ELECTRONIC HEALTH RECORDS (EHR) DATA:

METHODS AND APPLICATIONS

John D. Rice, PhD, and
Carter Sevick, MS

ACCORDS
ADULT AND CHILD CONSORTIUM FOR HEALTH OUTCOMES RESEARCH AND DELIVERY SCIENCE
UNIVERSITY OF COLORADO | CHILDREN'S HOSPITAL COLORADO

What is EHR data?

- EHR is an electronic, digital version of a patient's chart
- Can contain
  - Medical history
  - Diagnoses
  - Medications
  - Treatment plans
  - Immunization dates
  - Radiology images
  - Lab test results
- Can allow for use of data-driven tools to aid providers' decisions about a patient's care
- Can automate/streamline providers' workflows

(Source: https://www.healthit.gov/faq/what-electronic-health-record-ehr)

Uses of EHR

- Pharmacovigilance
- Phenotyping
- Natural Language Processing
- Data Application and Integration
- Clinical Decision Support
- Personal Monitoring and Social Media

Source: Ross MK, Wei W, Ohno-Machado L. "Big data" and the electronic health record. *Yearb Med Inform*. 2014;9(1):97-104. Published 2014 Aug 15. doi:10.15265/IY-2014-0003

Data availability

- Secondary use of data for research
  - Researchers must consider the circumstances in and purposes for which data was originally collected
  - Regulatory and administrative concerns can be mitigated by use of de-identified or “limited” data sets
- What data is available may be driven by convenience and/or policy
  - For example, squirmy babies may not have a height recorded, but weights are less of a problem
  - Low-cost procedures may go un-coded if it is perceived that the effort of coding would outweigh the monetary return
  - Some information may only be available as free text

Meaning of diagnoses and procedures

- Diagnoses
  - Possibly an original diagnosis
  - May be about a previously diagnosed condition
  - Justification for procedures
- Procedures
  - Services, in that visit, that were billed
- Guidelines on how to elicit certain information from medical records
  - Standardized methods may exist for diagnoses based on EHR
  - For example: birth defects, AHRQ, HEDIS (asthma), Seattle Children's Hospital
- Biases could exist due to certain billing practices
  - Coding for more expensive procedures
  - *Example*: billing for vaccines was skipped because procedure was not lucrative enough to bother

Necessary information about EHR for research

- Source system/s for each data element
- How close the source system is to the original point of charting
- Source of the data values
- Uniformity of the clinical work flow in which a data element was collected
- How the capture of the data has changed over time
- How relevant ICD and CPT codes are applied within the facility
- Consistency of charting the data element across the facility
- Whether the data element contains both data from devices and manual measurement and if so how these are differentiated
- Any cleaning, standardization or other transformations performed on the data element

(Source: https://rethinkingclinicaltrials.org/resources/acquiring-and-using-electronic-health-record-data/)

Statistical methods used with EHR data

- Observational data: need to account for non-randomized treatment in statistical analyses
  - Regression adjustment (limitations with too many covariates)
  - Propensity score methods (matching, weighting)
- Association versus causation
- Prediction versus inference

Prediction or inference?

- It is important to define the goal of your study early
- **Inference** is used to determine whether there is statistical evidence of (e.g.) the superiority of treatment A versus treatment B in a given data set
- **Prediction** is distinct from inference and is directed toward classifying/ future patients on the basis of models trained on a given data set

Causal inference

- Conventional method is **covariate adjustment**
  - Used to adjust for differences between treatment groups in observational databases
  - Problems can occur with overfitting (many covariates relative to # of patients in database)
- **Propensity score methods** offer an alternative approach
  - A propensity score (PS) is defined as the probability of a patient receiving an intervention, given a set of covariates
  - Often this is estimated using a logistic regression model with outcome being treatment

Further reading: Elze et al. (2017), Comparison of Propensity Score Methods and Covariate Adjustment: Evaluation in 4 Cardiovascular Studies, *Journal of the American College of Cardiology*, Volume 69, Issue 3, pages 345-357.

Propensity score methods

- Method is intended to achieve “covariate balance”: individuals with the same PS should have the same distribution of that covariate regardless of treatment (on average)
- After constructing/estimating PS for all participants, several methods may be used for analysis
  - **Stratification**: separate treatment effect estimated within strata defined by similar PS, then combined
  - **Matching**: attempts to find 1 (or more) individual in each of the treatment groups with similar PS, then conduct conventional regression analysis (possibly with adjustment for pairing)
  - **Weighting**: analysis is based on giving greater weight to individuals who received “surprising” treatments
  - **PS as covariate**: can include the estimated PS in regression models with treatment

Methods for prediction

- Often trying to predict binary outcomes (e.g., treatment success)
- Can use regression for prediction (overlaps with inference)
- More specific methods for prediction exist
  - Penalized regression (ridge, LASSO)
  - Multivariate adaptive regression splines (MARS)
  - Decision trees/random forests
  - Support vector machines
  - Neural networks
- Predictive modeling considerations
  - Multiplicity of good models
  - Simplicity versus accuracy

Further reading: Leo Breiman (2001), Statistical Modeling: The Two Cultures, *Statistical Science,* Vol. 16, No. 3, pages 199–231.

Issues with very large data

- Clustering (by site, hospital, etc.)
  - Occurs frequently with large EHR data sets
  - Standard statistical methods encounter computational difficulties
  - Other methods can underestimate variance with a small number of clusters
- P-values and inference
  - With very large sample sizes, many p-values will be small
  - Difference between **clinical significance** and **statistical significance**

Issues with very small data

- EHR is not only associated with big data
- EHR can also answer questions about rare diseases
  - May affect <1 per 10,000 population
  - Often interested in case detection from medical records
  - **Need a large pool of data to get even a handful of cases**
- Statistical methods: exact inference for small samples
  - Bootstrap
  - Fisher's exact test
  - *t* test
- Samples can be **deceptively** small: e.g., if cluster-randomized with small number of clusters
  - Total sample size may be large
  - Number of independent units (=number of clusters) will usually be much smaller

Final thoughts

- EHR can represent a wealth of information for researchers
- Need to be careful about interpretation when analyzing
- Always involve a statistician early on in when designing any study, including those using EHR!