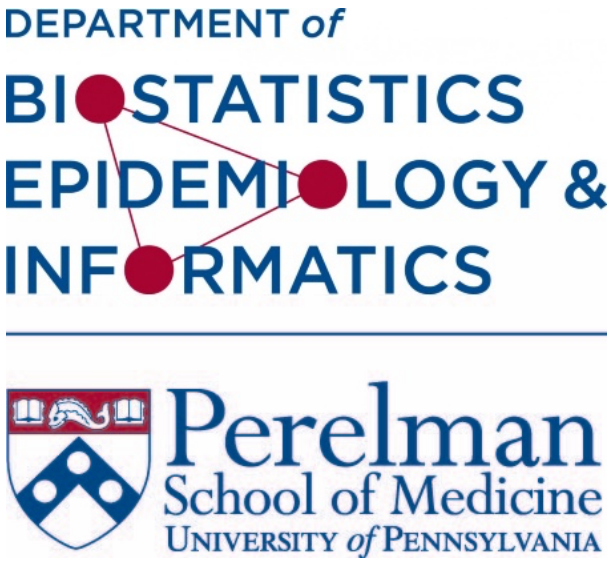
Clinical Trials in the Age of Data Science

Rebecca Hubbard, PhD

rhubb@upenn.edu
https://www.med.upenn.edu/ehr-stats/

April 17, 2019
UPENN Clinical Trials Conference



A fork in the road

Clinical trials

EHR + machine learning

Clinical trials

- Elegant study designs
- High-fidelity interventions
- Well-characterized patient population
- High quality outcome data
- Analytic approaches with well-understood statistical properties
- Statistically rigorous inference

EHR + machine learning

- Large patient populations
- Enormous numbers of potential exposures
- Patients, care, and outcomes as observed in the community
- Sophisticated tools for data processing and prediction

The central dilemma of the data science age

Zombie apocalypse

AI apocalypse

Clinical trials

- Fear of zombies arises from fear of science run amok
- Distrust of experts seen as out of touch
- Artificial conditions
  - inclusion criteria
  - intervention
  - care setting
- Expensive and slow
- Inefficient use of available data resources

EHR + machine learning

- Fear of AI driven by elimination of the human element from science
- Black box methods
- Recapitulation of our biases and prejudices
- Error-prone data may lead us to wrong answers

The middle path

Data Science = Data + Science

Bridging science and data

The challenge: Healthcare system-embedded trials

- Addresses many concerns about generalizability of clinical trials
- Accelerate translation into clinical practice
- Preliminary data from EHR facilitates more realistic design choices
- Expert knowledge of the healthcare system (enrollment/disenrollment, data collection and recording, coding) is key

Bridging science and data

The challenge: Phenotyping error

- Few data elements derived from EHR are “research quality”
- Use of anchor variables leverages clinical knowledge of disease and coding processes
- In combination with statistical methods for outcome misclassification yields unbiased and efficient estimates

Bridging science and data

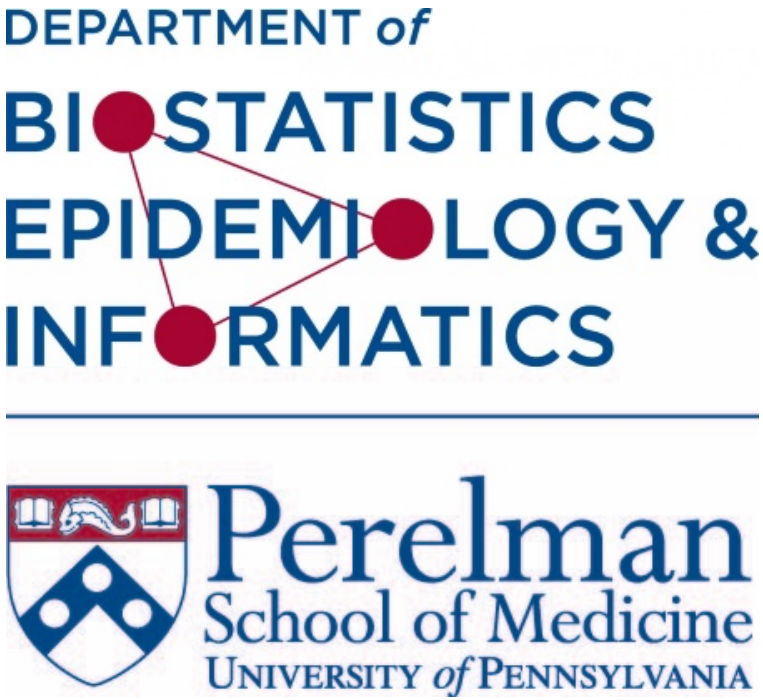
The challenge: Heterogeneity

- Heterogeneity is ubquitous in EHR-based research and can be seen as a feature or a bug
- Lack of fidelity in interventions can be seen as real-world performance
- Creates opportunities for precision medicine through availability of unique sub-groups
- Statistical approaches to detecting heterogeneity can highlight differences in coding or clinical practice

Concluding thoughts

- Statistical science has the tools to address many of the challenges presented by EHR data
- Data science is necessarily collaborative: domain expertise, informatics, computer science, statistics
- But... how do we build collaboration with individuals or disciplines that are committed to pursuing one path to the exclusion of the other?
  - *Why are you using EHR data when it has so many errors?*
  - *Why are you worried about error when we have so much data?*

- **Putting the science in data science can harness the power of data and mitigate the errors in EHR**


DEPARTMENT of
BIOSTATISTICS
EPIDEMIOLOGY &
INFORMATICS
Perelman
School of Medicine
University of Pennsylvania