Clinical Data Wrangling: Introduction

HIP523

Ted Laderas, PhD

Acknowledgements

- Some of these slides are adapted from Nicole Weiskopf from our clinical data wrangling workshop
- This dataset is adapted from the synthetic patient cohort used in BMI 569 Data Analytics

Please Note

- This session is subject to the BioData Club Code of Conduct: https://biodata-club.github.io/code_of_conduct/
- This session is meant to be a psychologically safe space to ask questions
- Please respect each other and interact with each other respectfully, or I will mute you or ask you to leave.

Introduction

- Assistant Professor, DMICE
- Bioinformatics and Interactive visualization
- Certified RStudio Instructor
- Founder, BioData Club and Cascadia R conference
- BMI 569 – Data Analytics
- BMI 507 – Ready for R

Short Introduction

- In Chat type:
  - Your Name
  - Your Department
  - Why are you interested in Data Analysis?
  - What do you hope to get out of this session?

Learning Objectives

- **Understand** basic issues with using clinical data and how it is collected
- **Learn** and **Apply** Basic Principles of Exploratory Data Analysis to assess whether data is fit for reuse
- **Identify** when missing values in data may affect using clinical data for reuse
- **Identify** possible predictors of an outcome using exploratory data analysis

Our Big Goal

- Reduce unnecessary hospital readmissions from poor inpatient or outpatient care
  - Reduce overall costs
  - Improve overall outcomes of our patient population
- Our metric: whether a patient has be readmitted to the hospital within 30 days

Our Analytic Goal

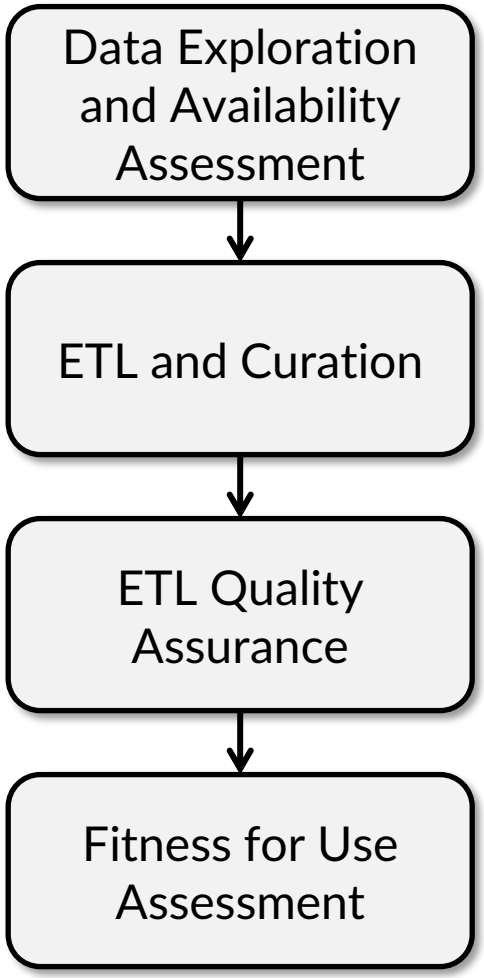
- Can we predict which patients in our cohort had 30 day hospital readmissions?
- Is our data fit for predicting this?
- Can we understand which of these predictors is helpful?
  - Comorbidities (diabetes complications and/or myocardial complications)
  - Length of Stay in Hospital
  - Age

Your Turn

- In chat, construct a hypothesis about one of these:
  - length of stay
  - History of diabetes complications
  - History of myocardial infarctions
  - Age

  And how they would impact whether a patient is readmitted to the hospital within 30 days

Using a systematic but flexible approach to "wrangling" your clinical data, combined with basic competencies in exploratory data analysis, will get you where you want to go.



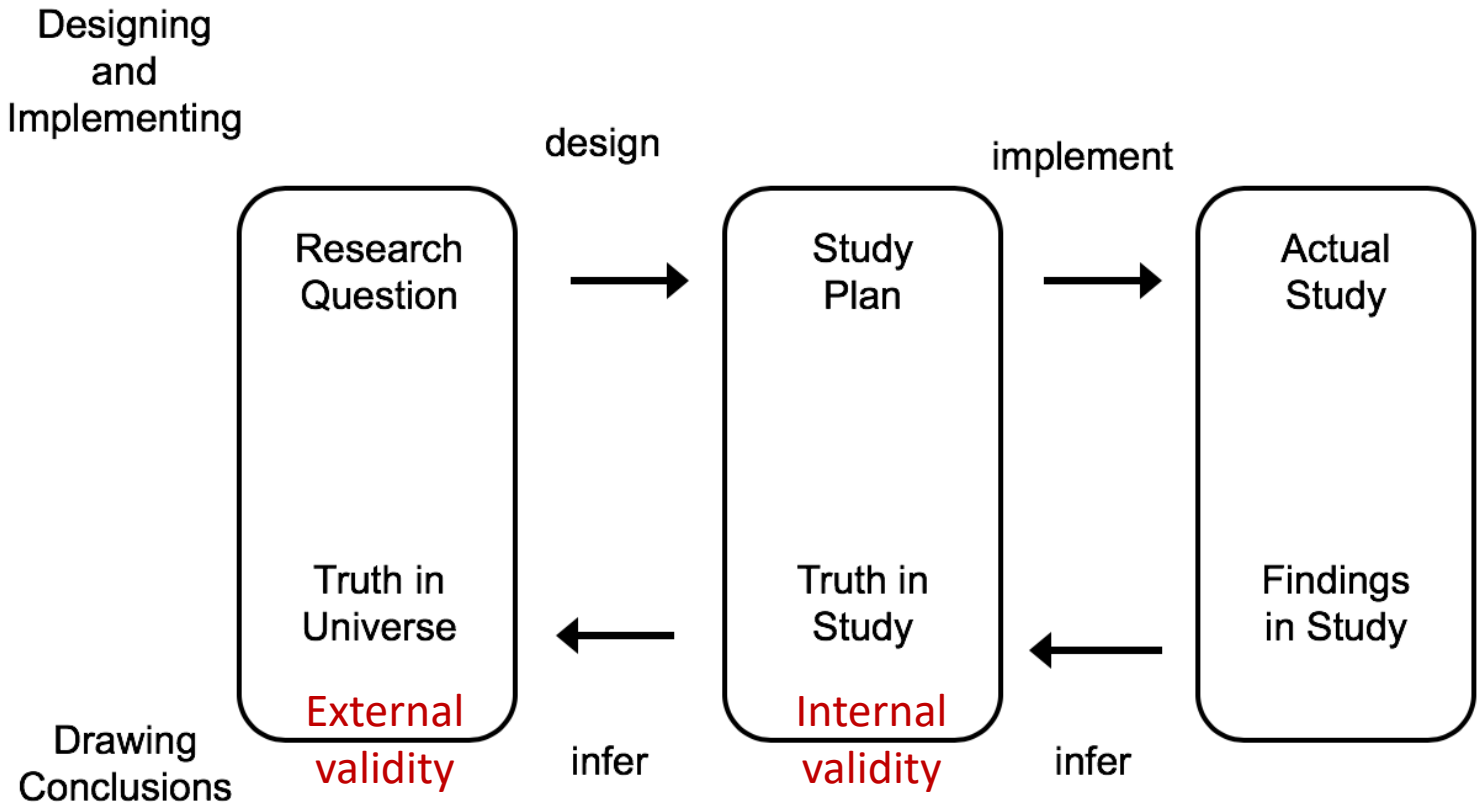
Defining “Clinical” Data

- Clinical documentation
  - Unstructured: Progress notes, Medical Hx, etc
  - Structured: Labs, Medications, Orders, Dx, etc
- Administrative data (often included in this definition)
  - Billing
  - CMS, Insurance
- Primary purpose is *not* research

Benefits of Clinical Data in Research

1. Decrease costs (time and money)
2. Enable recruitment and retention
   - Underrepresented populations
   - Rare diseases
3. Volume, variety, and velocity of data
   https://www.gartner.com/it-glossary/big-data
4. Increased representativeness (aka, generalizability or external validity)

Good research should give us broadly applicable truths



Thanks to Adam Wilcox

Credit to Designing Clinical Research: An Epidemiologic Approach, Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. Third edition, 2006

What *is* the quality of EHR data?

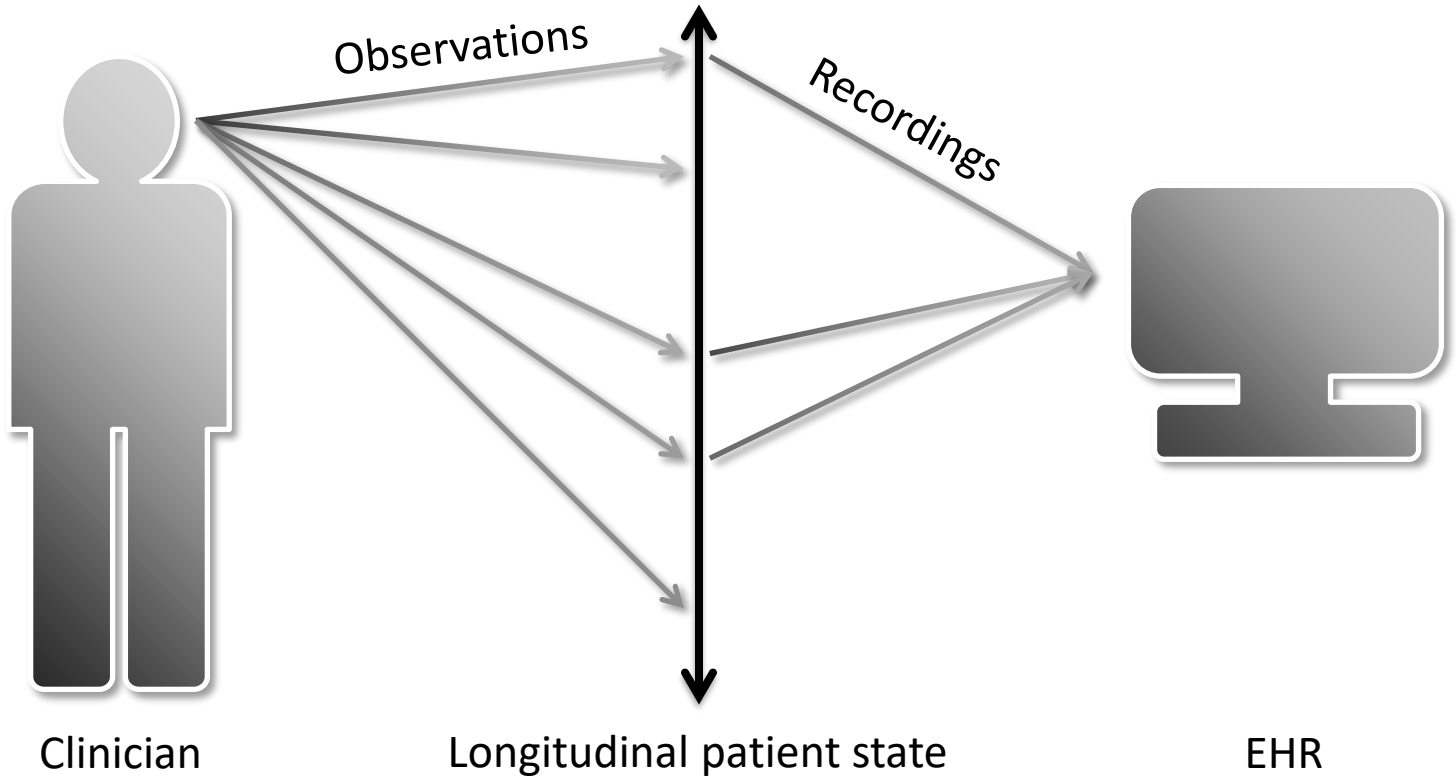
- Hogan and Wagner (1997)
  - Correctness: 44% - 100%
  - Completeness: 1.1% - 100%
- Chan et al. (2010)
  - Completeness of BP: 0.1% – 51%

Hogan & Wagner (1997) Accuracy of data in computer-based patient records.
Chan et al. (2010) EHRs and the reliability and validity of quality measures: a review of the literature.

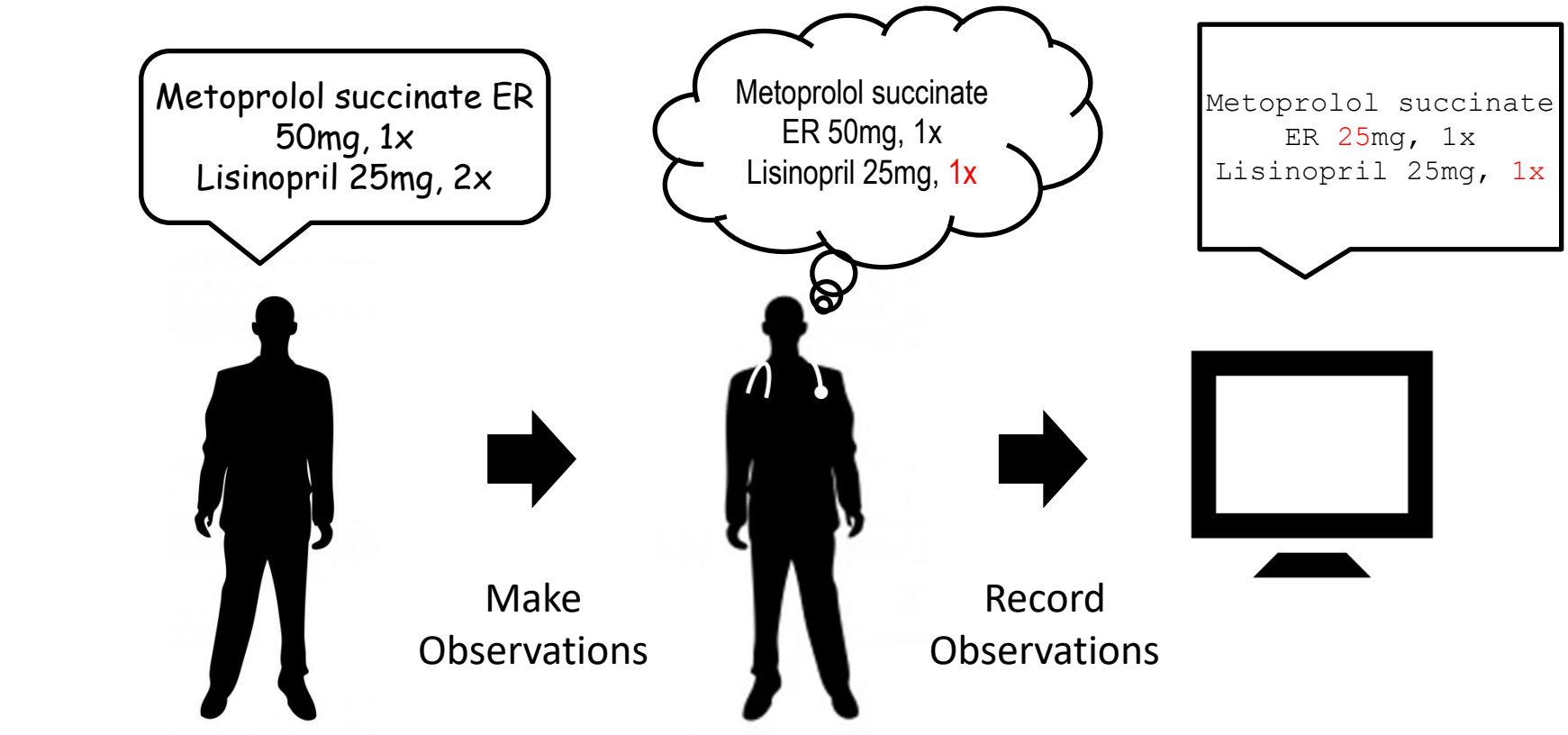
Why are EHR data of such variable and often poor quality?

- The quality of the data is defined with respect to the intended use of the data (fitness for use)
- Clinical data are collected for patient care and billing purposes, not for research
- The processes involved in taking a clinical truth about a patient all the way to a dataset being used for research is fraught with pitfalls

Not all clinical concepts are observed, and not all observations are recorded.



Weiskopf et al. (2013) Defining and measuring completeness of EHRs for secondary use


Metoprolol succinate ER 50mg, 1x
Lisinopril 25mg, 2x
Metoprolol succinate ER 50mg, 1x
Lisinopril 25mg, 1x
Metoprolol succinate ER 25mg, 1x
Lisinopril 25mg, 1x
Make Observations
Record Observations
Multi-vitamin, 1x
Metoprolol succinate ER 50mg, 1x
Lisinopril 25mg, 2x

A quick intro to missingness

There are three types of missingness, defined by Rubin.

- **MCAR** (missing completely at random): pattern of missingness is not related to any other data
- **MAR** (missing at random): the pattern of missingness is related to data that are *present*
- **MNAR** (missing not at random): the pattern of missingness is related to the values of the data that are *missing*

Rubin (1976) Inference and missing data

An overly simple example of missingness

	Population	Sample	MCAR	MAR	MNAR
Men	70.0	70.2	70.3	70.5	71.3
Women	64.0	64.2	64.1	64.2	65.4
Overall	67.0	67.2	67.1	66.3	68.4

An overly simple example of missingness

	Population	Sample	MCAR	MAR	MNAR
Men	70.0	70.2	70.3	70.5	71.3
Women	64.0	64.2	64.1	64.2	65.4
Overall	67.0	67.2	67.1	66.3	68.4

Sample of 200 men, 200 women

Rubin (1976) Inference and missing data

An overly simple example of missingness

- **MCAR** (missing <u>completely</u> at random): pattern of missingness is not related to any other data

	Population	Sample	MCAR	MAR	MNAR
Men	70.0	70.2	70.3	70.5	71.3
Women	64.0	64.2	64.1	64.2	65.4
Overall	67.0	67.2	67.1	66.3	68.4

25% of men and women didn't want to share their height

Rubin (1976) Inference and missing data

An overly simple example of missingness

- **MAR** (missing at random): the pattern of missingness is related to data that are *present*

	Population	Sample	MCAR	MAR	MNAR
Men	70.0	70.2	70.3	70.5	71.3
Women	64.0	64.2	64.1	64.2	65.4
Overall	67.0	67.2	67.1	66.3	68.4

50% of men didn't want to share their height

Rubin (1976) Inference and missing data

An overly simple example of missingness

- **MNAR** (missing <u>not</u> at random): the pattern of missingness is related to the values of the data that are *missing*

	Population	Sample	MCAR	MAR	MNAR
Men	70.0	70.2	70.3	70.5	71.3
Women	64.0	64.2	64.1	64.2	65.4
Overall	67.0	67.2	67.1	66.3	68.4

Half of the shortest 25% of men and women didn't share their height

Rubin (1976) Inference and missing data

Data quality is a large problem area that is still mostly unsolved. Ultimately we need to improve the source data, but until then:

- Understand the provenance of your data, especially in terms of system complexities and potential failure points
- Don’t think of data quality as an issue of right versus wrong values– the problem is generally more subjective (fitness for use)
- Data that are “bad” at random aren’t always an issue in research, but systematic data quality problems can drastically alter your results
- When you uncover potential data quality problems, be thoughtful in your attempts to compensate

Using a systematic but flexible approach to "wrangling" your clinical data, combined with basic competencies in exploratory data analysis, will get you part of the way there.