Clinical Data Wrangling

An Introduction

Aaron S Coyner, PhD

HIP 523

Acknowledgements

- Slides
  - Adapted from the Clinical Data Wrangling Workshop
  - Nicole Weiskopf, PhD
  - Ted Laderas, PhD
- Data
  - Adapted from the synthetic patient cohort used in BMI 569: Data Analytics

Introduction

Aaron S. Coyner, PhD

- Senior Computational Biologist
  - Casey Eye Institute
  - Data Scientist
  - Machine Learning Engineer
- Computer Vision + Clinical Data
- Bioinformatics + Clinical Informatics

Learning Objectives

Learning Objectives

- **Understand** basic issues with using clinical data and how it is collected

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

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

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

Overall Goal

Overall Goal

- **Reduce** unnecessary hospital readmissions from poor in/outpatient care

Overall Goal

- **Reduce** unnecessary hospital readmissions from poor in/outpatient care
    - **Improve** overall outcomes of our patient population

Overall Goal

- **Reduce** unnecessary hospital readmissions from poor in/outpatient care
  - **Improve** overall outcomes of our patient population
  - **Reduce** overall costs

Overall Goal

- **Reduce** unnecessary hospital readmissions from poor in/outpatient care
  - **Improve** overall outcomes of our patient population
  - **Reduce** overall costs
- **Metric:** whether a patient has be readmitted to the hospital within 30 days

Analytic Goal

Analytic Goal

- Can we **predict** which patients in our cohort had 30 day hospital readmissions?

Analytic Goal

- Can we **predict** which patients in our cohort had 30 day hospital readmissions?
  - Is our data **fit** for predicting this?

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 predictors are helpful?

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 predictors are helpful?
    - Comorbidities (e.g., diabetes complications, myocardial complications, etc.)

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 predictors are helpful?
    - Comorbidities (e.g., diabetes complications, myocardial complications, etc.)
    - Length of Stay in Hospital

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 predictors are helpful?
    - Comorbidities (e.g., diabetes complications, myocardial complications, etc.)
    - Length of Stay in Hospital
    - Age

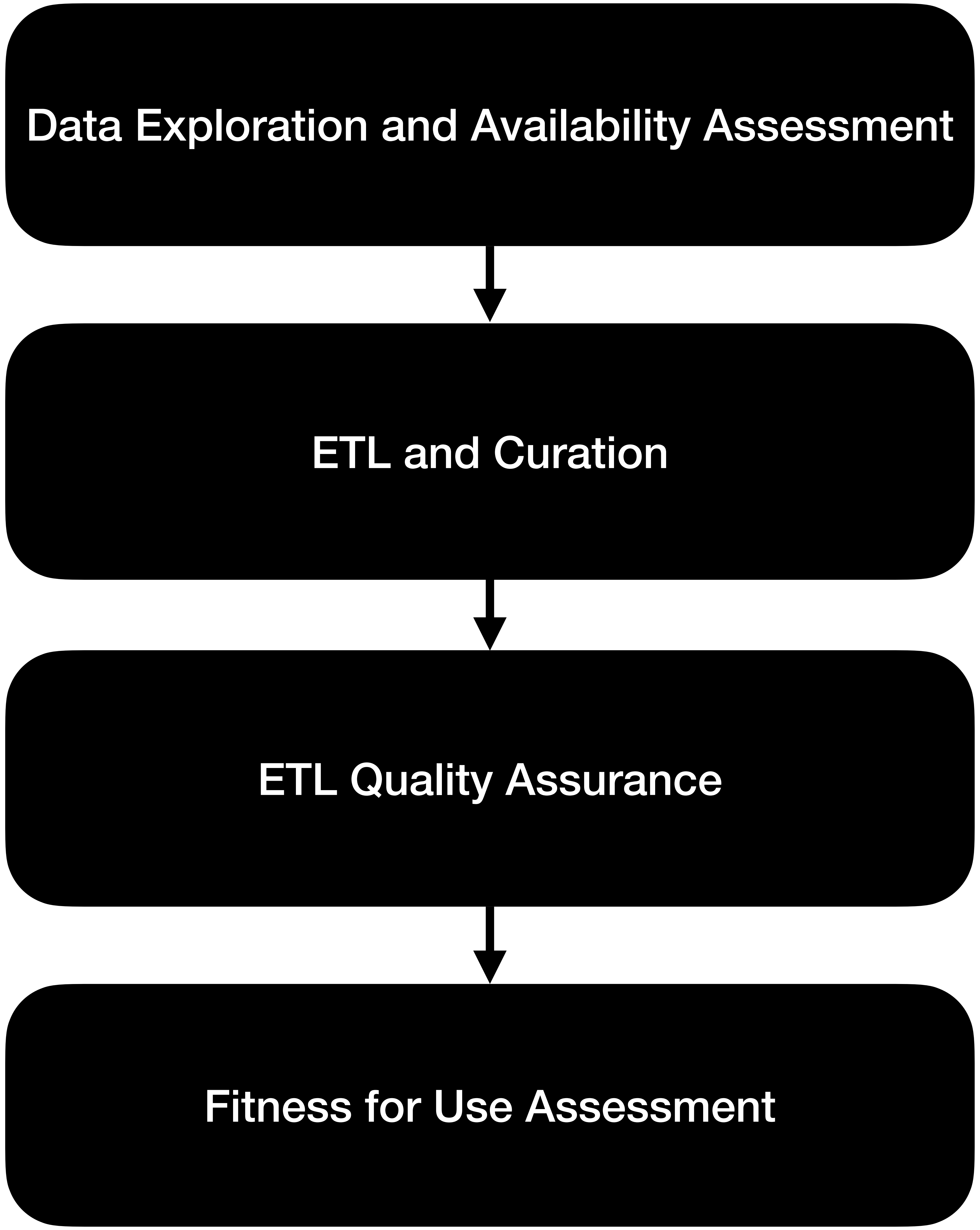
Construct a Hypothesis

- How would one of these potential predictors impact whether a patient is likely to be readmitted to the hospital within 30 days?
  - History of diabetes
  - History of myocardial infarctions
  - Age
  - Length of stay

Clinical Data
Processing

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.

ETL: Extract, Transform, Load



Clinical Data

What is it?

Clinical Data

What is it?

- Clinical documentation
  - Unstructured (e.g., progress notes, medical history, etc.)
  - Structured (e.g., labs, medications, orders, diagnoses, etc.)

Clinical Data

What is it?

- Clinical documentation
  - Unstructured (e.g., progress notes, medical history, etc.)
  - Structured (e.g., labs, medications, orders, diagnoses, etc.)
- Administrative data
  - Billing
  - CMS, Insurance

Clinical Data

What is it?

- Clinical documentation
    - Unstructured (e.g., progress notes, medical history, etc.)
    - Structured (e.g., labs, medications, orders, diagnoses, etc.)
- Administrative data
    - Billing
    - CMS, Insurance
- **Primary purpose is *not* research**

Clinical Data

What are its benefits?

Clinical Data

What are its benefits?

- Decrease costs (time and money)

Clinical Data

What are its benefits?

- Decrease costs (time and money)
- Enable recruitment and retention
  - Rare diseases
  - Underrepresented populations

Clinical Data

What are its benefits?

- Decrease costs (time and money)
- Enable recruitment and retention
  - Rare diseases
  - Underrepresented populations
- Volume, variety, and velocity of data
  - https://www.gartner.com/it-glossary/big-data

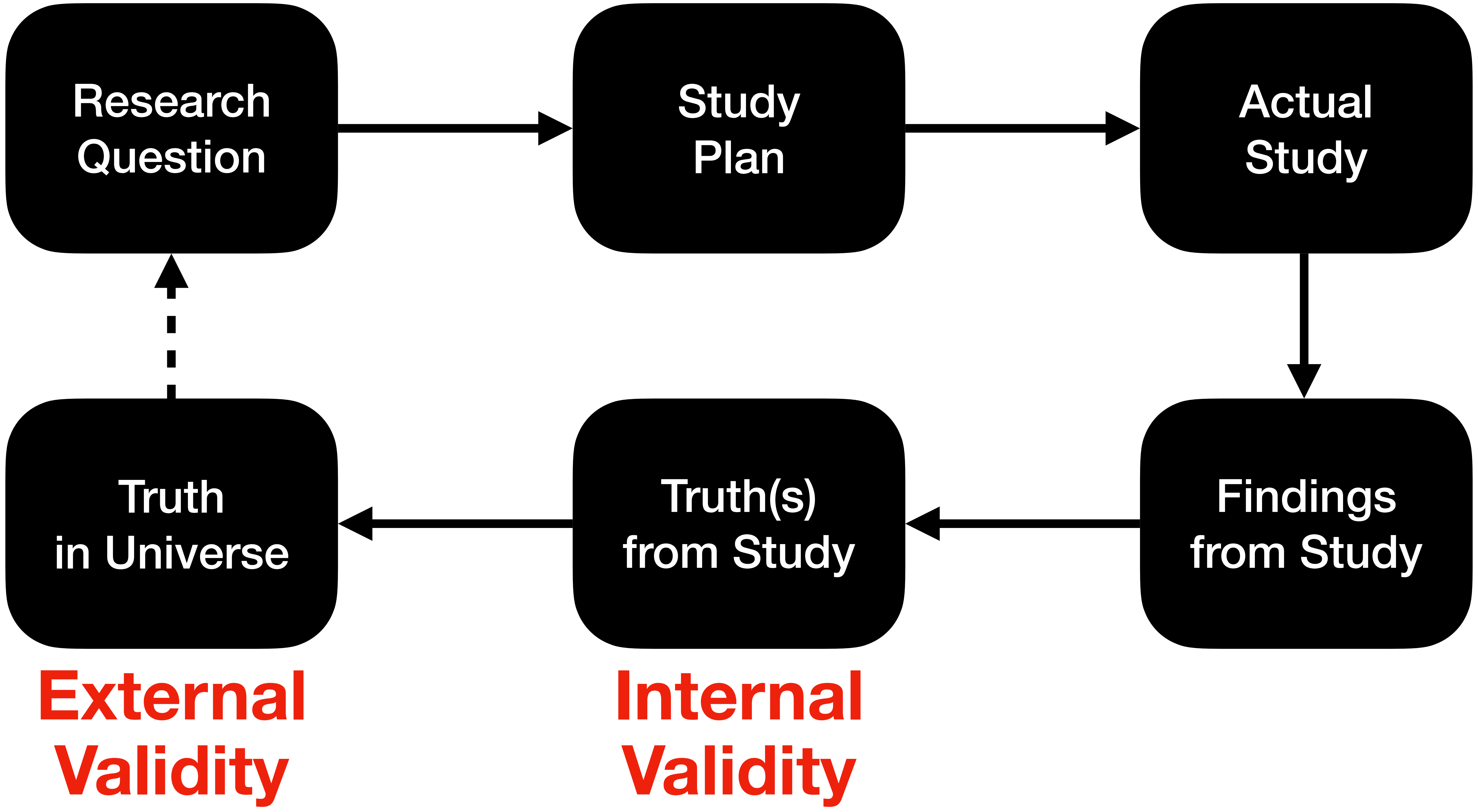
Clinical Data

What are its benefits?

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

Research

Good Research Should Provide Broadly-applicable Truths



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

Clinical Data

Electronic Health Record Data Quality

- **Correctness:** 44–100%
- **Completeness:** 1.1–100%
- Examples
  - Completeness of smoking status: 10–38%
  - Completeness of blood pressure: 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.

Clinical Data

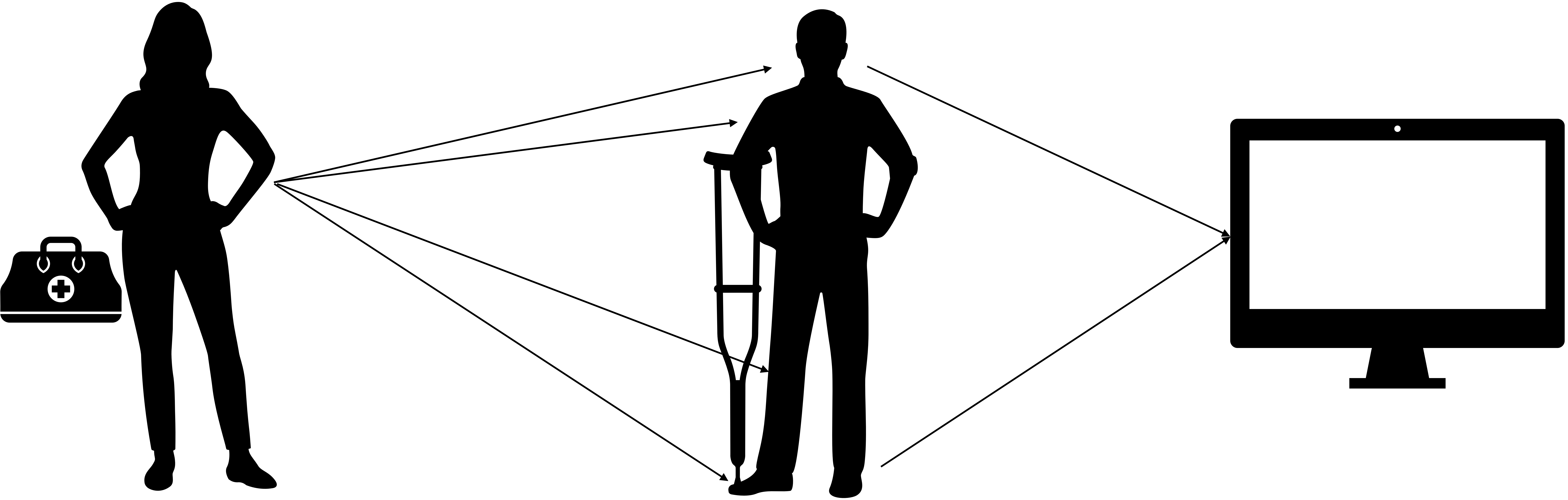
Electronic Health Record Data Quality

- Quality of data is **defined with respect to its intended use case**
  - Clinical data are collected for patient care and billing purposes
- 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

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.

Clinical Data

Not all clinical concepts are observed. Not all observations are recorded.



Missingness
A Brief Introduction

Rubin (1976) Inference and Missing Data.

Missingness
A Brief Introduction

- **MCAR** — Missing Completely at Random
  - Pattern of missingness is **not related** to any other data

Rubin (1976) Inference and Missing Data.

Missingness

A Brief Introduction

- **MCAR** — Missing Completely at Random
  - Pattern of missingness is **not related** to any other data
- **MAR** — Missing at Random
  - Pattern of missingness is **related** to data that are **present**
  - In essence, it is not “random” 🙄

Rubin (1976) Inference and Missing Data.

Missingness

A Brief Introduction

- **MCAR** — Missing Completely at Random
  - Pattern of missingness is **not related** to any other data
- **MAR** — Missing at Random
  - Pattern of missingness is **related** to data that are **present**
  - In essence, it is not “random” 🙄
- **MNAR** — Missing Not at Random
  - Pattern of missingness is **related** to the values of the data that are **missing**

Rubin (1976) Inference and Missing Data.

Missingness

Simplified Example: Height Measurements

	Population Mean	Sample Mean (No Missingness)	Sample Mean (MCAR)	Sample Mean (MAR)	Sample Mean (MNAR)
Men	70.4	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

Rubin (1976) Inference and Missing Data.

Missingness

Simplified Example: Height Measurements

Sample of 200 men and 200 women

	Population Mean	Sample Mean (No Missingness)	Sample Mean (MCAR)	Sample Mean (MAR)	Sample Mean (MNAR)
Men	70.4	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

Rubin (1976) Inference and Missing Data.

Missingness

Simplified Example: Height Measurements

25% of men and women
did not want to share their height

	Population Mean	Sample Mean (No Missingness)	Sample Mean (MCAR)	Sample Mean (MAR)	Sample Mean (MNAR)
Men	70.4	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

Rubin (1976) Inference and Missing Data.

Missingness

Simplified Example: Height Measurements

50% of men
did not want to share their height

	Population Mean	Sample Mean (No Missingness)	Sample Mean (MCAR)	Sample Mean (MAR)	Sample Mean (MNAR)
Men	70.4	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

Rubin (1976) Inference and Missing Data.

Missingness

Simplified Example: Height Measurements

Half of the shortest 25% of men and women
did not want to share their height

	Population Mean	Sample Mean (No Missingness)	Sample Mean (MCAR)	Sample Mean (MAR)	Sample Mean (MNAR)
Men	70.4	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

Rubin (1976) Inference and Missing Data.

Data Quality

An Unsolved Issue

Data Quality
An Unsolved Issue

- **Understand** the provenance of your data
  - **System complexities** and **potential failure points**

Data Quality

An Unsolved Issue

- **Understand** the provenance of your data
  - **System complexities** and **potential failure points**
- **Fitness for Use**
  - Do not think of data quality as an issue of right versus wrong values

Data Quality

An Unsolved Issue

- **Understand** the provenance of your data
  - **System complexities** and **potential failure points**
- **Fitness for Use**
  - Do not think of data quality as an issue of right versus wrong values
- **Systematic** data quality problems can drastically alter results
  - Data that are “bad” at random are not always an issue in research

Data Quality
An Unsolved Issue

- **Understand** the provenance of your data
  - **System complexities** and **potential failure points**
- **Fitness for Use**
  - Do not think of data quality as an issue of right versus wrong values
- **Systematic** data quality problems can drastically alter results
  - Data that are “bad” at random are not always an issue in research
- When you uncover potential data quality problems, be thoughtful in your attempts to compensate