# A Tour of Pragmatic Methods & Measures: Planning Your Data Collection Strategies

## Electronic Health Records (EHR) Data

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

**What is EHR data?**

EHR is an electronic, digital version of a patient's chart. This can contain

- Medical history
- Diagnoses
- Medications
- Treatment plans
- Immunization dates
- Radiology images
- Lab test results

**What can EHR data be used for?**

(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)

- **Pharmacovigilance** is the process by which drugs already brought to market are monitored for adverse events.
- **Phenotyping** is associating certain disease characteristics with genetic patterns.
- **Natural language processing** (NLP) is a set of biomedical informatics techniques used to extract and structure information from text.
- **Data Application and Integration** refers to the use of EHR to compare treatments in conditions closer to reality than clinical trials would be.
- **Clinical Decision Support** refers to the use of EHR to aid in improving clinical documentation, increasing adherence to clinical guidelines, helping predict outcomes, making diagnoses, and preventing errors.
- **Personal Monitoring and Social Media** refers to the application of mobile technology to healthcare.

**What metadata does a researcher need to be aware of when using EHR data?**

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

- 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

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

CoPRH Colorado Pragmatic Research in Health Conference 2020

# A Tour of Pragmatic Methods & Measures: Planning Your Data Collection Strategies

## Electronic Health Records (EHR) Data

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

**What should researchers keep in mind when analyzing EHR data?**

- Since EHR data is observational by nature, researchers need to account for non-randomized “treatment assignment” in statistical analyses. It is important not to confuse association and causation:
- **Association** can be measured in many ways, depending on the form of the variables and the type of relationship that is of interest (e.g., correlation coefficients for continuous data or odds ratios for binary data).
- **Causation** means that changes in one variable directly cause changes in another, which often can only be reliably established by random assignment to the conditions under study.
- Both **regression adjustment** and **propensity score analyses** provide methods to remove confounding bias when estimating the relationship between a given exposure and outcome from EHR data or other observational data sources.

**How can propensity scores (PS) be used to conduct causal inference?**

- **Stratification** involves defining 5-10 groups on the basis of similar propensity scores. The overall treatment effect is then calculated by taking the (weighted) average across strata.
- **Matching** study subjects on propensity scores, often in a 1:1 ratio, is followed by estimation of treatment effect and its significance using a statistical method appropriate for clustered data (e.g., paired *t* test).
- **Weighting** methods assign weights (usual weights correspond to 1/PS for the treated cohort and 1/(1-PS) for the control cohort), then conduct a weighted version of the usual analysis (e.g., linear regression model).
- Propensity scores can be **included as a covariate** in any regression models. This “summarizes” the effects of all other covariates on the treatment group membership probability with just one degree of freedom in the regression model.

**What methods are available for training predictive models?**

- **Regression models** may be used for prediction (logistic, linear, Cox, etc.). **Penalized regression** is a refinement of these general methods that induces shrinkage in regression coefficients to increase out-of-sample model performance.
- **Multivariate adaptive regression splines (MARS)** represent a regression-like method that allows for the accommodation of nonlinear relationships and interactions between variables, while retaining some of the interpretability of more traditional regression models.
- **Decision trees** are a method of classifying an outcome variable using a flowchart-like structure. **Random forests** aggregate many such trees based on random perturbations of the original data. (More on this momentarily.)
- **Support vector machines** and **neural networks** are techniques from machine learning that may lack the interpretability of methods such as regression but can in some cases attain greater accuracy.

**Notes:**

ACCORDS

ADULT AND CHILD CONSORTIUM FOR HEALTH OUTCOMES
RESEARCH AND DELIVERY SCIENCE

UNIVERSITY OF COLORADO | CHILDREN’S HOSPITAL COLORADO

CoPRH Colorado Pragmatic Research in Health Conference 2020