

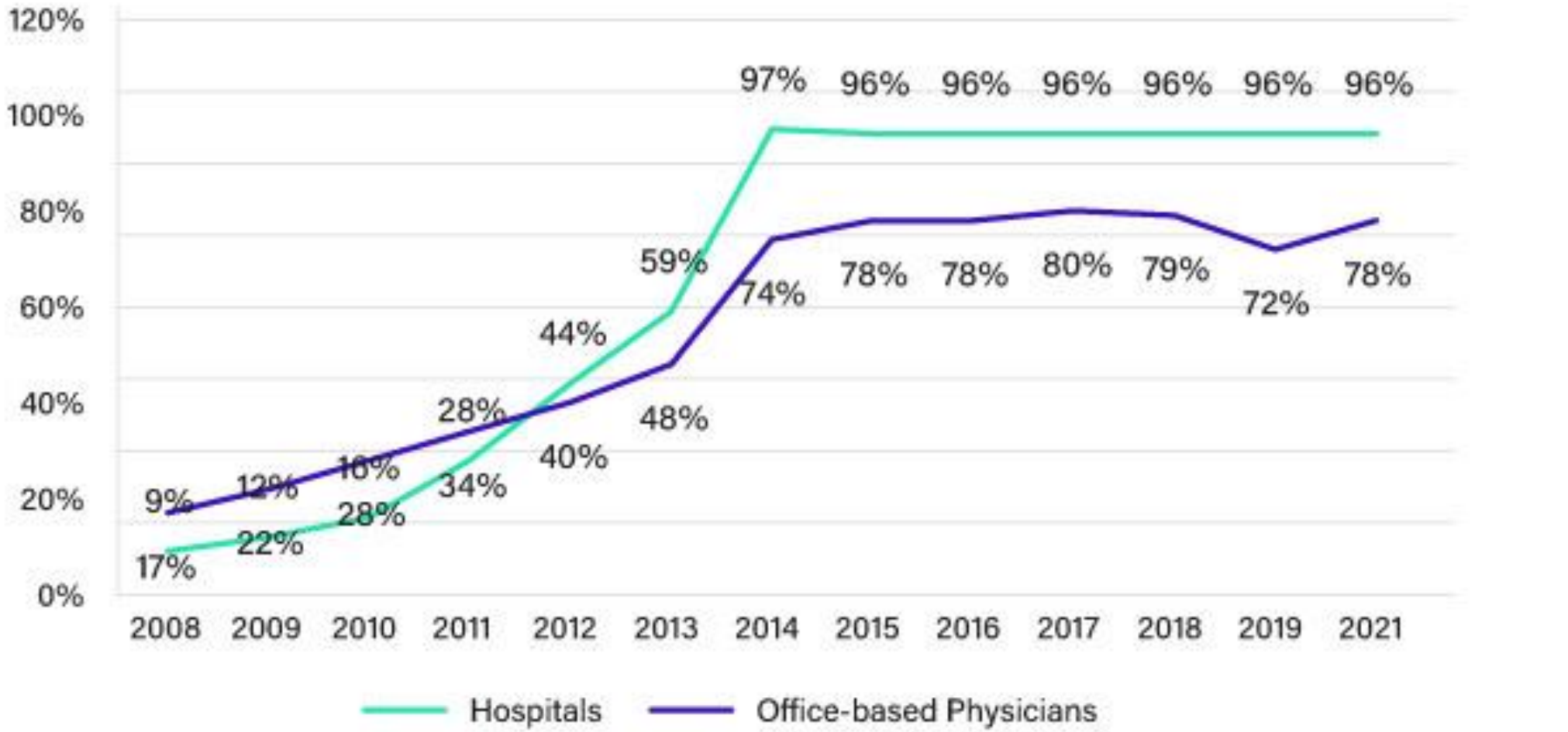
Data 146: Foundations for CPH

Health Data and Data Quality

Irene Y. Chen

Towards HEALTHCARE

Trend in U.S. Hospitals and Physician EHR Adoption



Source: www.towardshealthcare.com

Outline

- **EHR Adoption** (15 mins)
- **MIMIC Deep Dive** (15 mins)
- **Other Health Data** (20 mins)



How can we make Data 146 better for you?

Learning Objective: Understand how EHR data came to be and potential issues

Why did hospitals adopt the electronic health record?



HITECH Act of 2009

HITECH Act

- Health Information Technology for Economic and Clinical Health (HITECH) ACT of 2009

- Of $840 billion stimulus package in 2009, **$30 billion** was allotted for hospitals and medical providers for "**meaningful use**" of certified EHR systems
  - Payments linked to improvements in quality, safety, and efficiency



Health IT Policy Committee

- Created under the HITECH Act as an advisory committee
- Recommend policies and defined "**meaningful use**" of EHRs
- Included federal appointees, representatives from healthcare, technology, academia, and patient advocates
- Included **Judy Faulkner, CEO of Epic**

PROFILE

Judy Faulkner

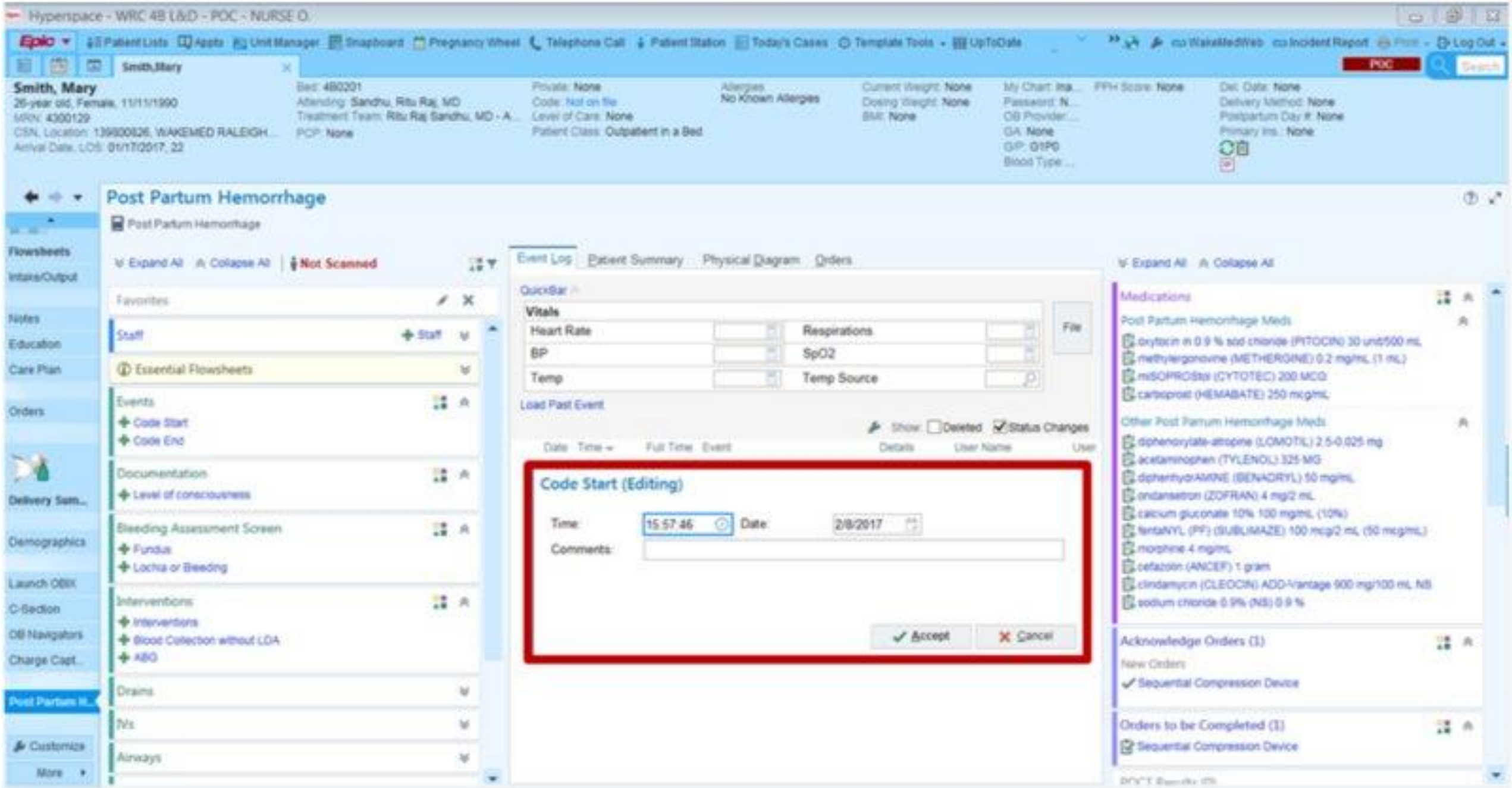
Founder & CEO, Epic Systems

$7.8B **Real Time Net Worth**
as of 10/28/25
#469 in the world today

$7.8B
2025 America's Richest Self-Made Women Net Worth
as of 6/3/25



PHOTO BY COURTESY JUDY FAULKNER


Hyperspace - WRC 4B L&D - POC - NURSE O
Epic
Patient Lists
Appts
Unit Manager
Snapboard
Pregnancy Wheel
Telephone Call
Patient Station
Today's Cases
Template Tools
UpToDate
Log Out
Smith,Mary
POC
Smith, Mary
26-year old, Female, 11/11/1990
MRN: 4300129
Allergies
No Known Allergies
Post Partum Hemorrhage
Post Partum Hemorrhage
Flowsheets
Intake/Output
Notes
Education
Care Plan
Orders
Delivery Sum...
Demographics
Launch OBIX
C-Section
OB Navigators
Charge Capt...
Post Partum H...
Customize
More
Expand All
Collapse All
Not Scanned
Favorites
Staff
Essential Flowsheets
Events
Code Start
Code End
Documentation
Level of consciousness
Bleeding Assessment Screen
Fundus
Lochia or Bleeding
Interventions
Interventions
Blood Collection without LOA
ABG
Drains
IVs
Airways
Event Log
Patient Summary
Physical Diagram
Orders
QuickBar
Vitals
Heart Rate
Respirations
BP
SpO2
Temp
Temp Source
File
Load Past Event
Show
Deleted
Status Changes
Date
Time
Full Time
Event
Details
User Name
User
Code Start (Editing)
Time:
15:57:46
Date:
2/8/2017
Comments:
Accept
Cancel
Expand All
Collapse All
Medications
Post Partum Hemorrhage Meds
oxytocin in 0.9 % sod chloride (PITOCIN) 30 unit/500 mL
methylergonovine (METHERGINE) 0.2 mg/mL (1 mL)
misoprostol (CYTOTEC) 200 MCG
carboprost (HEMABATE) 250 mcg/mL
Other Post Partum Hemorrhage Meds
diphenoxylate-atropine (LOMOTIL) 2.5-0.025 mg
acetaminophen (TYLENOL) 325 MG
diphenhydrAMINE (BENADRYL) 50 mg/mL
ondansetron (ZOFRAN) 4 mg/2 mL
calcium gluconate 10% 100 mg/mL (10%)
fentaNYL (PF) (SUBLIMAZE) 100 mcg/2 mL (50 mcg/mL)
morphine 4 mg/mL
cefazolin (ANCEF) 1 gram
clindamycin (CLEOCIN) ADD-Vantage 900 mg/100 mL NS
sodium chloride 0.9% (NS) 0.9 %
Acknowledge Orders (1)
New Orders
Sequential Compression Device
Orders to be Completed (1)
Sequential Compression Device

What is “meaningful use”?

- Data capture and sharing
- Hospital had to demonstrate basic capabilities about data entry and clinical quality measures
- Record demographics, problem lists, medications, lab results, electronic prescribing

What is missing here?
(Say out loud)

Problems with original recommendations

- **Limited data exchange**: No need to share data or facilitate exchange with patients or other hospitals
- **Vendor lock-in**: Each EHR vendor has its own data format. No cross-vendor interoperability means vendors are incentivized to silo in hospitals and have high switching costs
- **Improved clinical outcomes**: More documentation could slow down clinical workflows. How do you incentivize structures that actually improve clinical care?
- **Patient rights**: Hospitals control all of the data and nothing is written with the patient in mind.

PERSPECTIVE

Escaping the EHR Trap — The Future of Health IT

Authors: Kenneth D. Mandl, M.D., M.P.H., and Isaac S. Kohane, M.D., Ph.D. Author Info & Affiliations

Published June 14, 2012 | N Engl J Med 2012;366:2240-2242 | DOI: 10.1056/NEJMp1203102

VOL. 366 NO. 24 | Copyright © 2012

"In reality, diverse functionality needn't reside within single EHR systems, and there's a clear path toward better, safer, cheaper, and nimbler tools for managing health care's complex tasks."

21st Century Cures Act

- Merged and replaced previous committees with one that focuses on **interoperability, privacy, and patient empowerment**
- Explicitly prohibited “information blocking”, meaning EHR vendors or health systems that prevent data sharing
- Required development for standardized APIs (like FHIR) to allow patients and third party apps to securely access and exchanged health data

In 2015, lots of open EHR datasets

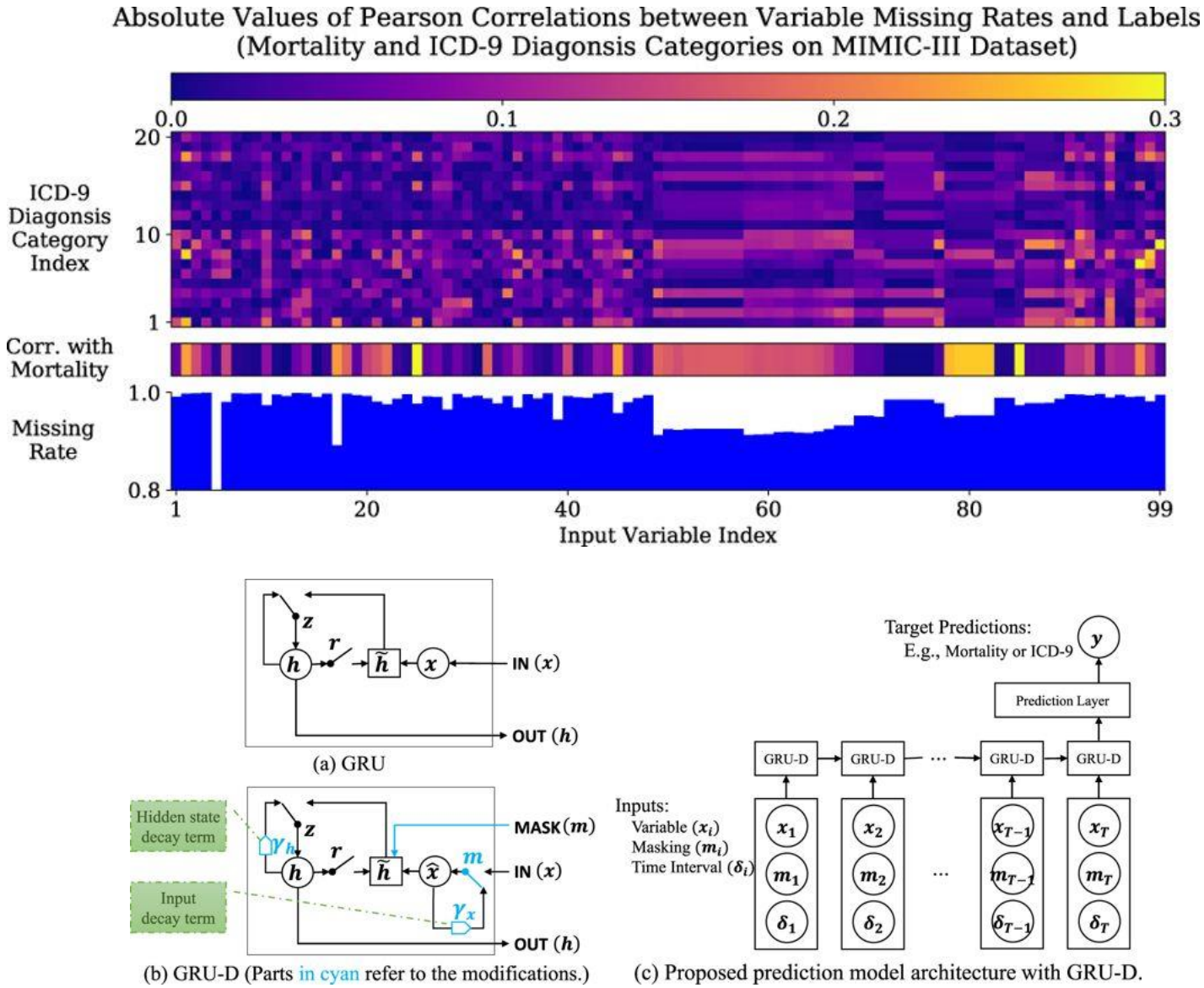
- **MIMIC-IV**: 65k patients intensive care unit (ICU) and emergency department (ED) data from Beth Israel Deaconess Medical Center in Boston
- **All of Us** research program: 200k patients including EHR, genomics, wearables, and patient surveys in standardized (OMOP) format from NIH
- **EHRSHOT**: 6k patients and 900k visits from Stanford

What is MIMIC?

- Largest open dataset for clinical healthcare (for authorized researchers)
- Dataset of 26 tables (e.g. admissions, patients)
- Maintained by Roger Mark's lab at MIT
- Potential place for **final project ideas**!

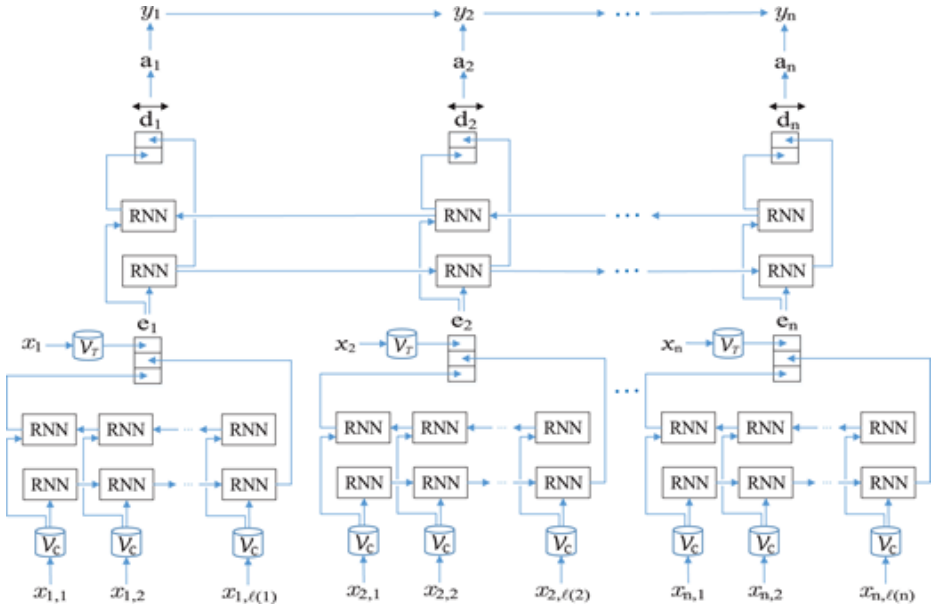
Recurrent Neural Networks for Multivariate Time Series with Missing Values

Che et al, 2018
(Nature Scientific Reports)



De-identification of patient notes with recurrent neural network

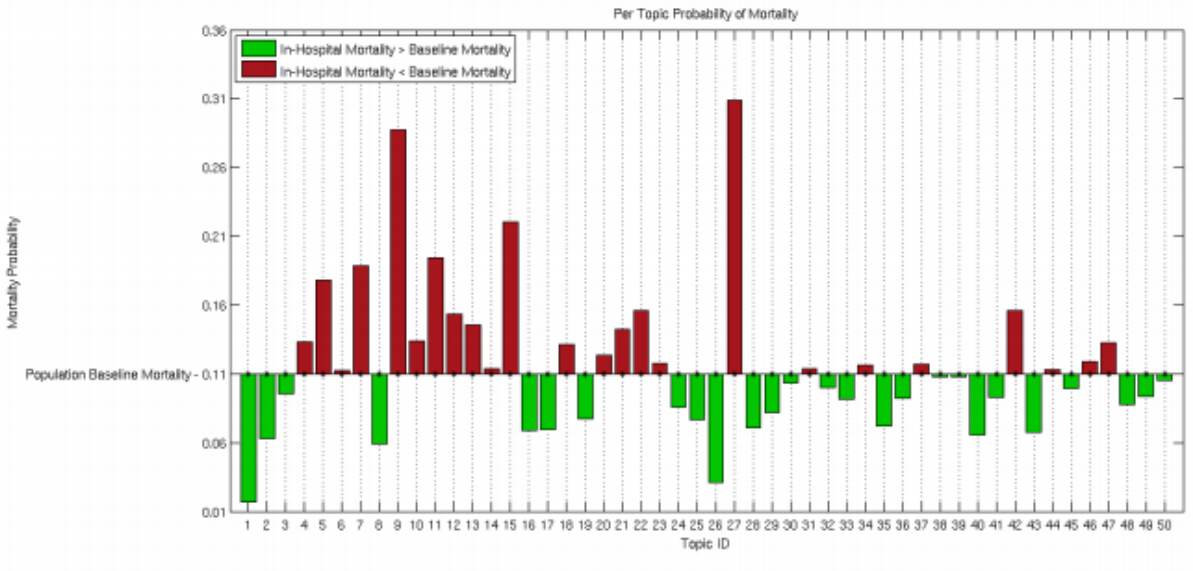
Dernoncourt et al, 2017 (JAMIA)



	i2b2			MIMIC		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Nottingham	**99.000**	96.400	97.680	–	–	–
MIST	91.445	92.745	92.090	95.867	98.346	97.091
CRF	98.560	96.528	97.533	99.060	98.987	99.023
ANN	98.320	97.380	97.848	**99.208**	99.251	**99.229**
CRF + ANN	97.920	**97.835**	**97.877**	98.820	**99.398**	99.108

Unfolding Physiological State: Mortality Modeling in Intensive Care Units

Ghassemi et al, 2014 (KDD)

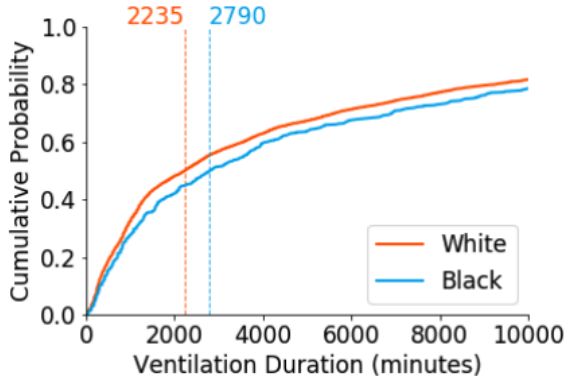


	Topic	Top Ten Words	Possible Topic
In-hospital Mortality	27	name, family, neuro, care, noted, status, plan, stitle, dr, remains	Discussion of end-of-life care
	15	intubated, vent, ett, secretions, propofol, abg, respiratory, resp, care, sedated	Respiratory failure
	7	thick, secretions, vent, trach, resp, tf, tube, coarse, cont, suctioned	Respiratory infection
	5	liver, renal, hepatic, ascites, dialysis, failure, flow, transplant, portal, ultrasound	Renal Failure

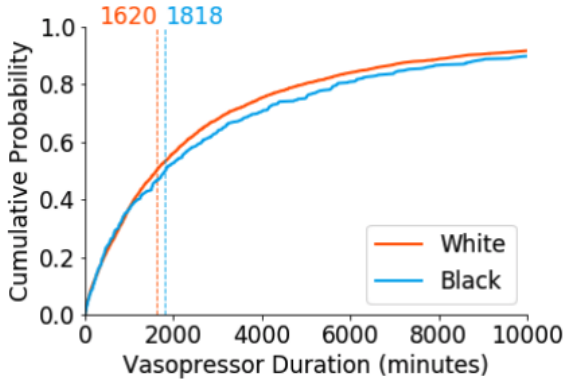
Racial Disparities and Mistrust in End-of-Life Care

Boag et al, 2018 (MLHC)

Social: Pt refused to sign ICU consent and expressed wishes to be DNR/DNI, seemingly very frustrated and mistrusting of healthcare system in relation to [redacted]. Also, w/ hx of poor medication compliance and follow-up



(a) CDF of ventilation duration by race ($p = .005$).



(b) CDF of vasopressor duration by race ($p = 0.12$).

Reproducibility in critical care: a mortality prediction case study

Johnson et al, 2017 (MLHC)

We reproduced datasets for 38 experiments corresponding to 28 published studies using MIMIC. In half of the experiments, the sample size we acquired was 25% greater or smaller than the sample size reported. The highest discrepancy was 11,767 patients. While accurate reproduction of each study cannot be guaranteed, we believe that these results highlight the need for more consistent reporting of model design and methodology to allow performance improvements to be compared. We discuss the challenges in reproducing the cohorts used in the studies, highlighting the importance of clearly reported methods (e.g. data cleansing, variable selection, cohort selection) and the need for open code and publicly available benchmarks.

Study	Window, W (hours)	Inclusion criteria
Caballero Barajas and Akella (2015)	24	Age>18, Random fixed size subsample
Caballero Barajas and Akella (2015)	48	Age>18, Random fixed size subsample
Caballero Barajas and Akella (2015)	72	Age>18, Random fixed size subsample
Calvert et al. (2016b)	5*	Age>18, In MICU, >1 obs. for all features, LOS $\geq$ 17hr, ICD-9 codes indicating alcohol withdrawal
Calvert et al. (2016a)	5*	Age>18, In MICU, >1 obs. for all features, 500hr $\geq$ LOS $\geq$ 17hr
Celi et al. (2012)	72	ICD-9 code 584.9
Celi et al. (2012)	24	ICD-9 code 430 or 852
Che et al. (2016) (b)	48	PhysioNet 2012 Challenge dataset
Ding et al. (2016)	48	PhysioNet 2012 Challenge dataset
Ghassemi et al. (2014)	12	Age>18, >100 words across all notes
Ghassemi et al. (2014)	24	Age>18, >100 words across all notes
Ghassemi et al. (2015)	24	Age>18, >100 words across all notes, >6 notes

EHR Safari: Data is Contextual

William Boag WBOAG@MIT.EDU
MIT CSAIL
Cambride, MA, USA

Mercy Oladipo MOLADIPO@MIT.EDU
MIT CSAIL
Cambride, MA, USA

Peter Szolovits PSZ@MIT.EDU
MIT CSAIL
Cambride, MA, USA

"we demonstrate numerous examples where a non-clinician's intuition may lead to incorrect – and potentially harmful – modeling assumptions"

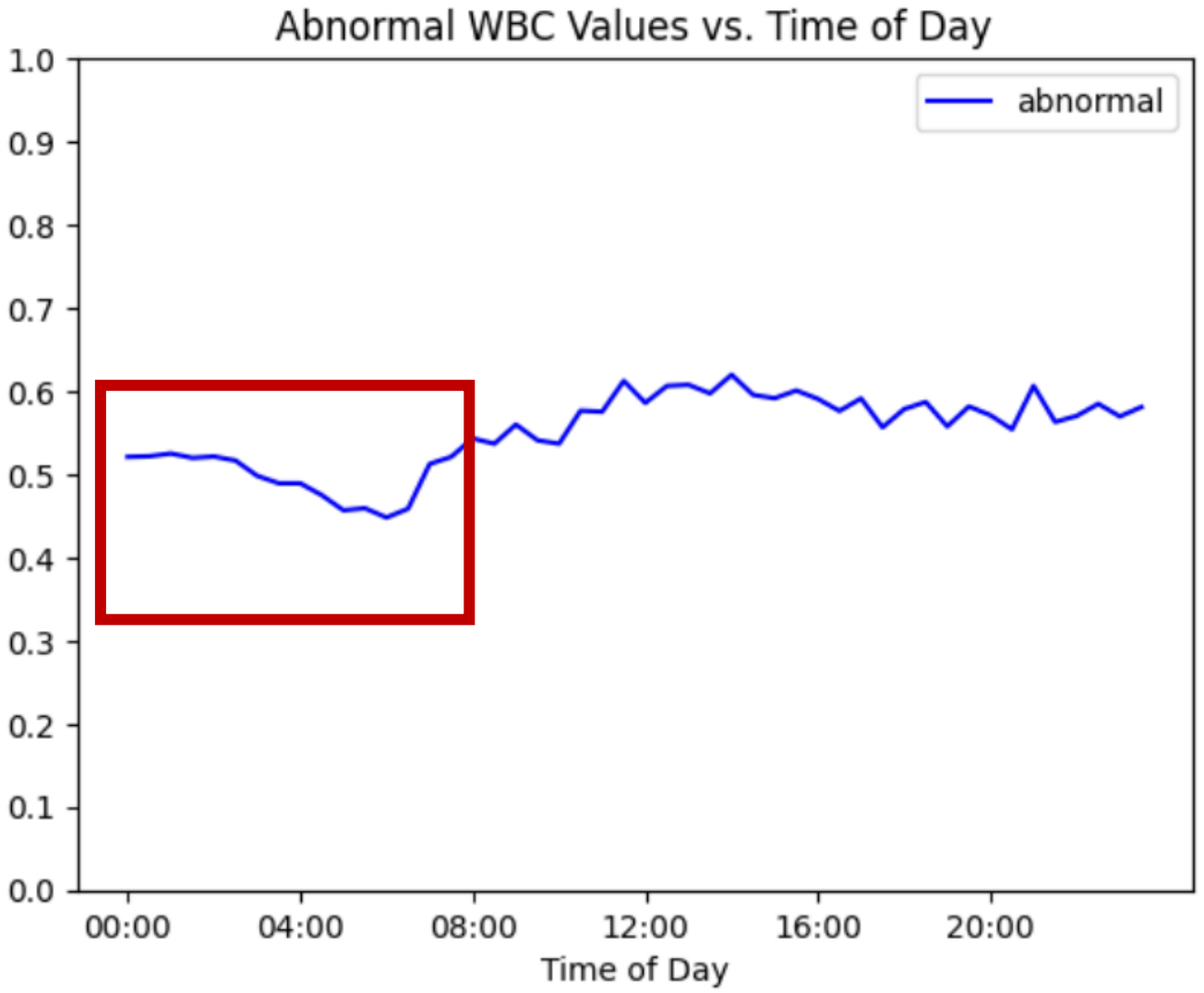
Problems with MIMIC

1. Inconsistent timestamps
2. Lab values vary by time of day
3. Multiple copies of provider notes
4. Missing death date collection

Inconsistent time stamps

- We want time stamps that show:
  - hospital admission time (admittime) ≤
  - patient enters ICU (intime) ≤
  - patient leaves ICU (outtime) ≤
  - patient discharged from hospital (dischtime)
- BUT, of 57k ICU stays, this ordering only holds **79% of the time**
  - Because hospital and ICU staff have their own paperwork, things can get filed out of sync

Lab values vary by time of day

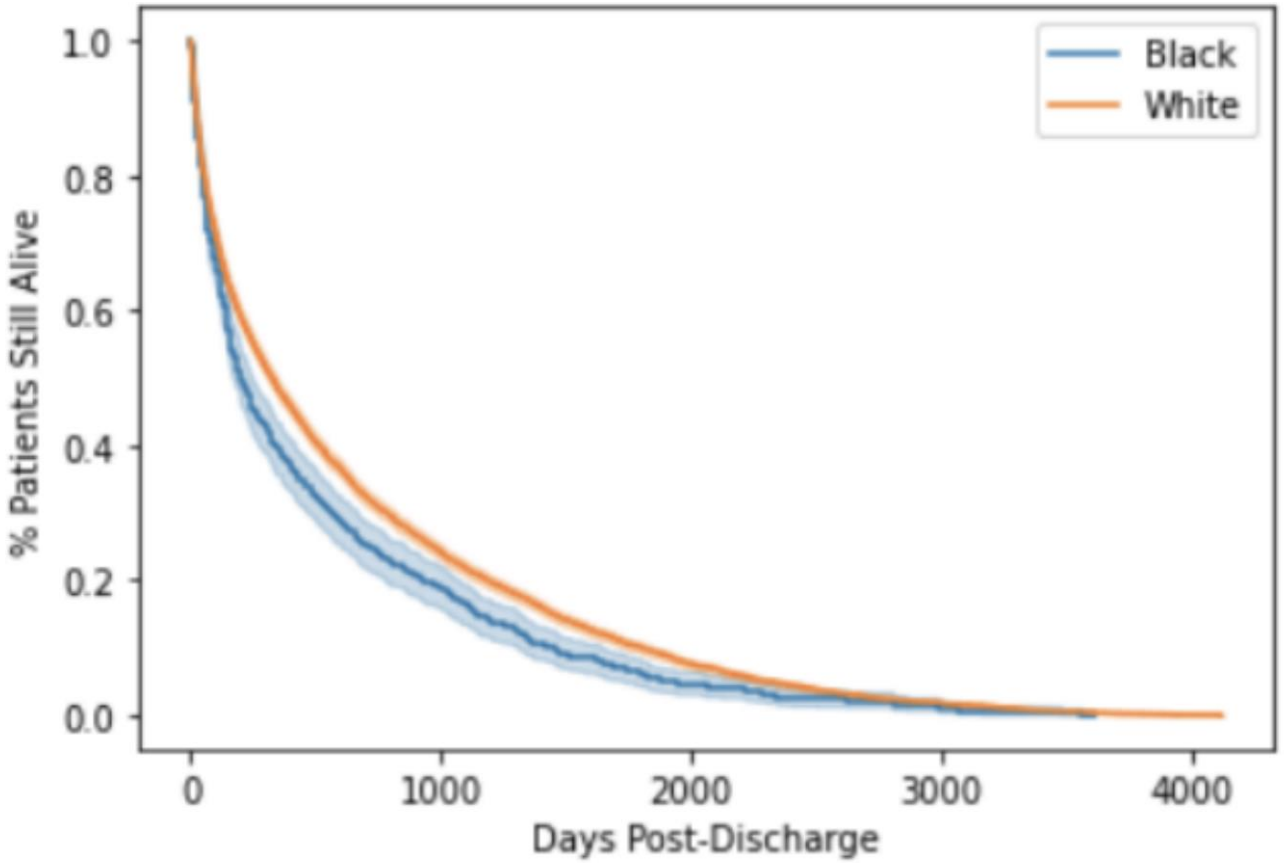


Multiple copies of provider notes

- NLP techniques use clinical notes to predict clinical outcomes
- MIMIC contains all note drafts that EHR autosaves
  - Disconnect between the chart time (completed) and the store time (partially written)

Missing death date collection

- Higher rate of mortality for white patients than black patients??
- Because 2013 legislative change forbid SSN collection from state records, this led to 40% change in capture of death information



More reading

- Caldwell, "We've Spent Billions to Fix Our Medical Records and They're Still a Mess", November 2015

- Boag et al, "EHR Safari: Data is Contextual", MLHC 2022.



Mother Jones

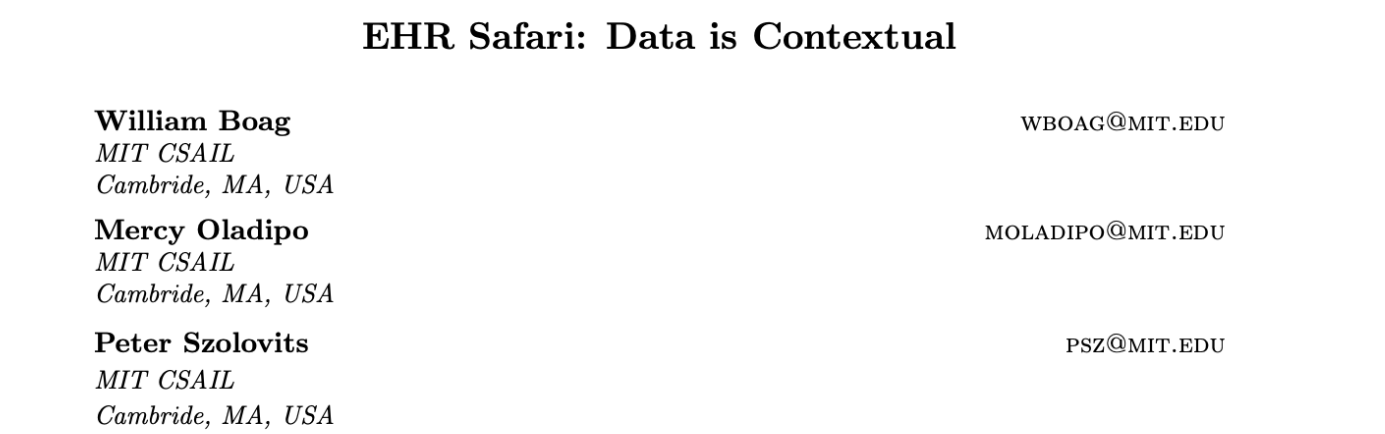
Smart, Fearless Journalism Donate

POLITICS ENVIRONMENT CRIMINAL JUSTICE GUNS RACE GENDER FOOD PODCAST VIDEO MAGAZINE

We've Spent Billions to Fix Our Medical Records, and They're Still a Mess. Here's Why.

Digitizing America's medical records was supposed to help patients and save money. Why hasn't that happened?

PATRICK CALDWELL NOVEMBER/DECEMBER 2015 ISSUE



EHR Safari: Data is Contextual

William Boag WBOAG@MIT.EDU
MIT CSAIL
Cambride, MA, USA

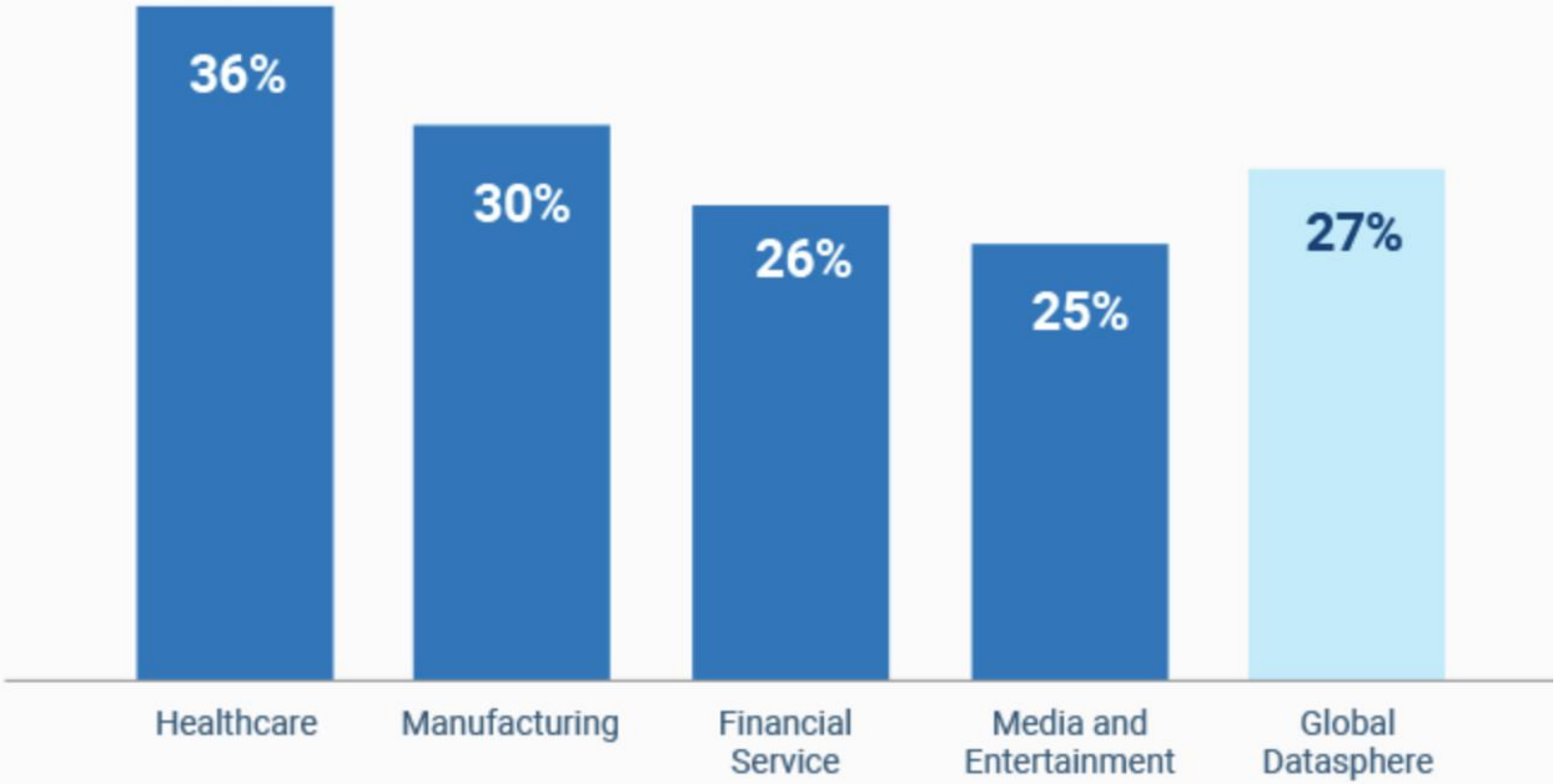
Mercy Oladipo MOLADIPO@MIT.EDU
MIT CSAIL
Cambride, MA, USA

Peter Szolovits PSZ@MIT.EDU
MIT CSAIL
Cambride, MA, USA

30% of world’s data volume is health data

Source: https://www.rbccm.com/en/gib/healthcare/episode/the_healthcare_data_explosion

2018-2025 Data – Compound Annual Growth Rate (CAGR)



Source: Coughlin et al Internal Medicine Journal article "Looking to tomorrow's healthcare today: a participatory health perspective". IDC White Paper, Doc# US44413318, November 2018: The Digitization of the World – From Edge to Core".

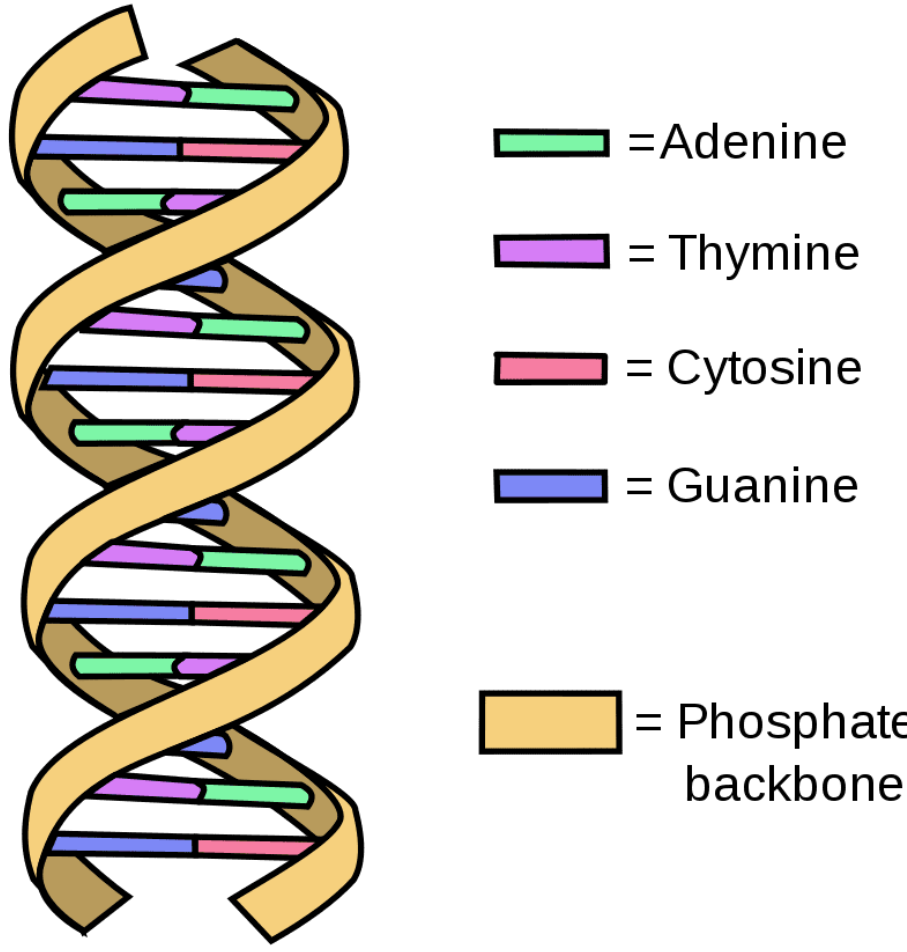
What is the goal of health data?

What is the goal of using health data?

1. **Clinical outcomes**: Given a label (e.g., diagnosis), predict patients most at risk
2. **Patient trajectories**: Given the beginning of a disease trajectory, predict future events over time
3. **Disease subtypes**: Unsupervised learning to determine heterogeneity in patient population
4. **Population monitoring**: Identify emergent public health concerns and where population-level interventions would be helpful

Genomic data

- Organism's complete set of DNA, including sequence of genes, functions, how they're regulated
- Usually includes:
  - Germline genotypes (SNPs)
  - Whole genome sequences
  - Polygenic risk scores (PRS)
- Example paper: Kooperberg etl a, "Risk Prediction using Genome-Wide Association Studies", *Genetic Epidemiology* 2011.



DNA

If you had EHR + genotype data, what would you try to predict? What would be the baseline you compare it to?

Risk Prediction using Genome-Wide Association Studies

- Showed effectiveness of using genetic markers in SNPs through Genome-Wide Association Studies (GWAS)
- Used **sparse regression methods**, i.e., lasso and elastic net regression, because of extremely high-dimensional data
- Wanted to individual disease risk for Crohn's disease, type 1 diabetes, and type 2 diabetes
- They found that using hundreds of SNPs improved prediction model

http://pmc.ncbi.nlm.nih.gov/articles/PMC2964405/

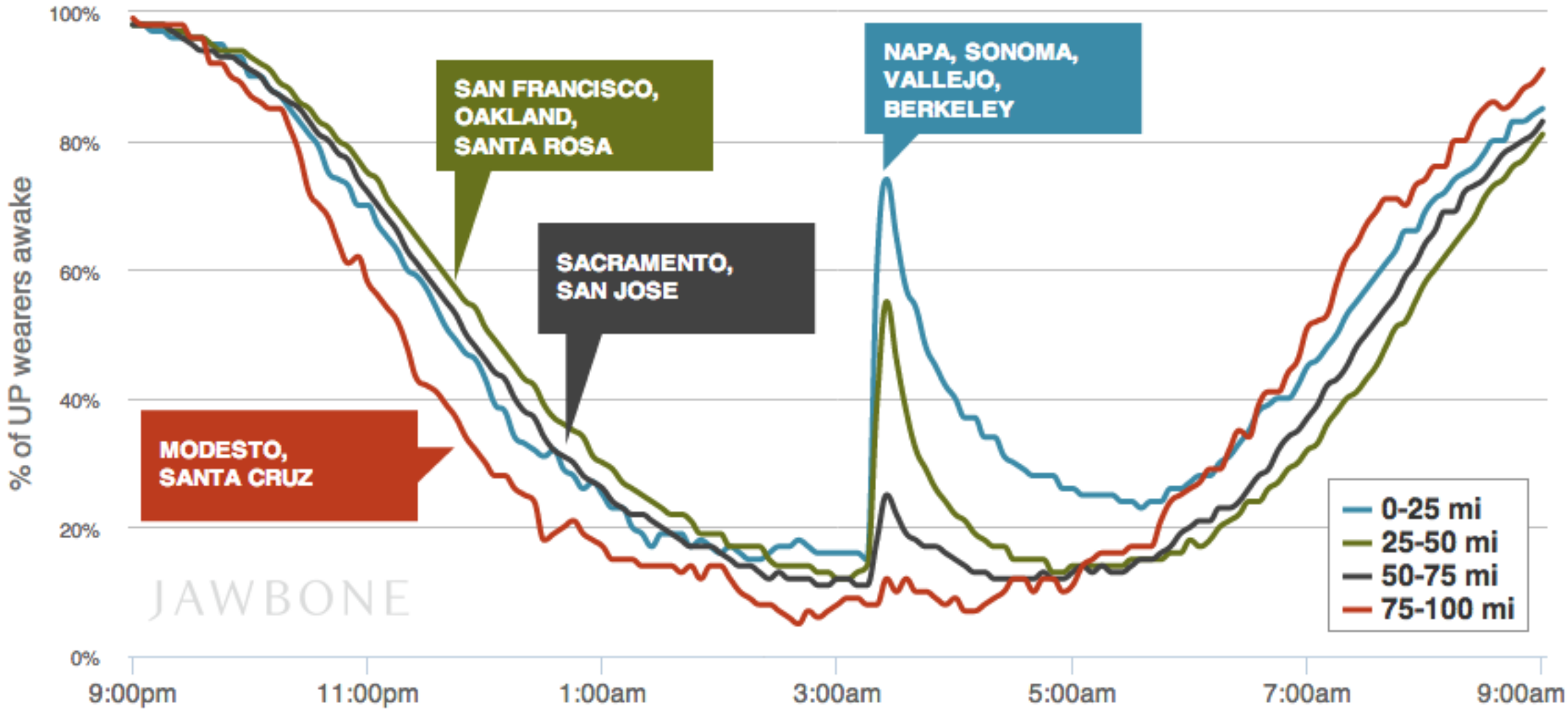
Genomic data

- Pros:
  - High dimensional data
  - Lifelong prediction
  - Could enable precision medicine
  - Growing number of datasets
- Cons:
  - Effect sizes for many common diseases are small
  - Many studies focus only on people of European descent (78%)
  - Data privacy issues

Martin et al, “Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations”

Wearables

- Data collected from sensors that people wear (e.g., wrist bands, smart socks, chest patches)
- Variables include **step count, heart rate, sleep duration**

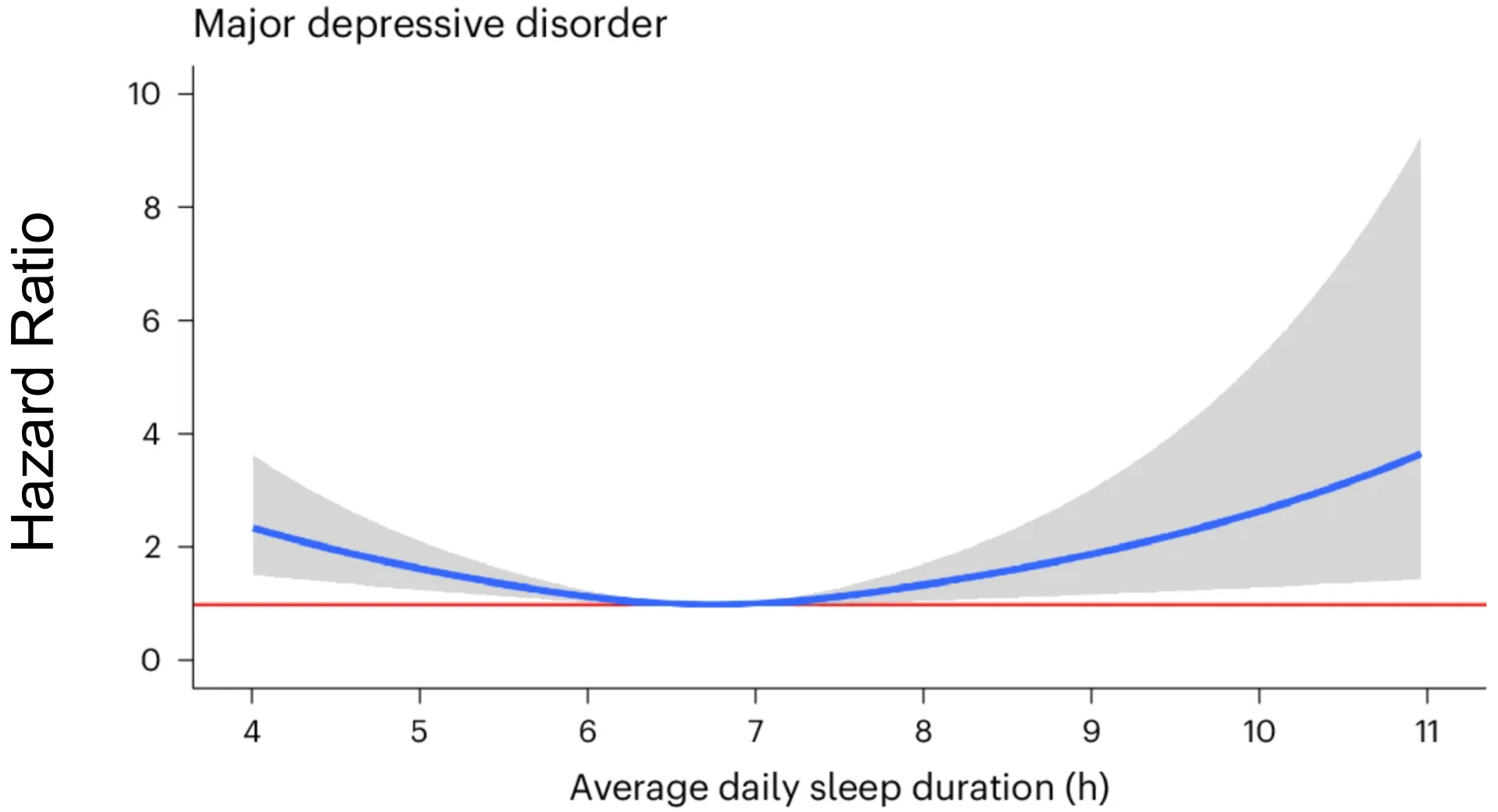

100%
80%
60%
40%
20%
0%
% of UP wearers awake
SAN FRANCISCO, OAKLAND, SANTA ROSA
SACRAMENTO, SAN JOSE
NAPA, SONOMA, VALLEJO, BERKELEY
MODESTO, SANTA CRUZ
JAWBONE
0-25 mi
25-50 mi
50-75 mi
75-100 mi
9:00pm
11:00pm
1:00am
3:00am
5:00am
7:00am
9:00am

Example papers:

- Quer et al, “Wearable sensor data and self-reported symptoms for COVID-19 detection”, *Nature Medicine* 2020.
  - AUC of 0.80
  - Combination of wearable sensor and app-solicited symptoms (e.g., body aches)

- Zheng et al, “Sleep patterns and risk of chronic disease as measured by long-term monitoring with commercial wearable devices in the All of Us Research Program”, *Nature Medicine* 2024.

N=6,785 participants over 4.5 years in All of Us datasets



What kind of issues might come up for using wearable data?

Wearables

- Pros:
  - Continuous measurements outside of the healthcare system
  - Low friction to gather longitudinal data
  - Opportunity for early detection
- Cons:
  - Data quality or missingness issues (e.g., people not wearing)
  - Equity issues: wearable users skew certain socioeconomic groups
  - Interoperability concerns
  - Ground-truth labeling

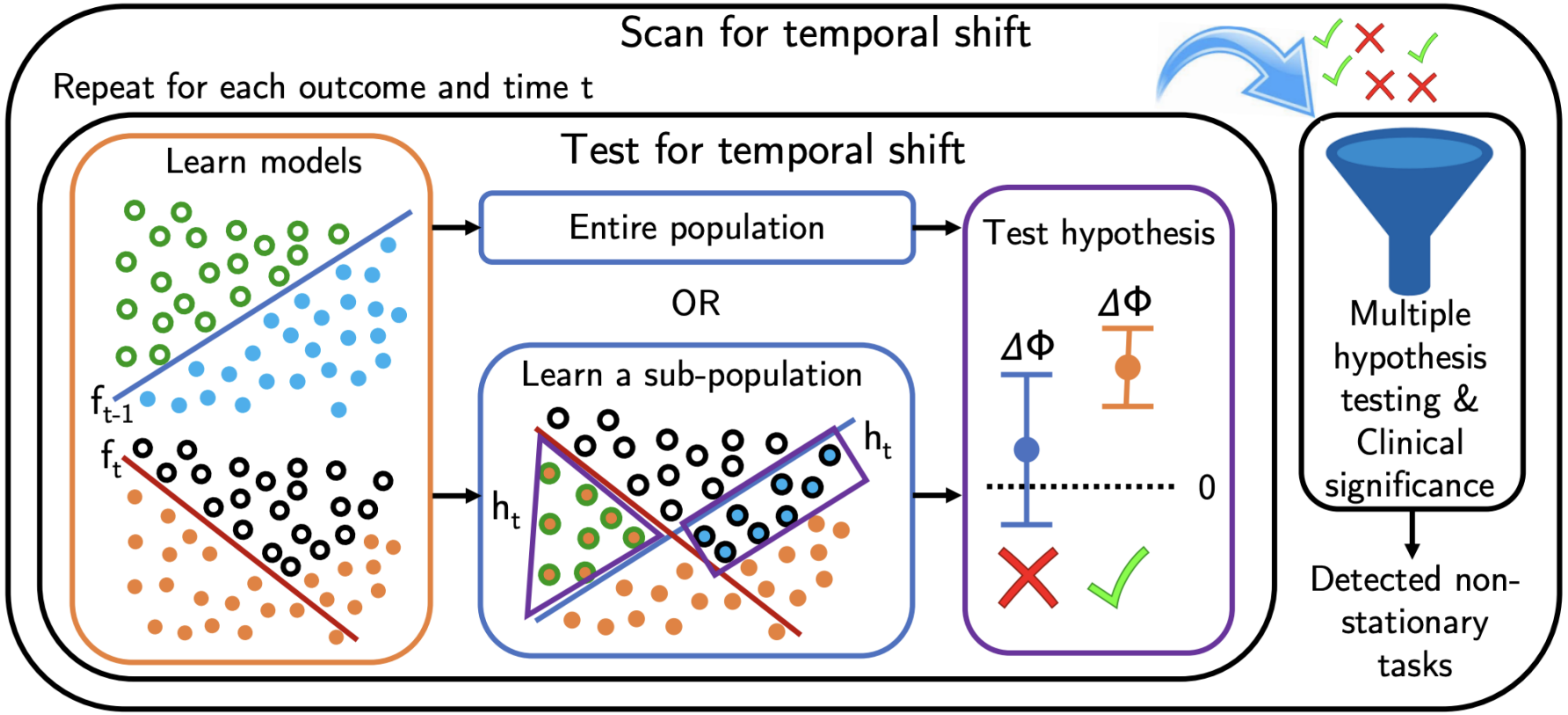
Insurance claims

- Claims data are administrative data generated for billing purposes (e.g., Medicare/Medicaid, commercial insurance companies)
- Similar to EHR data, contains diagnoses, procedures, prescriptions – and sometimes cost data
- Example paper: Ji et al, “Large-Scale Study of Temporal Shift in Health Insurance Claims”, CHIL 2023

Insurance claims

- Pros:
  - Large scale, often larger than single hospital, usually nationwide
  - Longitudinal claims across trajectories of care
  - Well-structured
- Cons:
  - Used for billing, not research, so diagnoses may be changed to justify reimbursement
  - No clinical details: lab values, vital signs, imaging, severity of disease
  - Censoring: patients might change insurers
  - Confounding: claims reflect treated populations, omitting uninsured

Large-Scale Study of Temporal Shift in Health Insurance Claims



Large-Scale Study of Temporal Shift in Health Insurance Claims

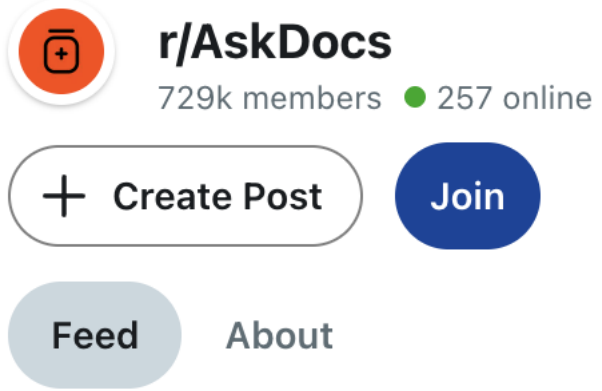
- Large private health insurance claims dataset (**1.6 million** patients, **15k features**) from 2015 to 2020
- Defined 1010 prediction tasks across 242 health outcomes, predicting if outcome would occur in 3 months
- Interested in determining if temporal shift would occur
  - **9.7% of tasks** had a shift for entire population
  - **93.0% of tasks** had a shift for specific groups
- Majority of shifts (62 out of 98) happened in 2020 during the pandemic

How would you account for temporal changes (coding, policy, population risk)?

Social Media Data

- Data from platforms where users generate content (e.g., Twitter, Reddit, Facebook, Youtube)

- Example paper: Eichstaedt et al, “Facebook language predicts depression in medical records”, PNAS 2018



Facebook language predicts depression in medical records

- Used history of **Facebook status updates** from 638 consenting patients
  - 114 had documented degression diagnosis
  - 524 did not
- Analyzed over **524k Facebook updates**
- Could predict depressed patients with AUC of 0.69, 3 months before first documentation of diagnosis
- Key predictors included phrases reflecting sadness, loneliness, and increased **use of first-person pronouns**

What biases might arise from using social media data to understand health?

Social Media Data

- Pros:
  - No standardization, authentic “patient voice”
  - Captures people who are not well-served by healthcare system
  - Potential large volume, potentially real-time
- Cons:
  - Selection bias: social media users aren’t representative of population
  - Noise and confounding: many posts are ambiguous
  - Ethical/privacy issues: consent, de-identification, platform policy
  - Hard to get ground truth

Other health sources

- Voice/speech data
- Facial and video data
- Environmental and geospatial data
- Mobility and transportation data
- LLM chat logs

Discussion

- What other data sources could you use for health data?
- What is the tradeoff between richness and reliability (e.g., genomics and wearables data)?
- Who is missing from each data type? How can we measure those populations better?
- What is the difference between understanding and predicting health?

Outline

✓**EHR Adoption**
✓**MIMIC Deep Dive**
✓**Other Health Data**



How can we make Data 146 better for you?

Next class: Radiology foundation model evaluation