Data 146: Foundations for CPH

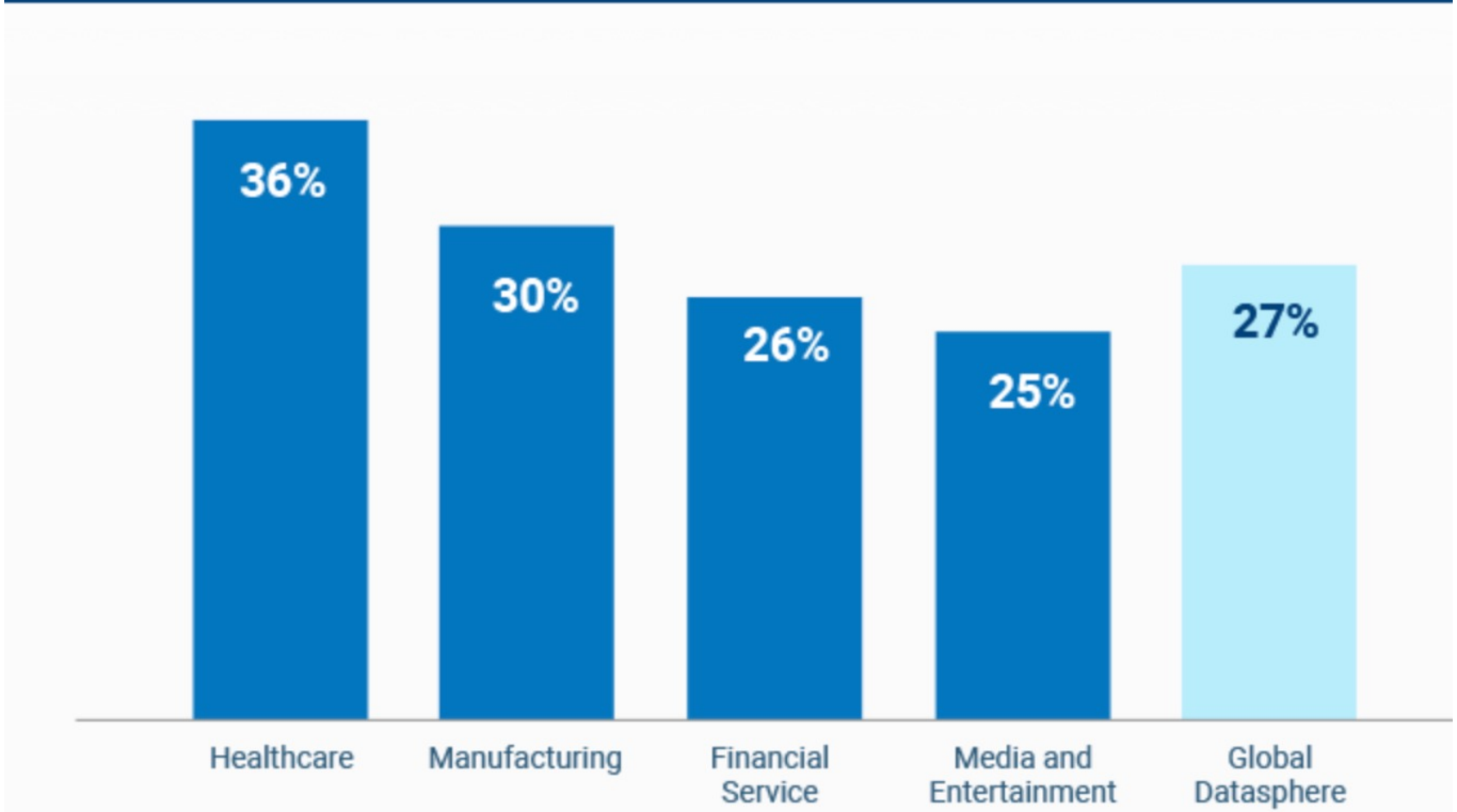
Other Health Data

Irene Y. Chen

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 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

Outline

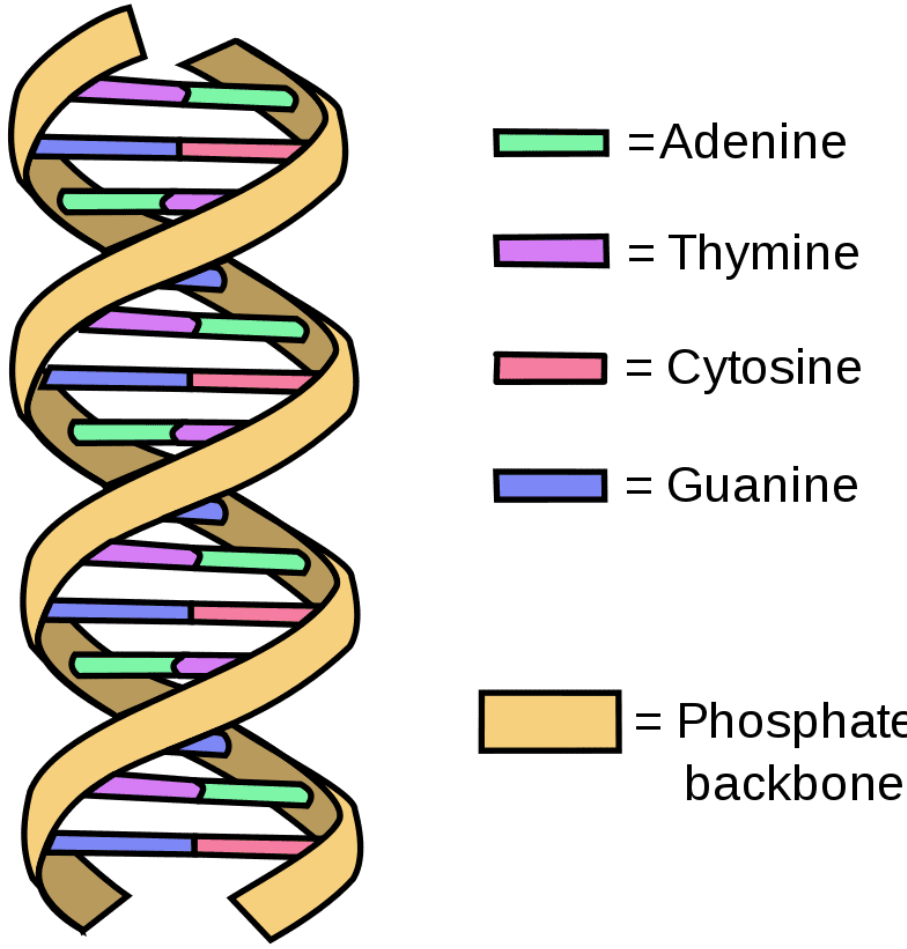
- **Genomic data** (10 mins)
- **Wearables** (10 mins)
- **Insurance claims** (10 mins)
- **Social media data** (10 mins)
- **Discussion** (10 mins)

How can we make Data 146 better for you?

Learning Objective: Understand health data sources and potential challenges

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? (Partner discussion)

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/

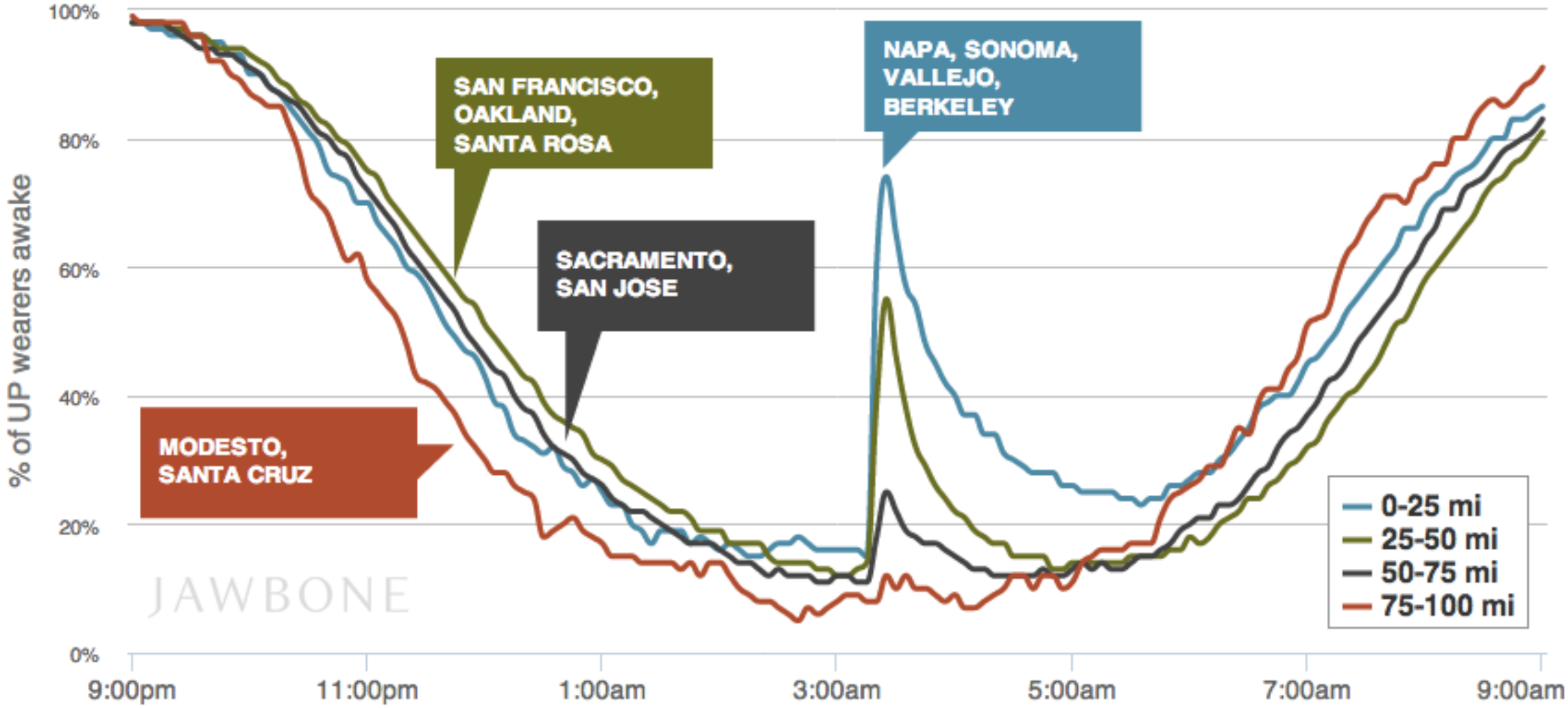
Genomic data

- 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**

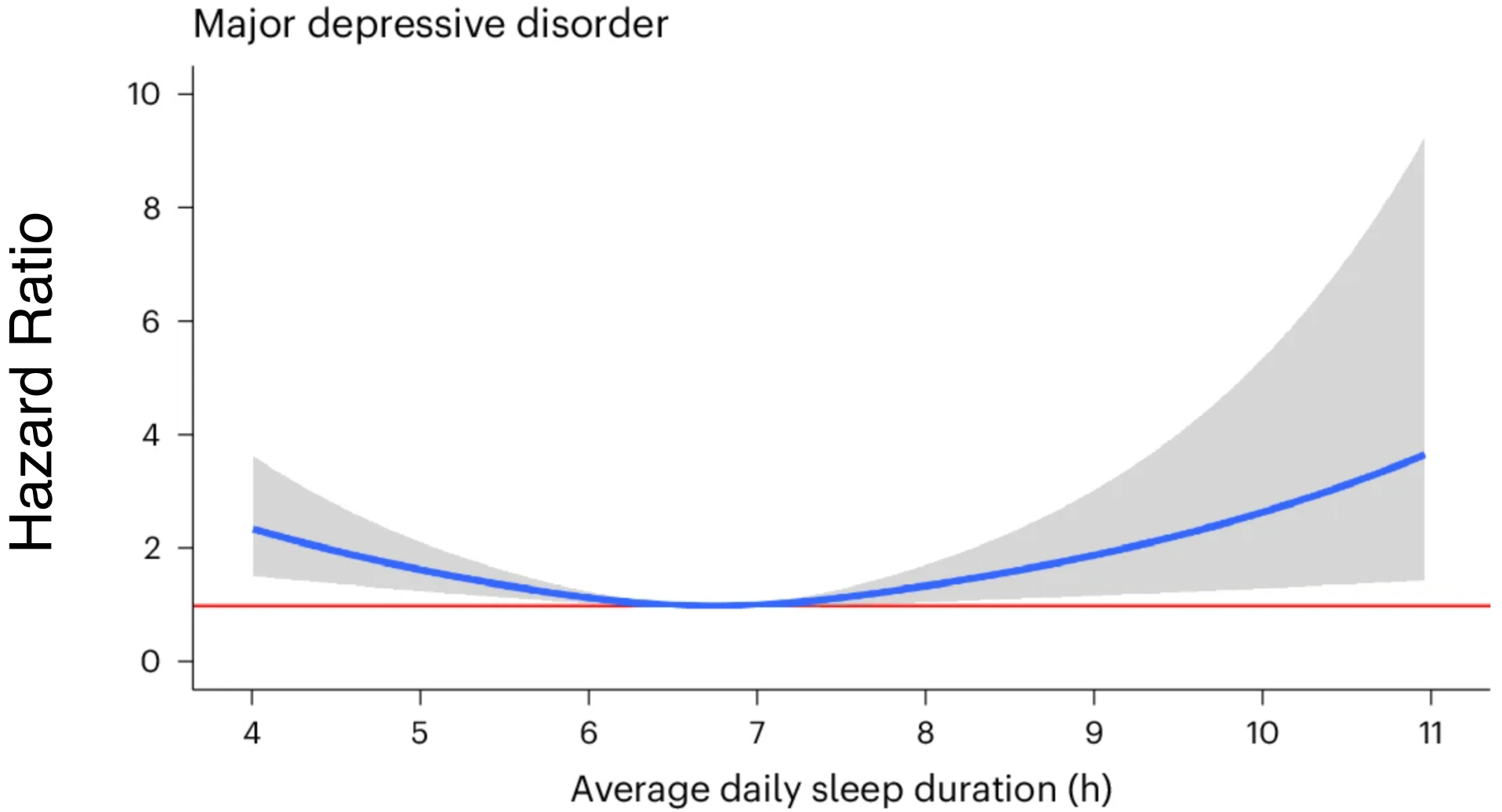

100%
80%
60%
40%
20%
0%
% of UP wearers awake
SAN FRANCISCO, OAKLAND, SANTA ROSA
SACRAMENTO, SAN JOSE
NAPA, SONOMA, VALLEJO, BERKELEY
MODESTO, SANTA CRUZ
JAWBONE
0-25 mi
25-50 mi
50-75 mi
75-100 mi
9:00pm
11:00pm
1:00am
3:00am
5:00am
7:00am
9:00am

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

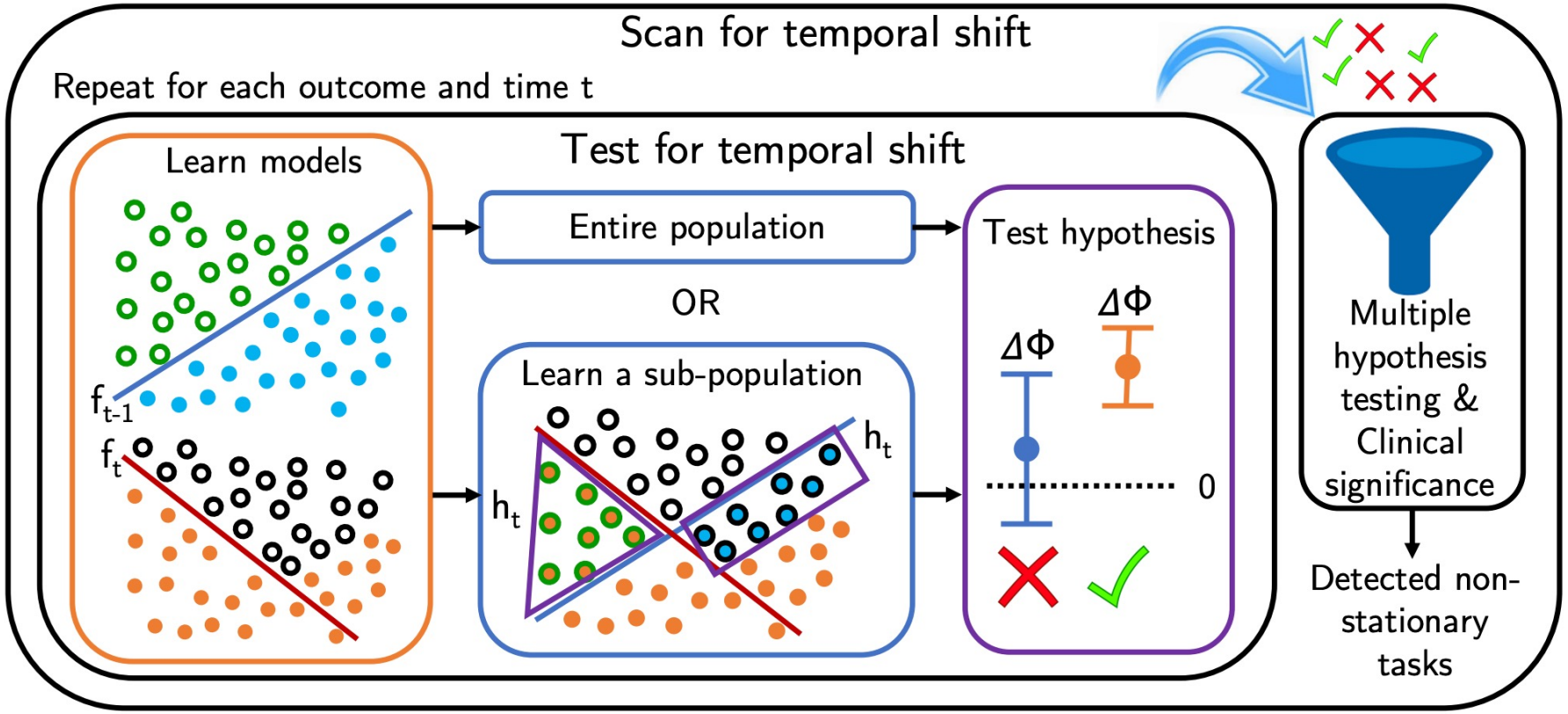
Insurance claims

- 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

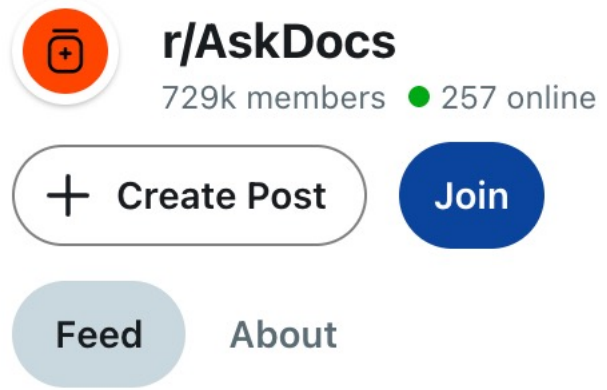
- Large private health insurance claims dataset (**1.6 million** patients, **15k features**) from 2015 to 2020
- Defined 1010 prediction tasks across 242 health outcomes, predicting if outcome would occur in 3 months
- Interested in determining if temporal shift would occur
  - **9.7% of tasks** had 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?

Summary

- ✓**Genomic data** (10 mins)
- ✓**Wearables** (10 mins)
- ✓**Insurance claims** (10 mins)
- ✓**Social media data** (10 mins)
- ✓**Discussion** (10 mins)

How can we make Data 146 better for you?

Next class: Evaluation and benchmarking