BIG DATA: OPPORTUNITIES & CHALLENGES

Donna L. Coffman, Ph.D.
Research Associate Professor
College of Health & Human Development

Outline

- Challenge of “secondary” big data analysis
- Association vs. causation vs. prediction and the role of theory
- Opportunities

“Secondary” Analysis: Big Effort

- Extracting big data ≈ collecting data
- Data prep: 1 year or more
- Data often not in useable form
  - Electronic health records
  - Network data

Challenge 1

Communicate need for extensive data prep period to reviewers

Is theory needed?

- *"Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all."* (WIRED magazine)
- Popsicle sales strongly correlate with shark attacks.
- Ice cream sales strongly correlate with forest fires.
- Correlation is not enough

Challenge 2: Low signal-to-noise ratio

- Many spurious relationships
- Many irrelevant variables
- Quantity ≠ quality
- Difficult to separate out the garbage

Social Networks and Causation

You are who your friends are

Birds of a feather flock together

- Peer influence vs. homophily
- Does obesity spread through networks?
- Loneliness? Cigarette smoking?
- Implications for intervention

Potential Value of Prediction?

- Friday night at convenience stores: beer and diaper sales highly associated
  - Should we place advertisements for treatment facilities in the diaper aisle?
  - Should we refer people who buy lots of diapers for treatment?
  - If we are a beer distributor should we start advertising beer on diaper packages?

NIH…Turning Discovery into Health

- Associations should be examined
- Prediction can help us target populations
- But in order to develop effective interventions or policies, we need to establish causation.

Opportunities

- Data mining can provide clues for generating hypotheses that can be followed up with experimental designs (whenever possible).
- If randomization is not possible, it is critical to train on one data set, test on another, and evaluate on third and use methods that allow the strongest possible causal inferences.

Summary

- Time is needed to prep data
- Having more data does not eliminate the need for good data analysis skills. It is still critical to think.
  - *Statistical Thinking: The Bedrock of Data Science* (Joel B. Greenhouse)