Statistics for Large Database Research

Graciela Mentz, PhD

Statistician Lead

University of Michigan

MPOG

MULTICENTER PERIOPERATIVE OUTCOMES GROUP

Outline

- Use and types of Statistical Models
- Variable Selection
- Evaluating Model Performance- Measures
- Risk adjustment/Propensity Matching
- Handling Missing Data
- Quasi-experiment design (Difference-in-Difference)

Common Uses for Statistical Models

Prediction

- **Goal**: predict a dependent variable
- Diagnosis, prognosis, or outcome
- Reporting guideline: **TRIPOD**

Association

- **Goal**: understand association of independent variable
- Independent risk factors
- Reporting guideline: **STROBE**

Types of Models / Variable Selection

Generalized Linear Mixed Models (GLMM)

- Fixed effects vs. Random Effects
- Longitudinal vs. Cross-sectional
- Linear vs. Logistic models

Ridge/LASSO/ElasticNet regression

Quantile Regression

Evaluating Model Performance

Internal Validity

- Is the observation **reproducible**?
- Techniques: Cross-validation, bootstrapping

External Validity

- Is the observation **generalizable**?
- Technique: Model discrimination in validation cohort

Overall Model Performance*

Discrimination

- Separate cases with/without a disease or outcome
- Concordance ("c-") statistic (**AUROC**)
- Precision-recall curve (**AUPRC**)

Calibration

- Agreement between observed and predicted risk
- **Calibration Plot**

*Applies to binary outcomes

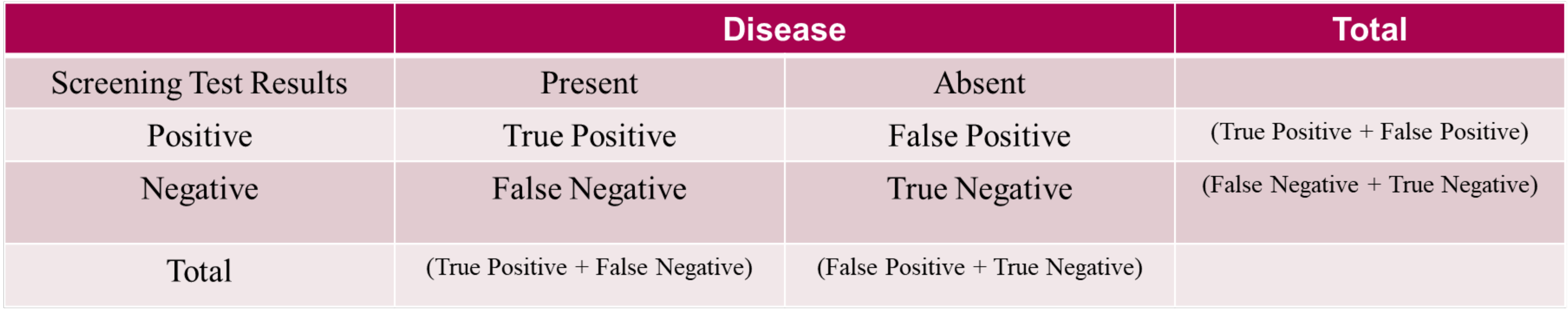
Overall Model Performance*

Net reclassification index (NRI)

- Improvement in prediction between models
- Used to understand **incremental value** of new marker when added to a prediction model

*Applies to binary outcomes

Measures of Model Performance



	Disease		Total
Screening Test Results	Present	Absent	
Positive	True Positive	False Positive	(True Positive + False Positive)
Negative	False Negative	True Negative	(False Negative + True Negative)
Total	(True Positive + False Negative)	(False Positive + True Negative)	

$$\text{Sensitivity}=\frac{True\ Positive}{(True\ positive+False\ Negative)}$$

$$\text{Specificity}=\frac{True\ Negative}{(True\ Negative+False\ Positive)};$$

$$\text{PPV}=\frac{True\ Positive}{(True\ positive+False\ Positive)}\ ;$$

$$\text{NPV}=\frac{True\ Negative}{(True\ Negative+False\ Negative)};$$

Risk Adjustment

Purpose

- To inform decision-making concerning individual welfare.
- Identifying and analyzing potential factors that may negatively impact individual’s health.

Methods

- Statistical Modeling Strategy (Logistic regression)

Propensity Matching

Why?

- Reduce bias due to confounding

How?

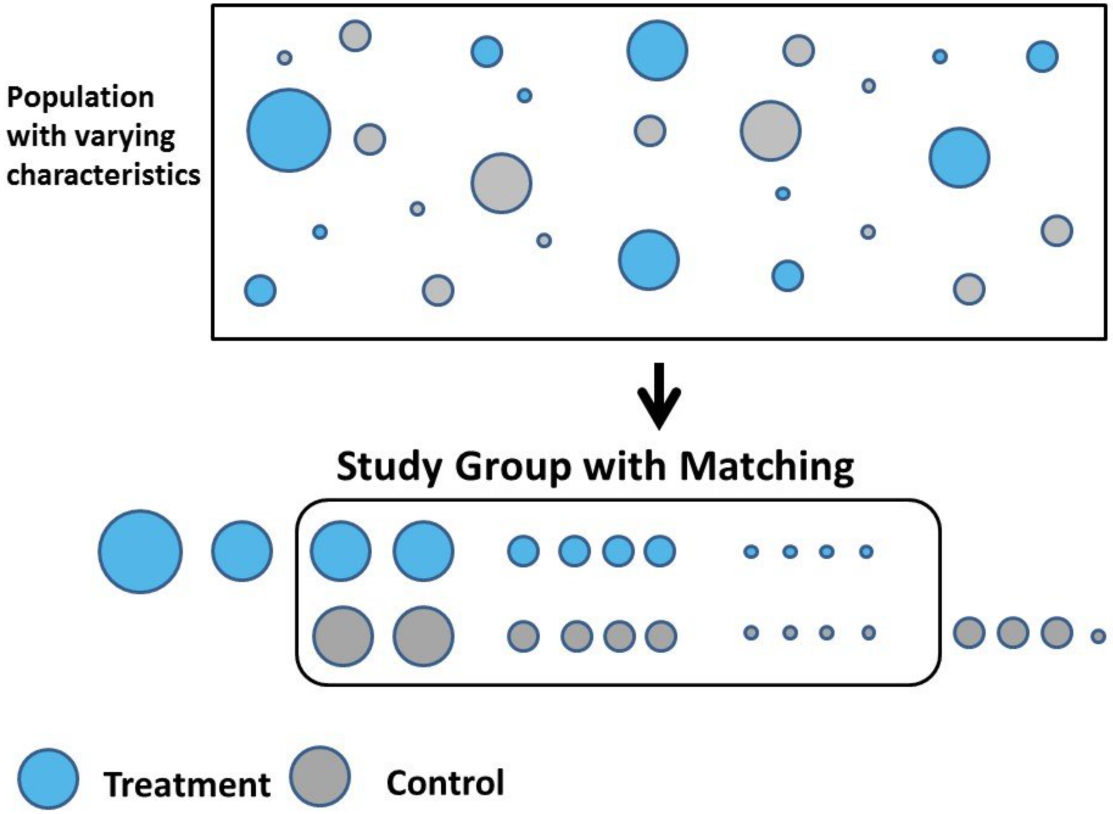
- Using statistical model

Requirements

- Very large dataset

Assumptions

- Outcome is independent of treatment status


Population with varying characteristics
Study Group with Matching
Treatment
Control

Handling Missing data

Types of missing data:

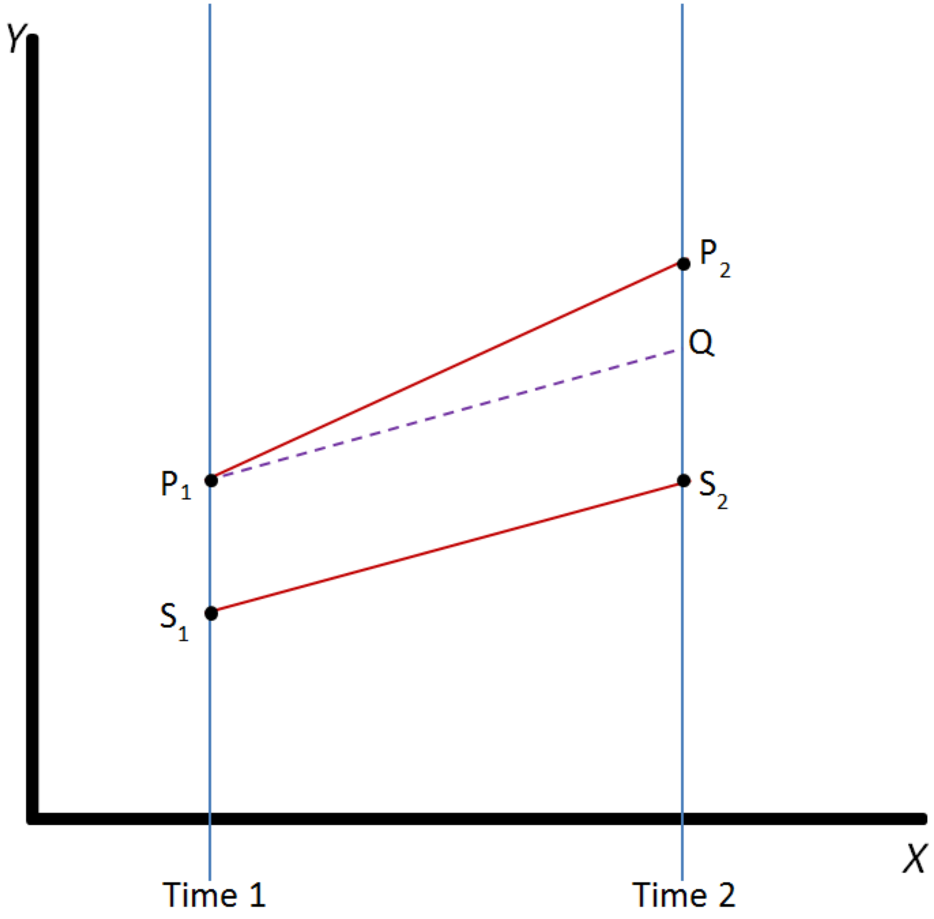
- Missing Not at Random (MNAR)
- Missing at Random (MAR)
- Missing Completely at Random (MCAR)

Methods to handle:

- Complete case analysis
- Multiple imputation

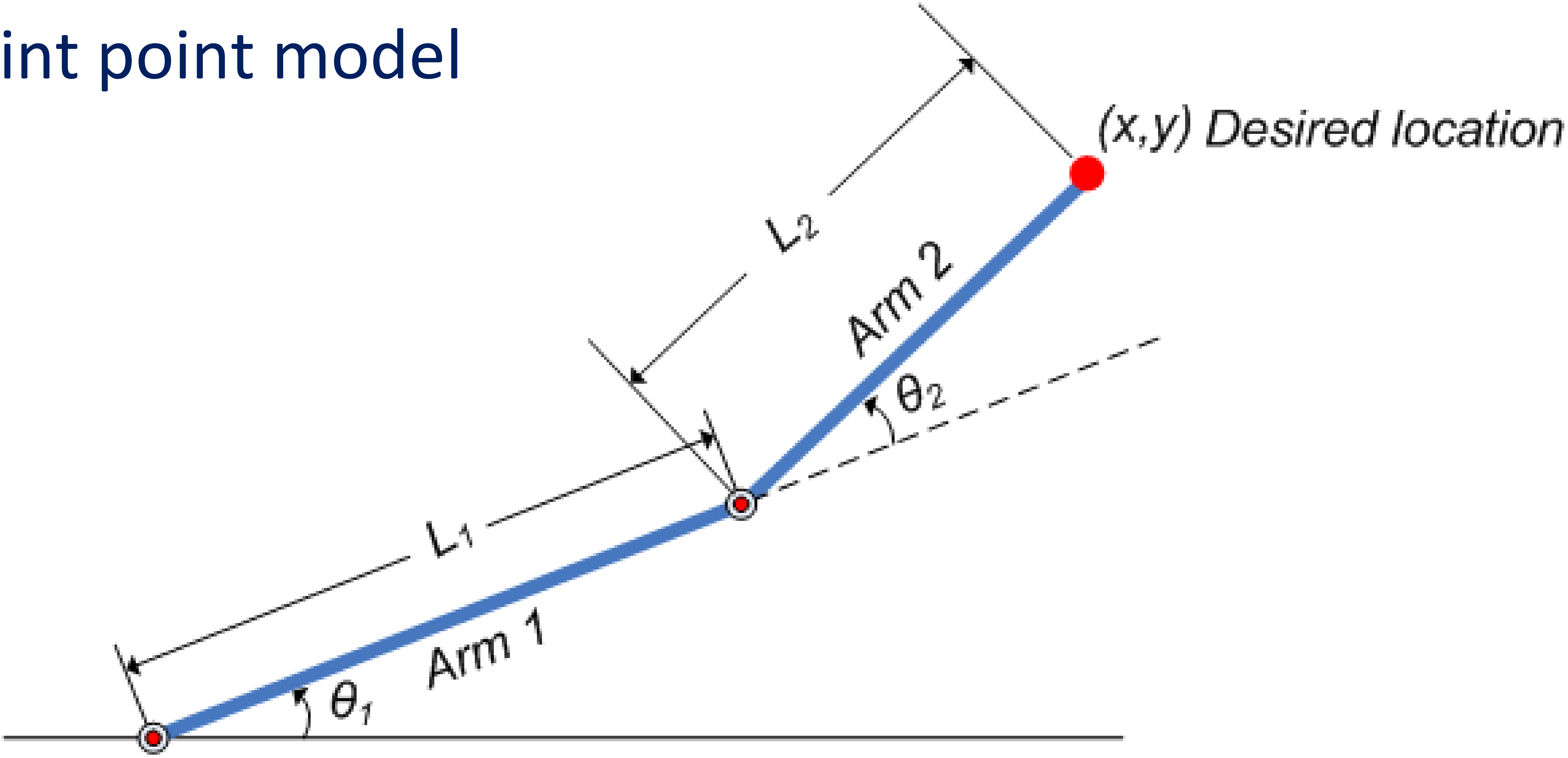
Assessing Trends / Quasi-Experimental Design

- Difference in Differences (DID)



Assessing Trends / Quasi-Experimental Design

- Joint point model



Helpful Statistical Articles / Textbooks

1. Peck R, et al. *Introduction to Statistics and Data Analysis* 6th ed. eTextbook.
2. Heumann C, et al. *Introduction to Statistics and Data Analysis*. Springer.
3. Steyerberg EW, et al. *Assessing the performance of prediction models: a framework for some traditional and novel measures*. Epidemiology. 2010 January ; 21(1): 128–138. oi:10.1097/EDE.0b013e3181c30fb2
4. Holland P. *Statistics and Causal Inference*. Journal of the American Statistical Association December 1986, Vol. 81, No. 396, Theory and Methods.
5. Ibrahim J, et al. *Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data*. Journal of Clinical Oncology, 2010.

Other Recommended Sources of readings

1. Anesthesiology Reader's Toolbox
2. Anesthesia & Analgesia Statistical Minute
3. JAMA Users' Guides to Medical Literature