estimation
  - Variance estimation requires the use of weights to compensate for variable probabilities of selection, differential nonresponse rates, and possible deficiencies in the sampling frame (e.g., undercoverage of certain population groups).
- Variance Estimation for longitudinal samples
  - Use all waves data and consider the within-subject correlations.

PATH example SAS code for analysis

proc surveyfreq data=*analysis_dataset* varmethod=BRR (fay=**0.3**);

tables *var1 var2 var3 var4* / cl(type=Wilson);

weight R01_A_PWGT;

repweights R01_A_PWGT1 - R01_A_PWGT100;

title "PROC SURVEYFREQ using BRR-Fay Replication";

Run;

The Fay's method, a variant of the balanced repeated replication (BRR) method, was used to form replicate weights in variance estimation in all the PATH survey data analysis.

Data analysis: NYTS Survey

- Survey Study Design
  - NYTS is a cross-sectional national survey data with target population of middle school and high school students in the United States.
  - 2019 NYTS use a stratified, three-stage cluster sample design
- Variance Estimation
  - Variance estimation need to consider primary sampling units (PSU), Strata, and weight.
- Variance Estimation for pooled multiple years samples
  - Divide weights by the number of years being pooled - simple and defensible
  - Example: 2004-2008 pooled analysis (5 years): divide weights by 5

NYTS example: SAS code for analysis

Proc Surveymeans Data=nyts2019 mean;

Var eelcigt ecigt ecigar eslt ehookah celcigt ccigt ccigar cslt chookah;

Class eelcigt ecigt ecigar eslt ehookah celcigt ccigt ccigar cslt chookah;

Stratum v_stratum2;

Cluster psu2;

Weight finwgt;

Domain SCHOOLTYPE SCHOOLTYPE*Sex SCHOOLTYPE*Race_S;

Title "NYTS **2019**, Tobacco Product Use Estimates by School Type, by School Type and Sex Cross-Classified, and by School Type and Race/Ethnicity Cross-Classified"; **run;**

Data Analysis: Comparative Study

- Depending on the type of outcome measurements, we could choose appropriate statistical methods to analyze the data.
- For categorical outcome measurements, we could choose logistic regression models, cumulative logistic regression models, or generalized linear models or generalized estimating equation models with log or logit link functions.
- For continuous outcome measurements, we could choose ANOVA, ANCOVA, general linear model, generalized linear model with identity link function, generalized estimating equation models.
- For longitudinal studies, either generalized estimating equation models or liner mixed effects models could be used to consider the within-subject correlations.

Data Interpretation

Recommendations for data interpretation

- Choose different data visualization methods for different types of measurements
  - Histogram for continuous data
  - Bar chart, pie chart for categorical data
  - Heatmap for correlations
- Use point estimates plus their 95% confidence intervals to measure the effects of intervention or association.
- Interpret the data within the contents
- Report results from analysis with and without outliers if deleting outliers generates different results.

Questions?

Contact Information:

Dongmei Li, PhD

Associate Professor, Clinical and Translational Research

Program Director, Biomedical Data Science Certificate

Director, Biostatistics & Informatics Core, WNY Center for Research on Flavored Tobacco Products (CRoFT)

University of Rochester School of Medicine and Dentistry

Email: Dongmei_Li@urmc.rochester.edu

UNIVERSITY of ROCHESTER
MEDICAL CENTER

MEDICINE of THE HIGHEST ORDER