Clinical Data Wrangling

Session 2: Understanding the Data (Problems)

Introduction to EHR Data Quality

Nicole G Weiskopf, 8/21/18

Learning Objectives

- What is “data wrangling?”
- Role of data wrangling in clinical data reuse
- Why data wrangling and data quality matter
- What “data quality” means
- Potential impact of data quality
- Basics of data quality assessment

What is data wrangling?

Very broadly, data wrangling is the process of making your source data actionable.

In our case, that means taking clinical data from the EHR and getting it into the proper state for clinical research.

Data wrangling is largely “hidden”

- There is *a lot* of pre-processing involved in the reuse of EHR data, but most “consumers” don’t know about it
  - E.g., data mapping, transformation, and cleaning
- This is somewhat analagous to wet lab work, but with some key differences
  - Data wrangling is often ad hoc
  - Limited transparency

Y
k

because there isn’t a right way. But we are going to teach you the basics of a systematic approach and get you thinking about the

d
s

process and underlying data issues may have on your findings.

A Real Life Example

Increase in rates of maternal mortality in Texas reported in 2016.

> *"The rate of Texas women who died from complications related to pregnancy doubled from 2010 to 2014, a new study has found, for an estimated maternal mortality rate that is unmatched in any other state and the rest of the developed world."*

The Guardian, 2016: https://www.theguardian.com/us-news/2016/aug/20/texas-maternal-mortality-rate-health-clinics-funding

A Real Life Example

Original Research

Recent Increases in the U.S. Maternal Mortality Rate

Disentangling Trends From Measurement Issues

Marian F. MacDorman, PhD, Eugene Declercq, PhD, Howard Cabral, PhD, and Christine Morton, PhD

RESULTS: The estimated maternal mortality rate (per 100,000 live births) for 48 states and Washington, DC (excluding California and Texas, analyzed separately) increased by 26.6%, from 18.8 in 2000 to 23.8 in 2014. California showed a declining trend, whereas Texas had a sudden increase in 2011–2012. Analysis of the measurement change suggests that U.S. rates in the early 2000s were higher than previously reported.

MacDorman MF et al. Is the United States Maternal Mortality Rate Increasing? Disentangling trends from measurement issues Short title: US Maternal Mortality Trends. Obstetrics and gynecology. 2016 Sep;128(3):447.

A Real Life Example

The Guardian

Texas has highest maternal mortality rate in developed world, study finds

San Antonio Express-News

Alarming number of black Texas moms dying in childbirth

HPPR

After Texas Slashed Women's Health Funding, Pregnancy-Related Deaths Doubled

From Twitter

Wil Not Be Refereed
@WilGafney

Heard an ad from Beto O'Roark today. It was good, strong. He called Texas out for being the most underinsured state in the nation, for its high maternal mortality rates, and their disproportionate effect on black women.

Twitter • 2 retweets • 9/14/18 6:36 PM

All coverage

The Huffington Post

It's Not Just Texas. Maternal Deaths Are High Across The U.S.

New Scientist

US pregnancy-related deaths are rising and have doubled in Texas

Texas Tribune

Texas Sees "Unusual" Spike in Pregnancy-Related Deaths, Study Finds

Huffington Post

Pregnancy-Related Deaths Nearly Doubled In Texas After Cuts To Women's Health

A Real Life Example

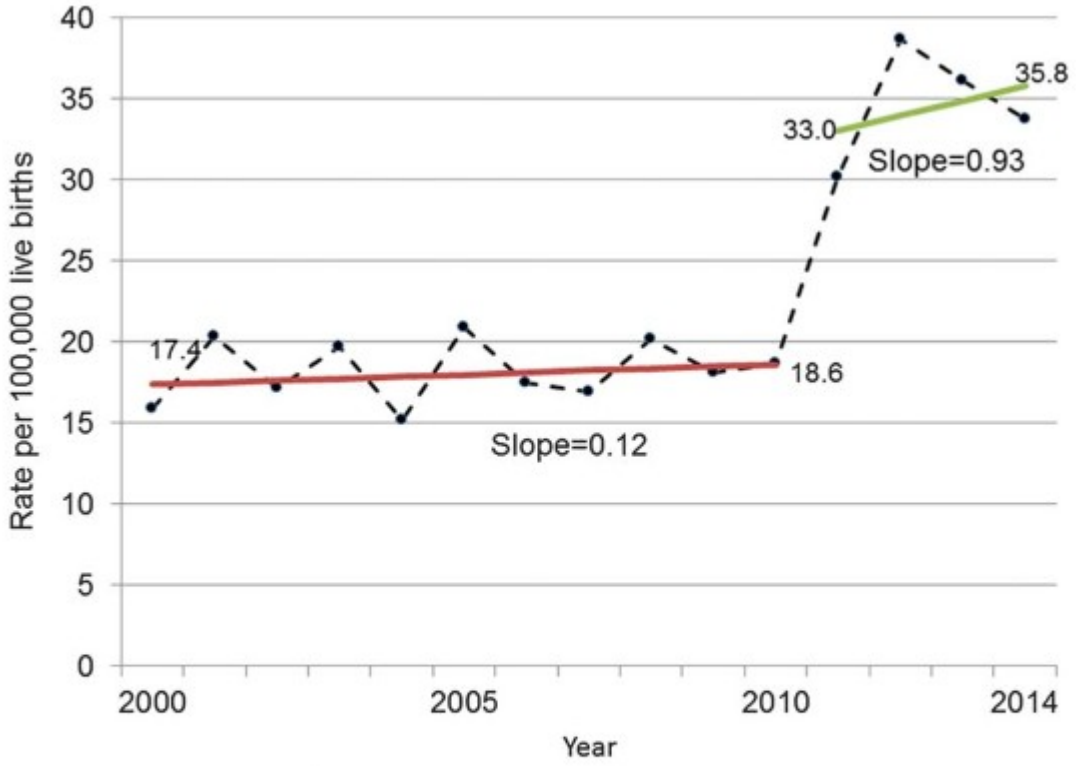


Fig. 4. Adjusted maternal mortality rates, Texas, 2000–2014. Texas revised to the U.S. standard pregnancy question in 2006. The unrevised question asked about pregnancies within the past 12 months.

MacDorman MF et al. Is the United States Maternal Mortality Rate Increasing? Disentangling trends from measurement issues. Obstetrics and gynecology. 2016 Sep;128(3):447.

A Real Life Example

WaPo: Texas's maternal mortality rate was unbelievably high. Now we know why

"….the Texas Maternal Mortality and Morbidity Task Force …. cross-referenced death certificates, birth certificates and a year's worth of medical records for all 147 women in the state's records. They found that, in fact, there were 56 deaths that fell under the definition of maternal mortality — any pregnancy-related death while a woman is pregnant or within 42 days of giving birth, excluding accidental or incidental causes such as car crashes or homicide.

"After all of the data-collection errors were excluded, Texas's 2012 maternal mortality rate was corrected from 38.4 deaths per 100,000 live births to 14.6 per 100,000 live births."

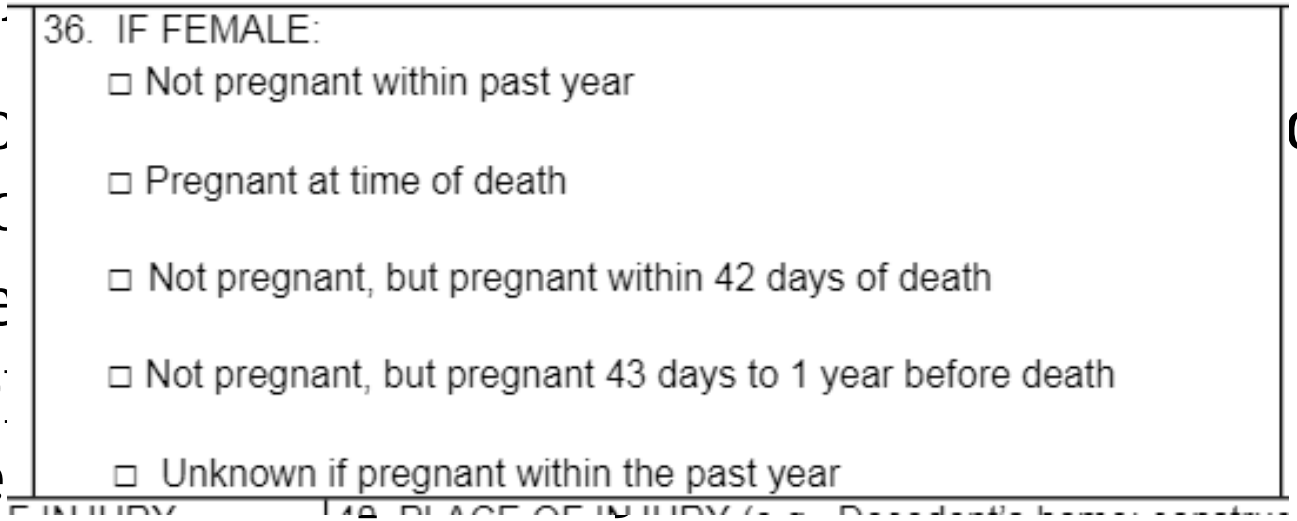
https://www.washingtonpost.com/news/morning-mix/wp/2018/04/11/texas-maternal-mortality-rate-was-unbelievably-high-now-we-know-why/?noredirect=on&utm_term=.a037fddba059

How Many American Women Die From Causes Related to Pregnancy or Childbirth? No One Knows.

Data collection on maternal deaths is so flawed and under-funded that the federal government no longer even publishes an official death rate.

by Robin Fields and Joe Sexton, Oct. 23, 2017, 8 a.m. EDT

- Historically, maternal death data come from dea
- Pric d to
  rec
- Afte
  ado
- The
  negatives, but also increased false positives


36. IF FEMALE:

☐ Not pregnant within past year

☐ Pregnant at time of death

☐ Not pregnant, but pregnant within 42 days of death

☐ Not pregnant, but pregnant 43 days to 1 year before death

☐ Unknown if pregnant within the past year

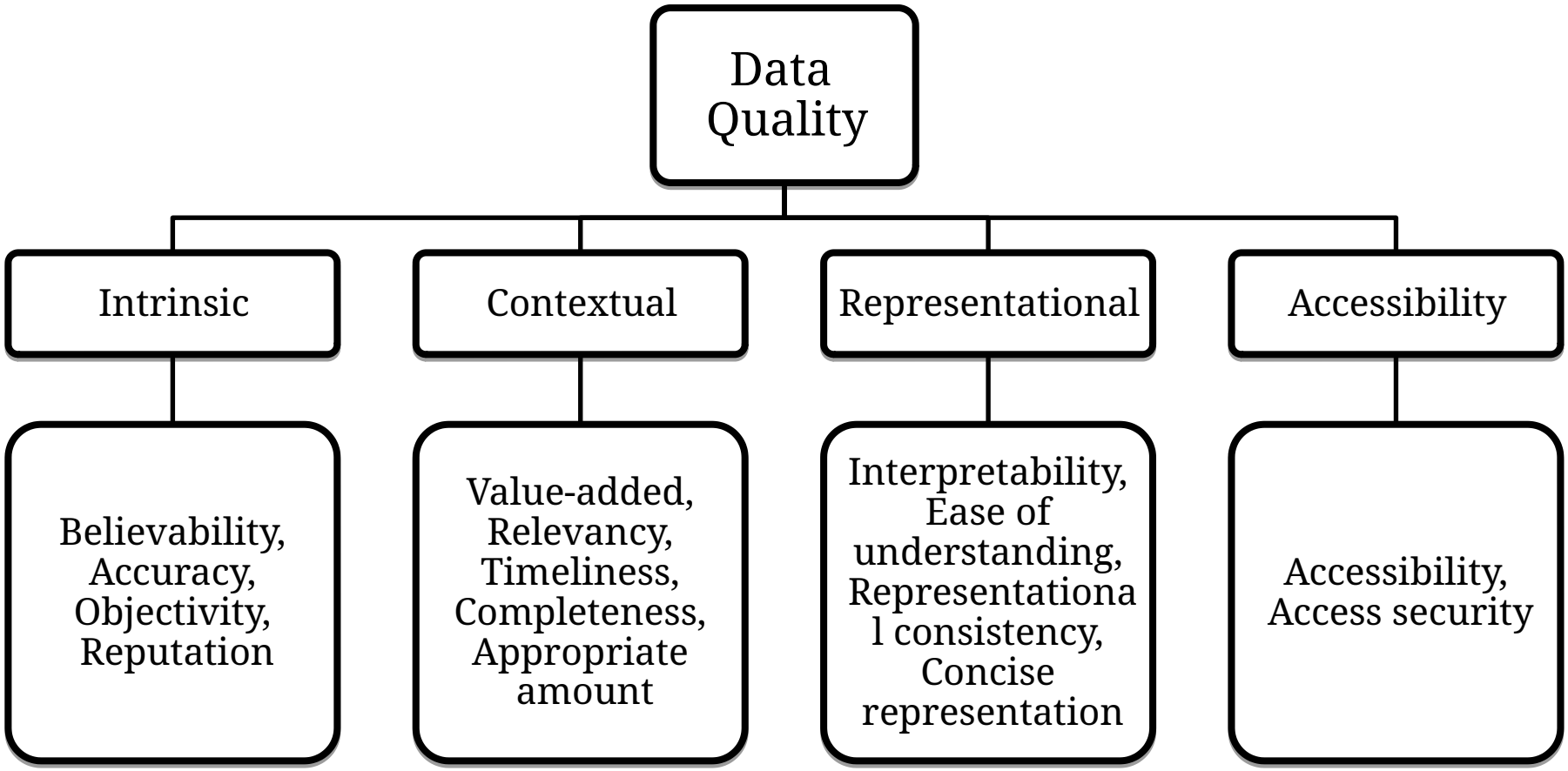
https://www.propublica.org/article/how-many-american-women-die-from-causes-related-to-pregnancy-or-childbirth

Hopefully I've convinced you that data quality matters, but what does it actually mean?

"Data are of high quality if they are fit for their intended uses in operations, decision making, and planning. Data are fit for use if they are free of defects and possess desired features."

Redman, T (2001) Data quality: the field guide. Based on Juran's work.

Wang & Strong (1996)
Beyond accuracy:
What data quality means to data consumers



Wang & Strong (1996)
Beyond accuracy:
What data quality means to data consumers

Data wrangling processes that take highly complex EHR data and transform them into flat files also transform underlying data quality problems related to structure, representation, and accessibility to **presence or absence** of data. This is why EHR-focused models of data quality are generally simpler than, for example, Wang and Strong's.

(If you talk to clinicians, who deal with the upstream data, you're likely to hear a lot about issues relating to data overload, unstructured text, fragmentation, etc.)

What *is* the quality of EHR data?

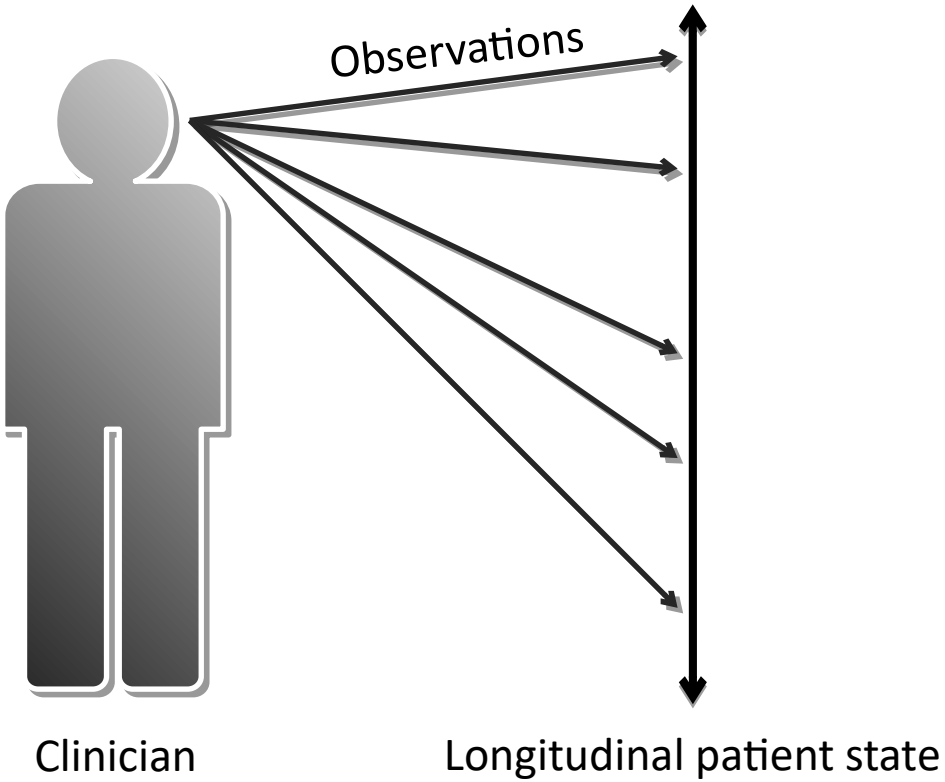
- Hogan and Wagner (1997)
  - Correctness: 44% - 100%
  - Completeness: 1.1% - 100%
- Chan et al. (2010)
  - Completeness of BP: 0.1% – 51%

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

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

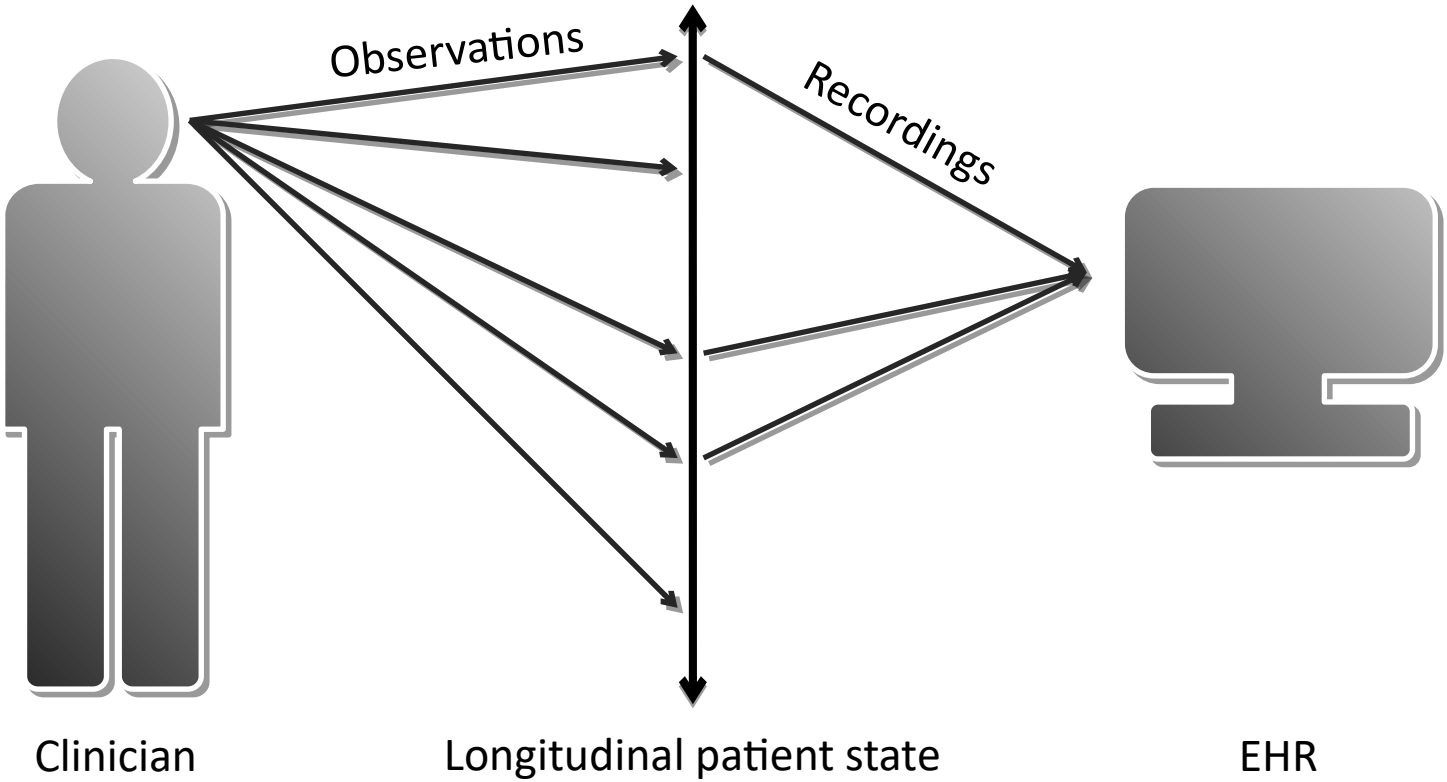
- A lot of this is because the quality of the data is defined with respect to the intended use of the data (fitness for use)
- But also because the processes involved in taking a clinical truth about a patient all the way to a dataset being used for research is fraught with pitfalls

Data can be observed or unobserved…

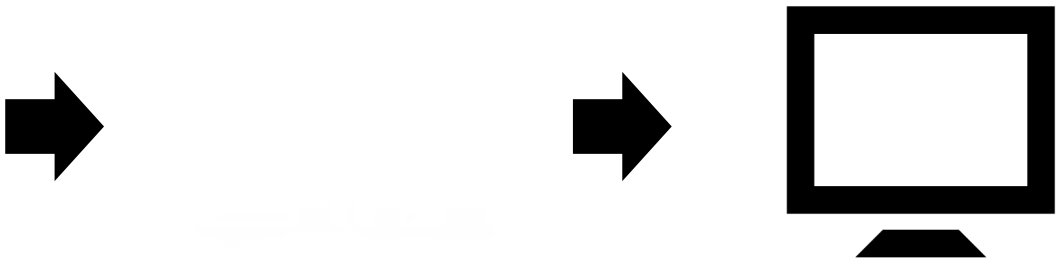


Weiskopf et al. (2013) Defining and measuring completeness of EHRs for secondary use

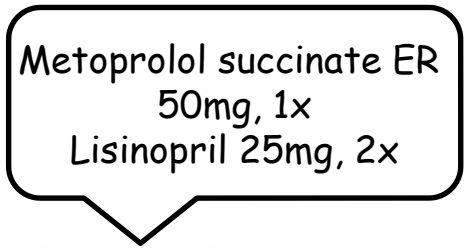
…and recorded or unrecorded

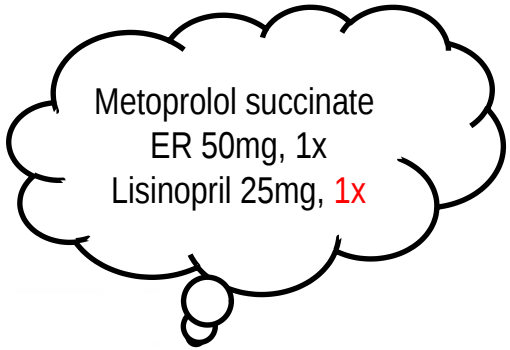


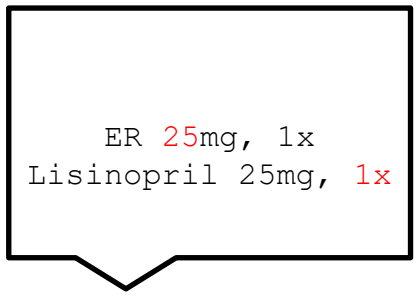
Weiskopf et al. (2013) Defining and measuring completeness of EHRs for secondary use

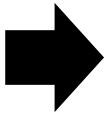

Make
Observations
Record
Observations

M

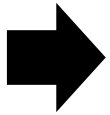








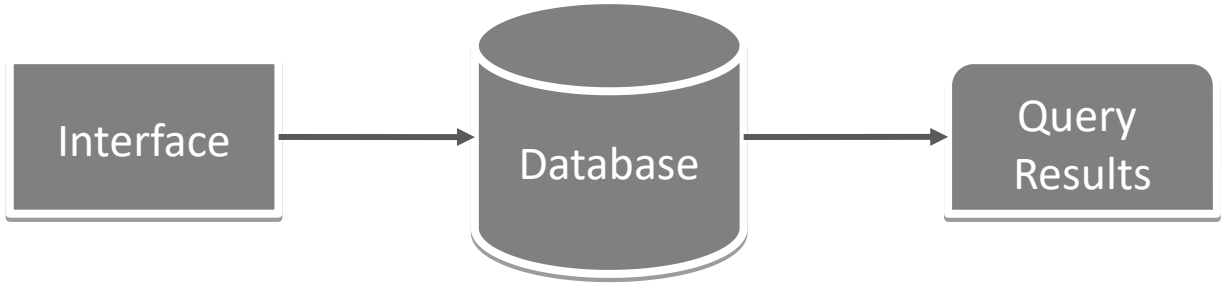
Make
Observations



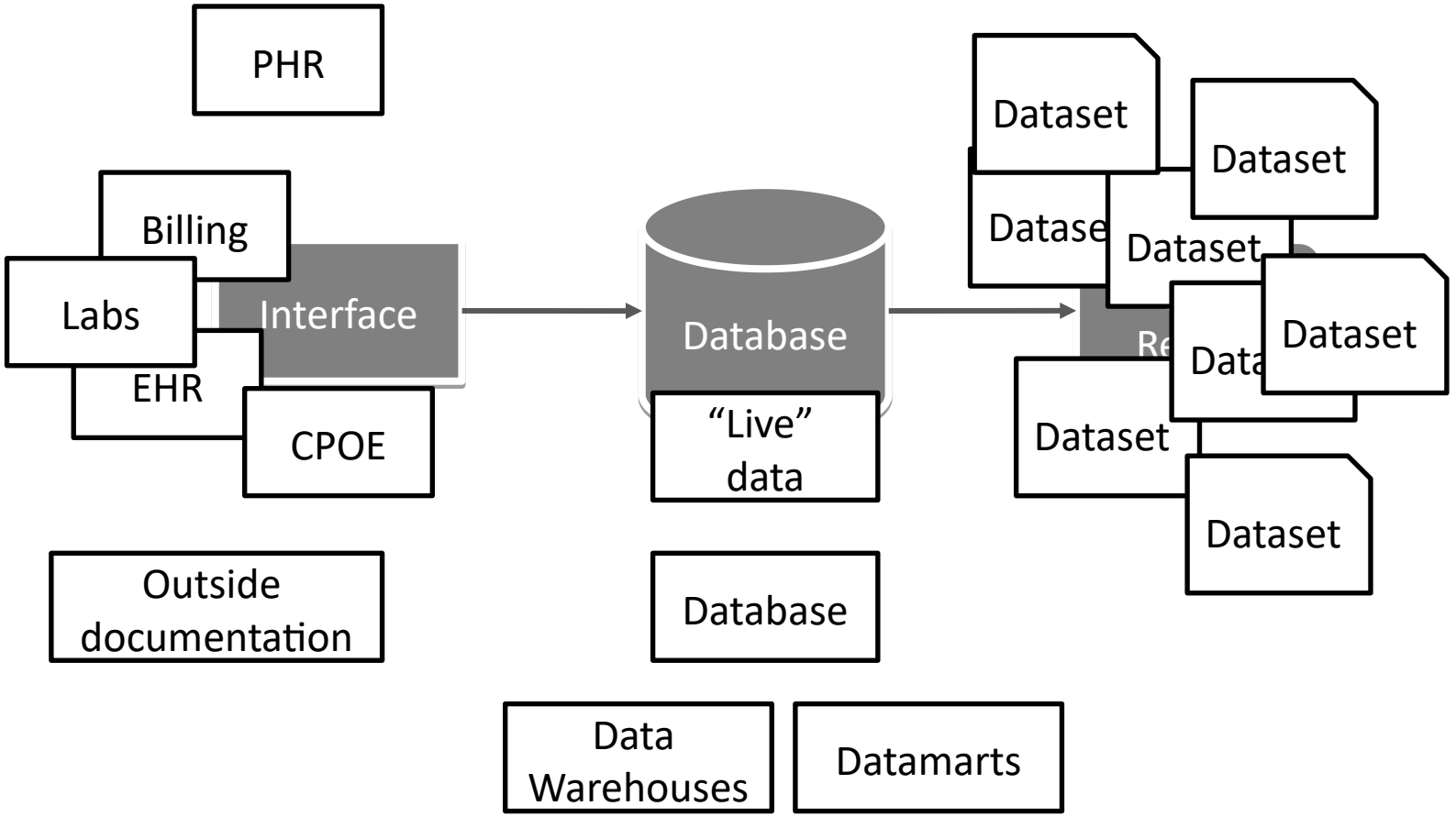
Record
Observations

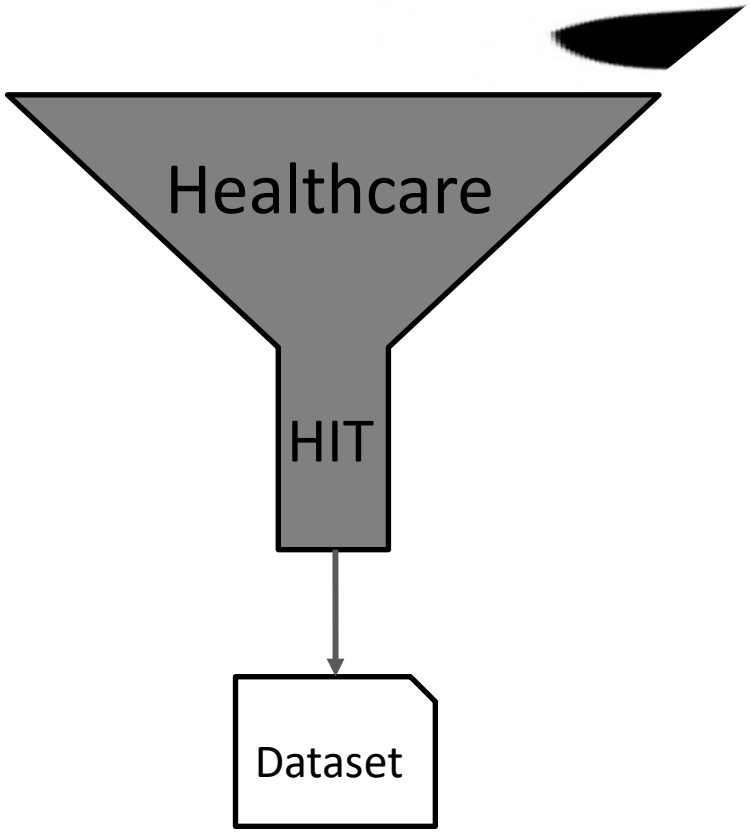
Multi-vitamin, 1x
Metoprolol succinate ER 50mg, 1x
Lisinopril 25mg, 2x

“Traditional” Data



Healthcare Data




Healthcare
HIT
Dataset

As an aside, deep understanding of how and when bias is introduced may lead to methods to “undo” that bias

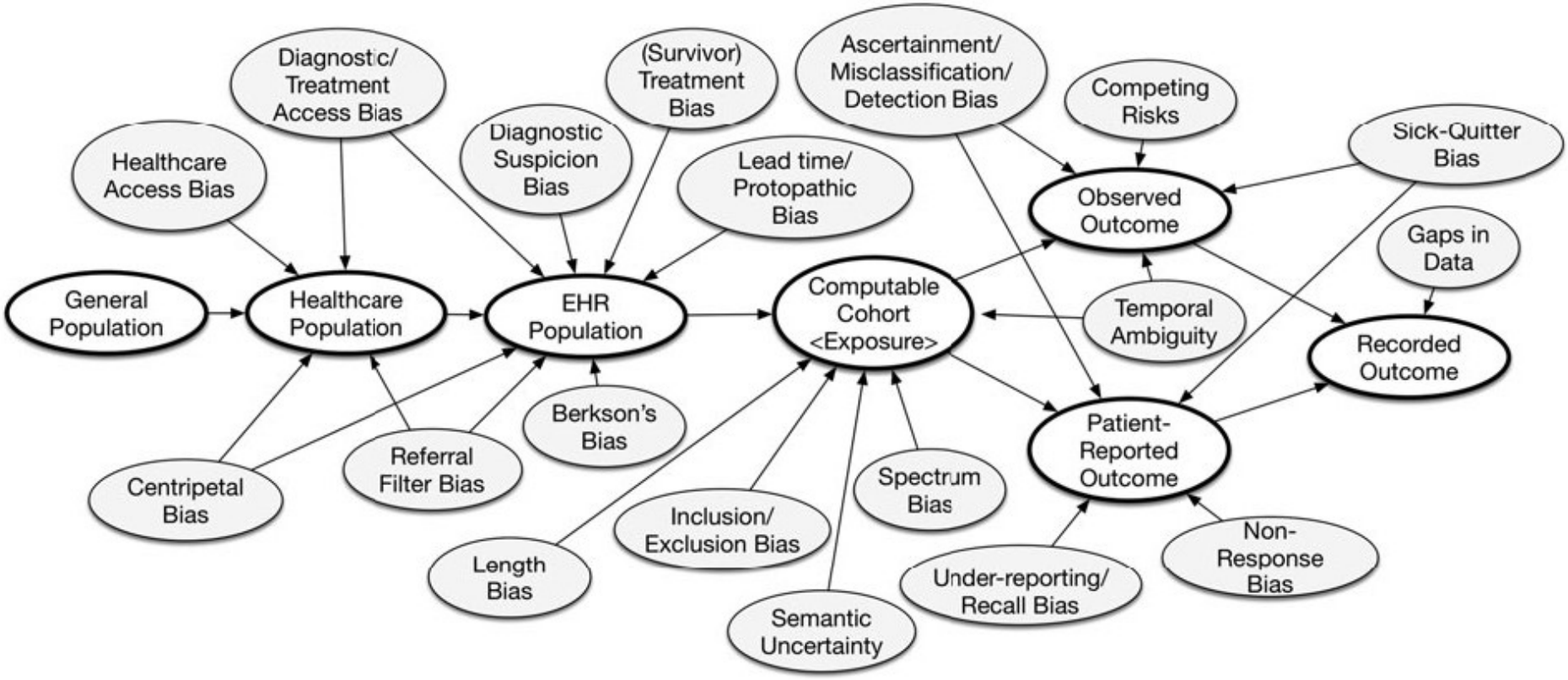
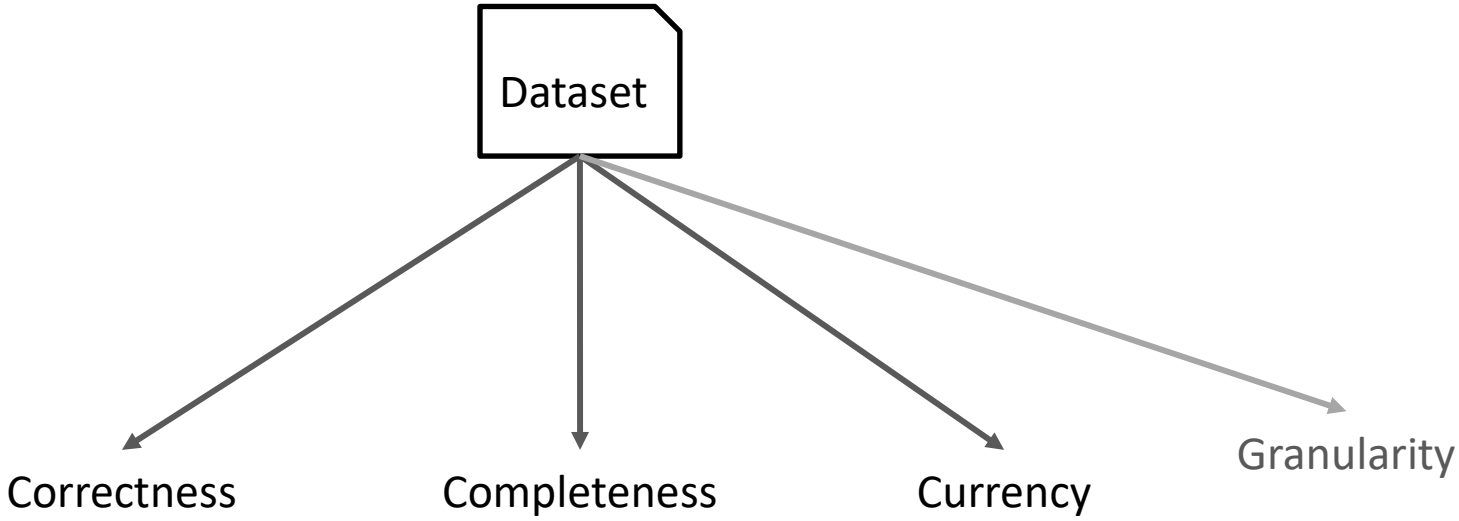
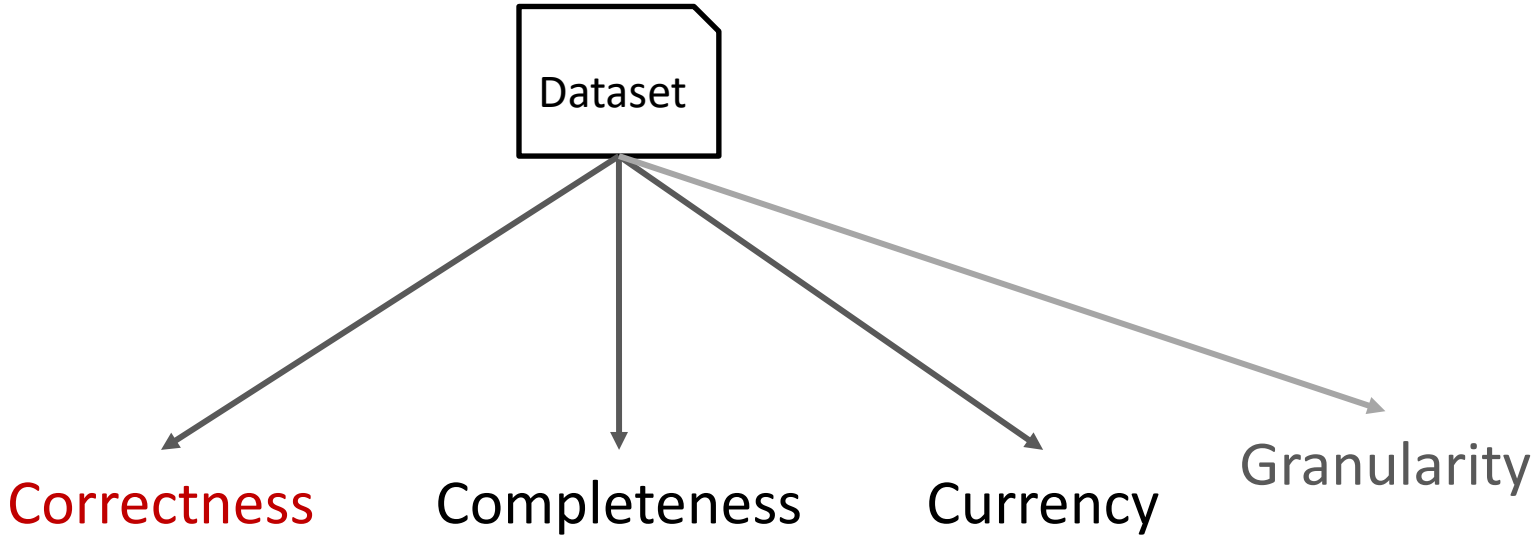


FIGURE 7 Bayesian meta-model for debiasing. Specific biases come from Delgado-Rodriguez[56]

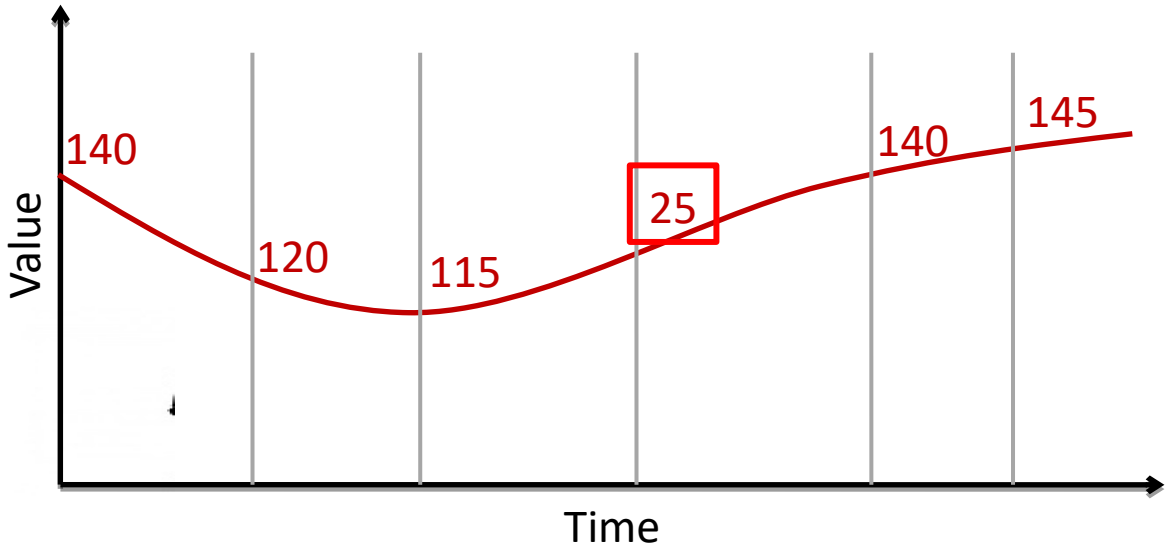
Lehmann HP, Downs SM. Desiderata for Computable Biomedical Knowledge for Learning Health Systems. Learn Heal Syst. 2018;e10065:1–9.

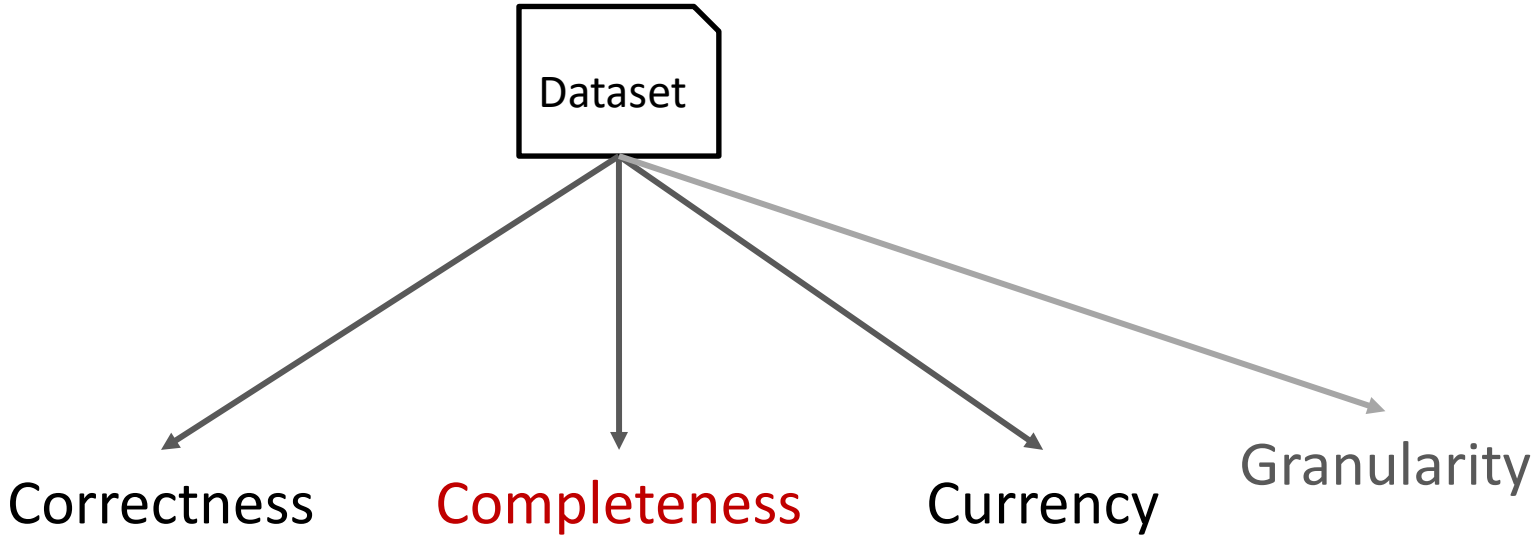
What types of data quality problems do we run into when we reuse clinical data?


Dataset
Correctness
Completeness
Currency
Granularity

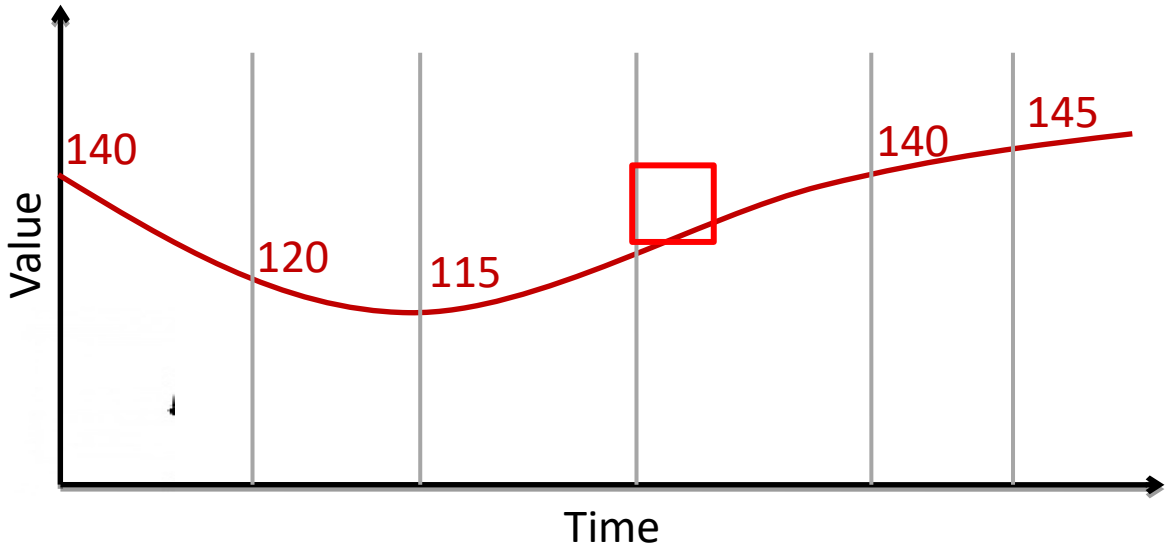


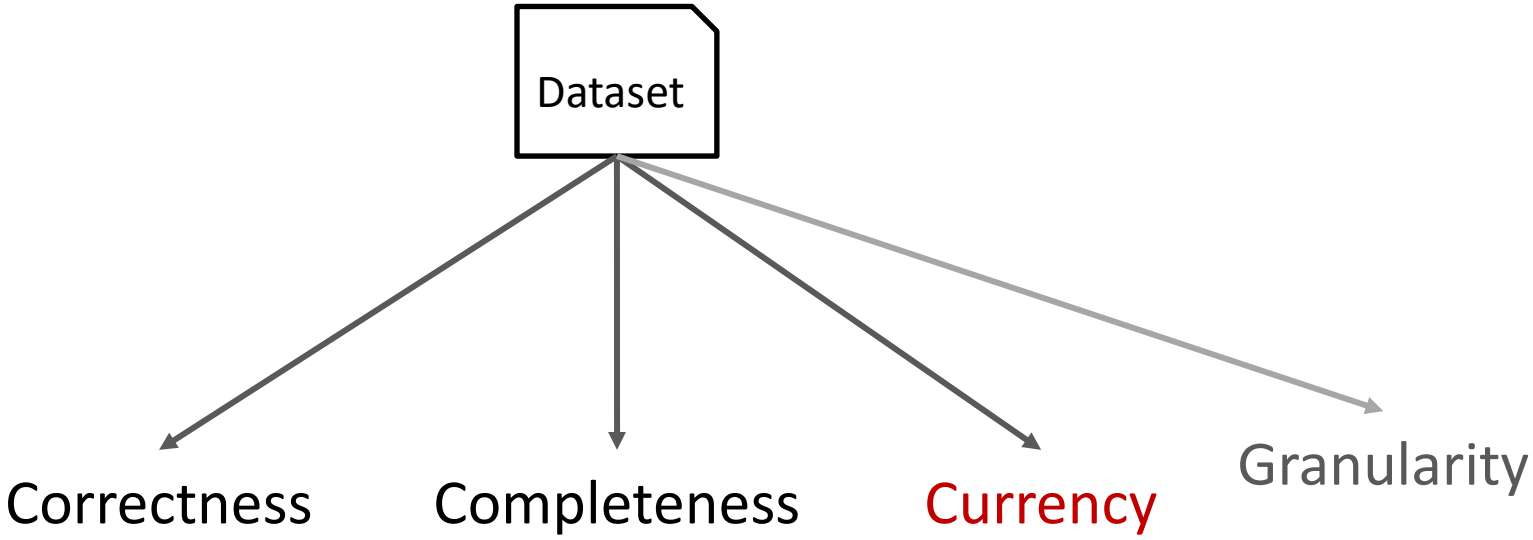
An element that is present in the EHR is true.



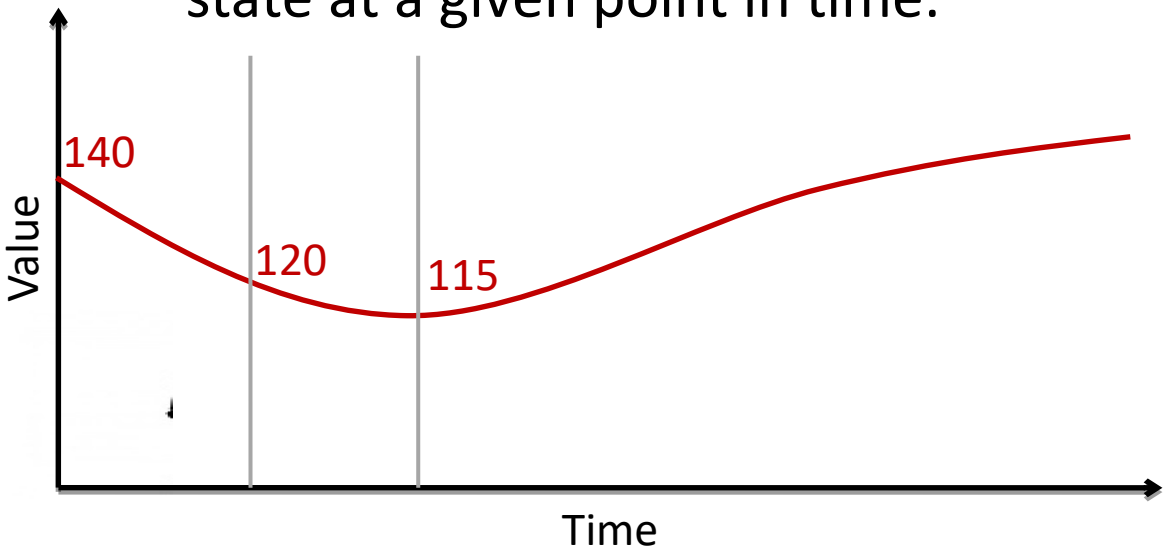


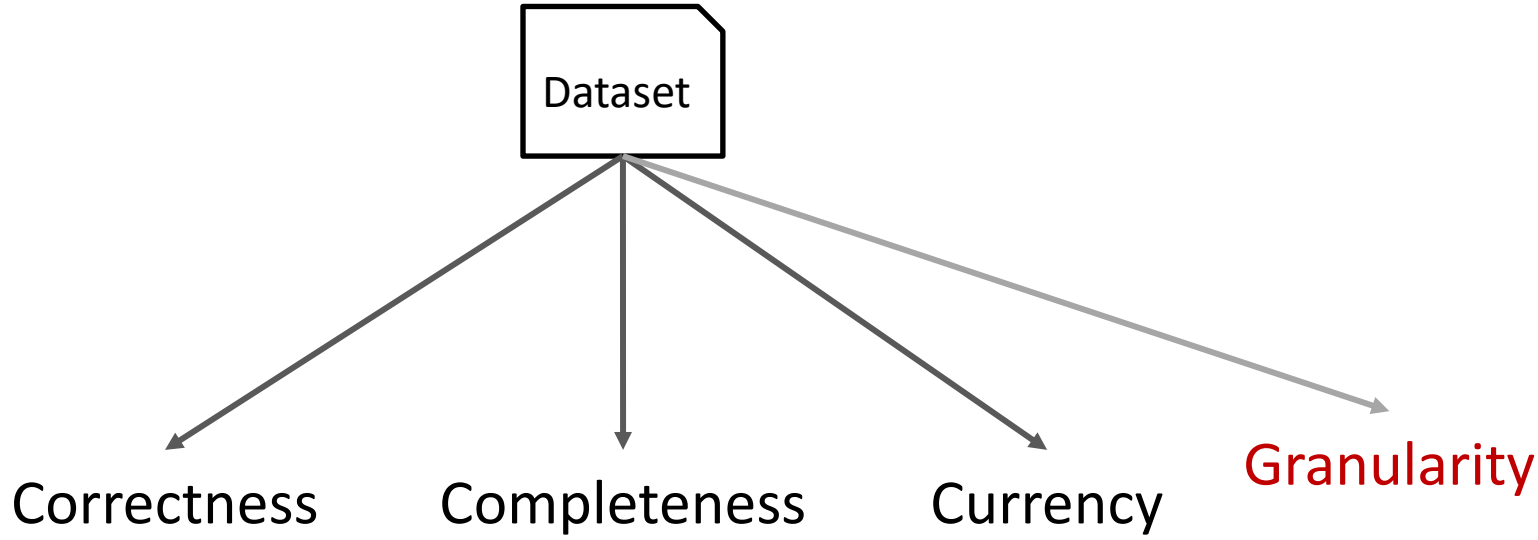
A truth about a patient is present in the EHR.



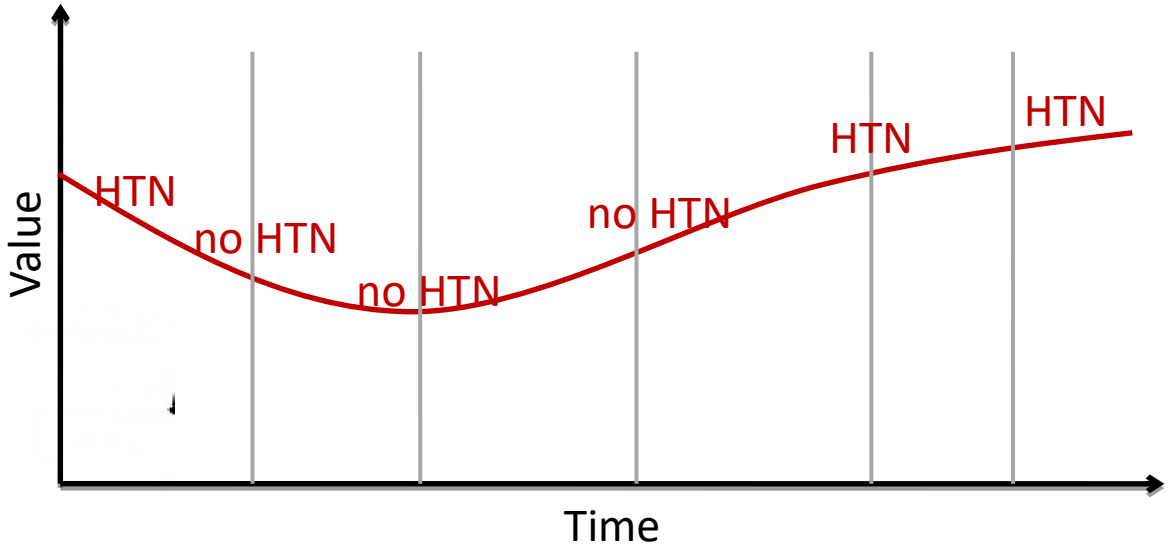


An element in the EHR a relevant representation of the patient state at a given point in time.



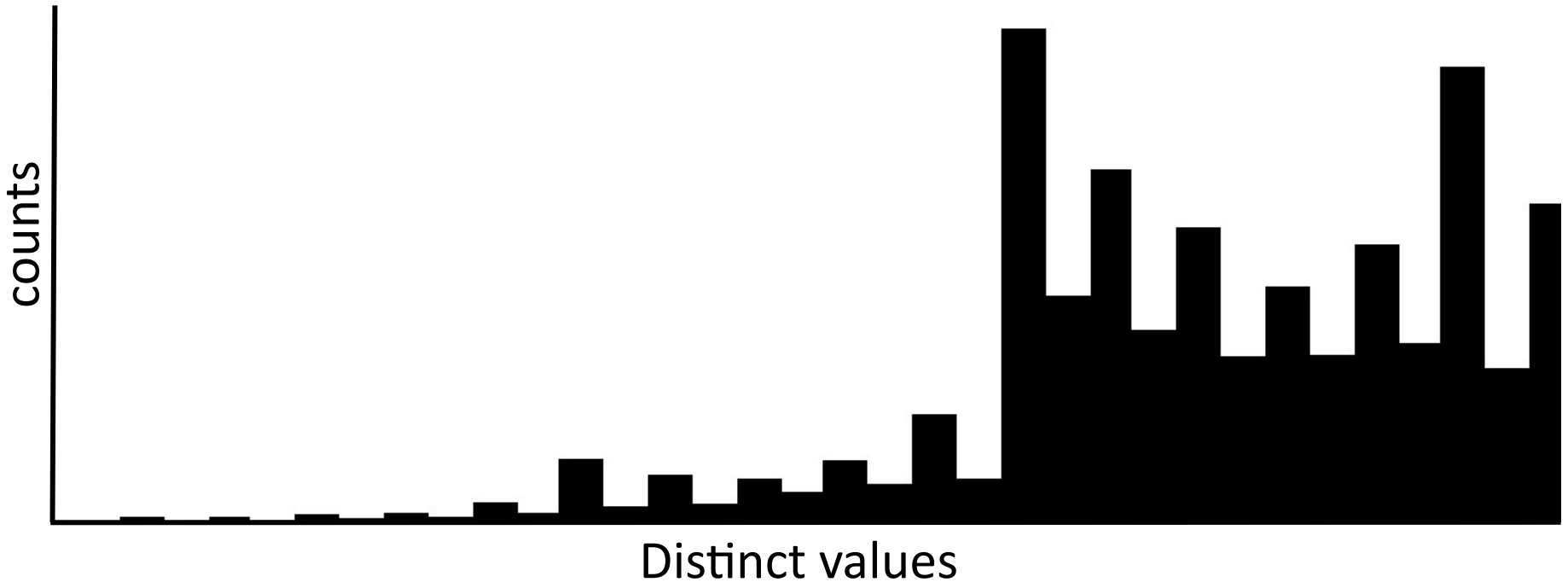


An element in the EHR contains the appropriate amount of information.



When you seek to understand the quality
data, quantification of the problem (errors,
m

think about the actual impact.



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

Not Missing

RID	systolic	diastolic	age
000000	120	90	50
111111	125	100	45
222222	100	80	38
333333	105	75	36
444444	85	60	32
555555	90	65	42
666666	135	95	64
777777	87	59	52
888888	120	80	47
999999	115	75	43

Actual Averages
Systolic: 108
Diastolic: 80

Missing Completely at Random

RID	systolic	diastolic	Age
000000	120	90	50
111111	125	100	45
222222	100	80	38
333333	105	75	36
444444	85	60	32
555555	90	65	42
666666	135	95	64
777777	87	59	52
888888	120	80	47
999999	115	75	43

Actual Averages
Systolic: 108
Diastolic: 80

MCAR Obs. Averages
Systolic: 111
Diastolic: 76

Missing at Random
(conditioned on age)

RID	systolic	diastolic	age
000000	120	90	50
111111	125	100	45
222222	100	80	38
333333	105	95	36
444444	85	60	32
555555	90	65	42
666666	135	95	64
777777	87	59	52
888888	120	80	47
999999	115	75	43

You can control for the effect of age.

Actual Averages
Systolic: 108
Diastolic: 80

MCAR Obs. Averages
Systolic: 111
Diastolic: 76

MAR Obs. Averages
Systolic: 113
Diastolic: 81

Missing Not at Random
(conditioned on missing data)

RID	systolic	diastolic	age
000000	120	90	50
111111	125	100	45
222222	100	80	38
333333	105	75	36
444444	85	60	32
555555	90	65	42
666666	135	95	64
777777	87	59	52
888888	120	80	47
999999	115	75	43

You can control for the effect of data that aren't there.

Actual Averages
Systolic: 108
Diastolic: 80

MCAR Obs. Averages
Systolic: 111
Diastolic: 76

MAR Obs. Averages
Systolic: 113
Diastolic: 81

MNAR Obs. Averages
Systolic: 117
Diastolic: 85

So what should we do about all of this?

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, 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.