Data

Day 2 — The foundation, constraint, and failure mode of health AI

TOM POLLARD · MIT · JUNE 23, 2026

Where we are on the spine

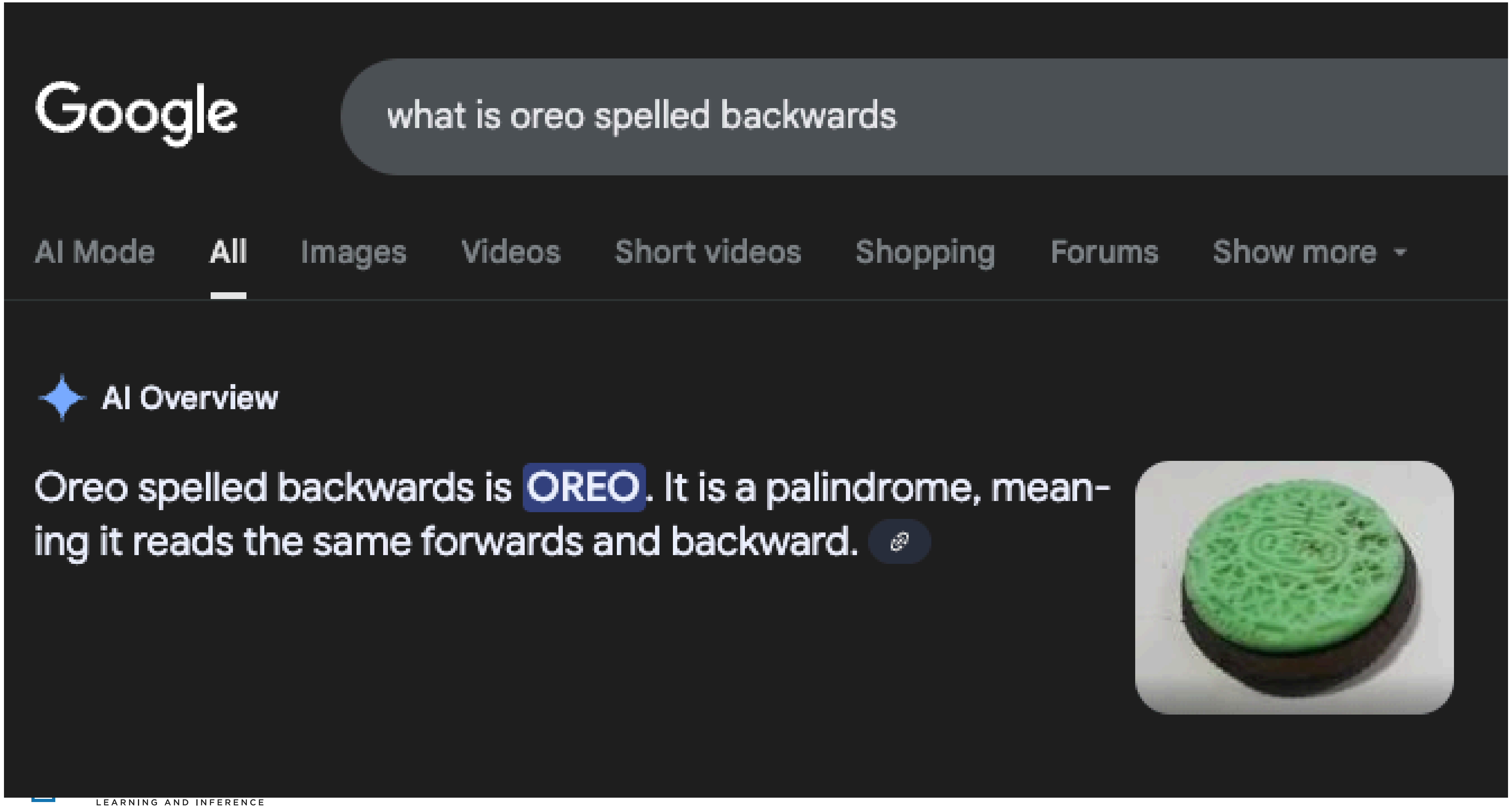
- **Yesterday:** an interesting topic → a well-posed problem.
- **Today:** a problem → the data that *actually exists* to attack it.

Recap problem statements

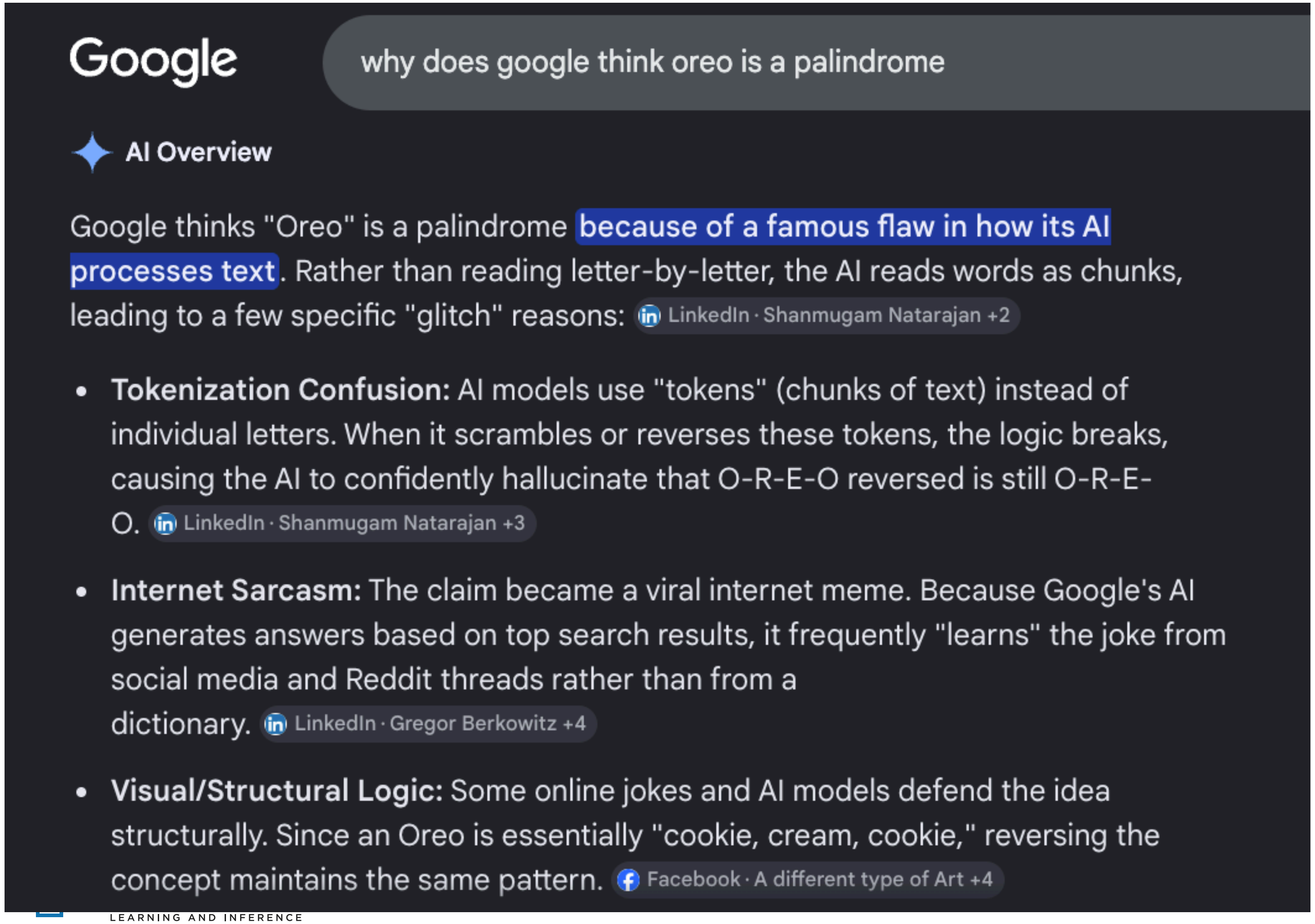
The last decade was dominated by models.

The next will be shaped by data, governance, and deployment.

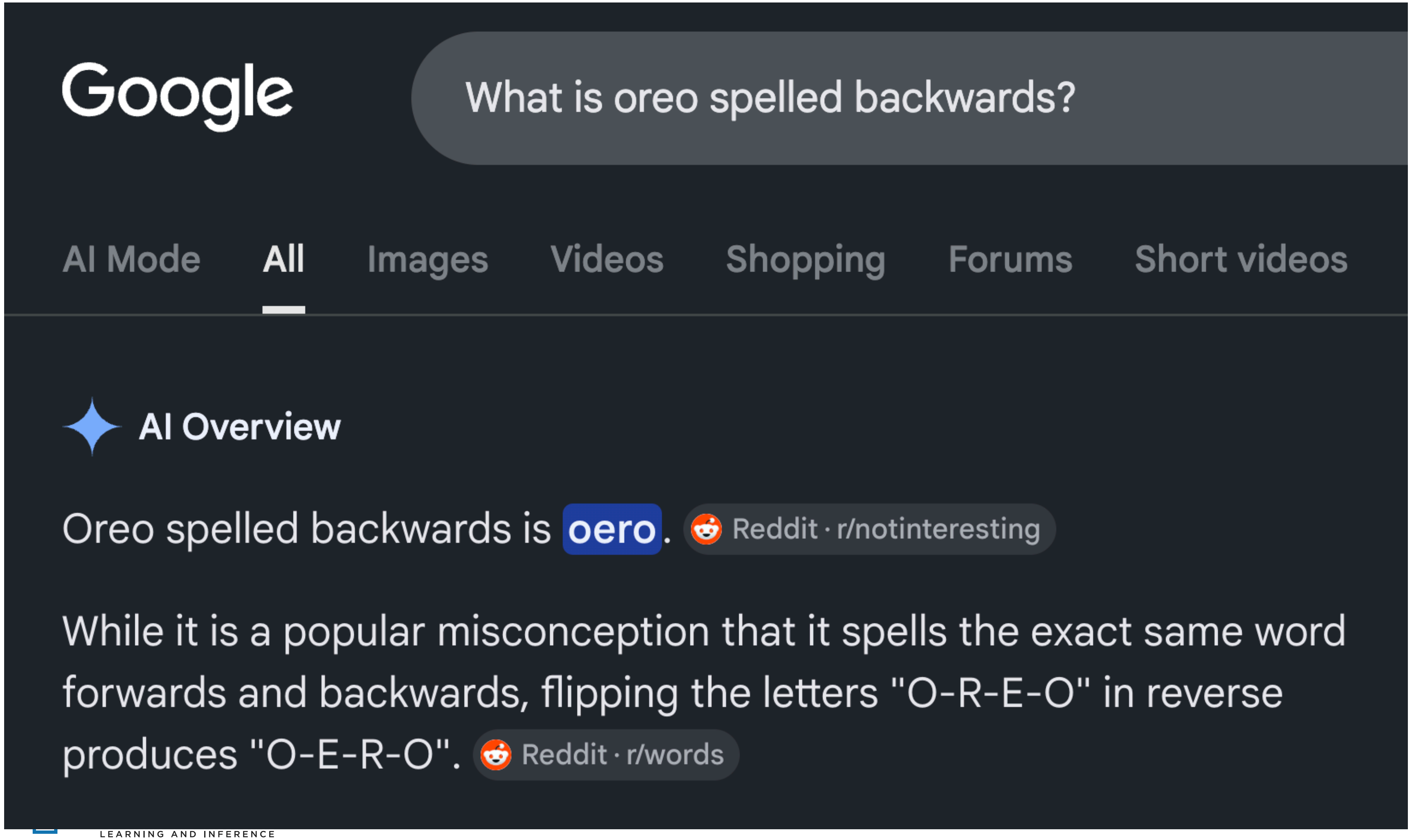
Are we okay with this?



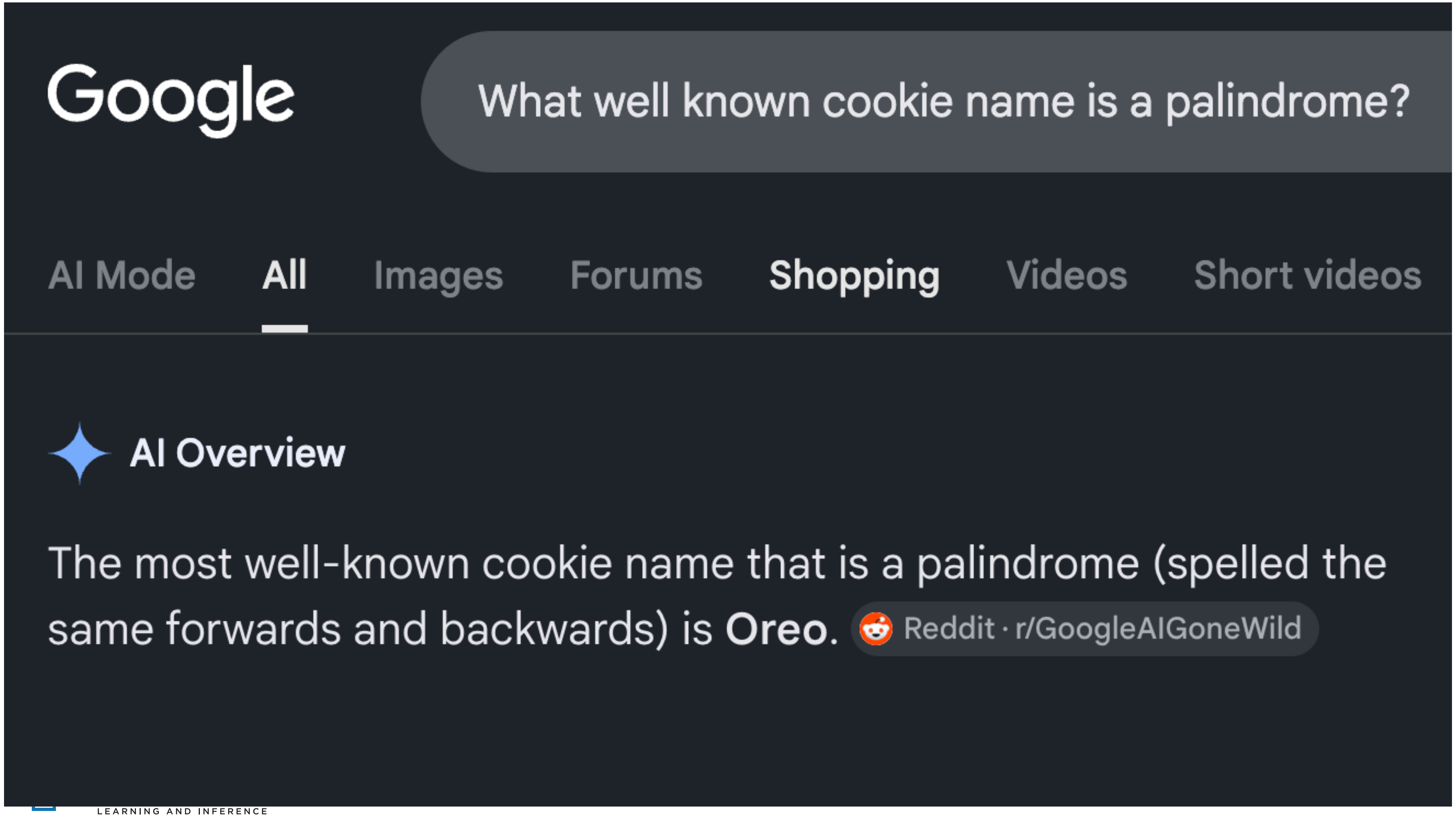
Why does Google think Oreo is a palindrome?



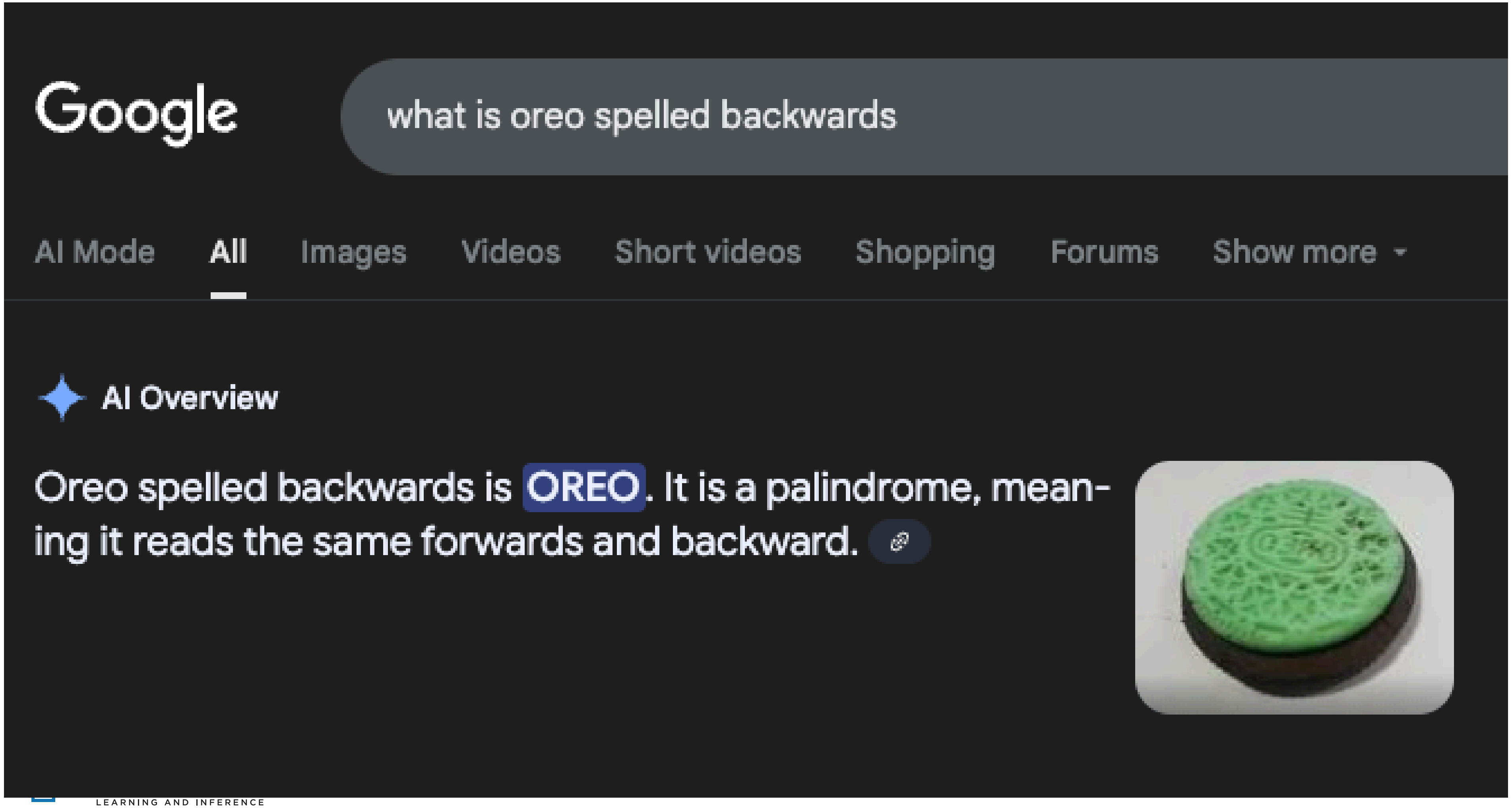
They fixed it!



They fixed it?



Are we okay with this in healthcare?

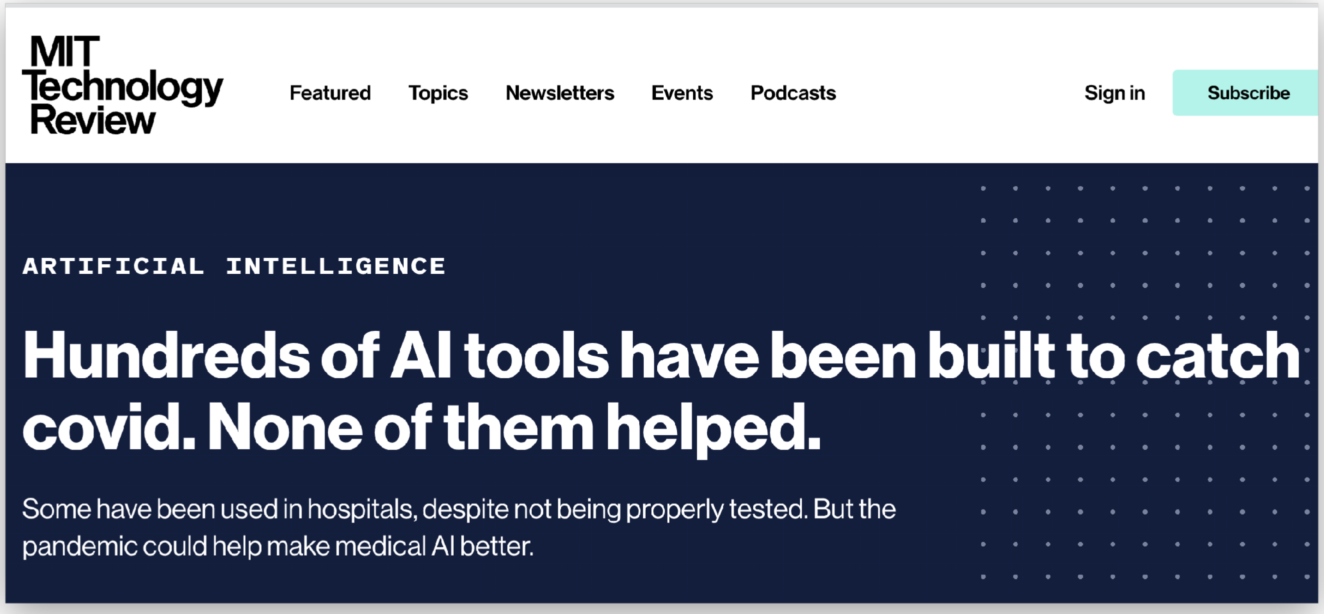


Public health data has enabled reproducible research, benchmarks, education, and community science.

Has it enabled trustworthy, meaningful, and deployable AI for health?

COVID was a stress test

Researchers around the world turned their focus to building predictive models to beat covid.


MIT Technology Review

Featured Topics Newsletters Events Podcasts Sign in Subscribe

ARTIFICIAL INTELLIGENCE

Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

AHLI HEALTH AI SUMMER CAMP 2026

What went wrong?

> "...**datasets spliced together** from multiple sources and [containing] duplicates."

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

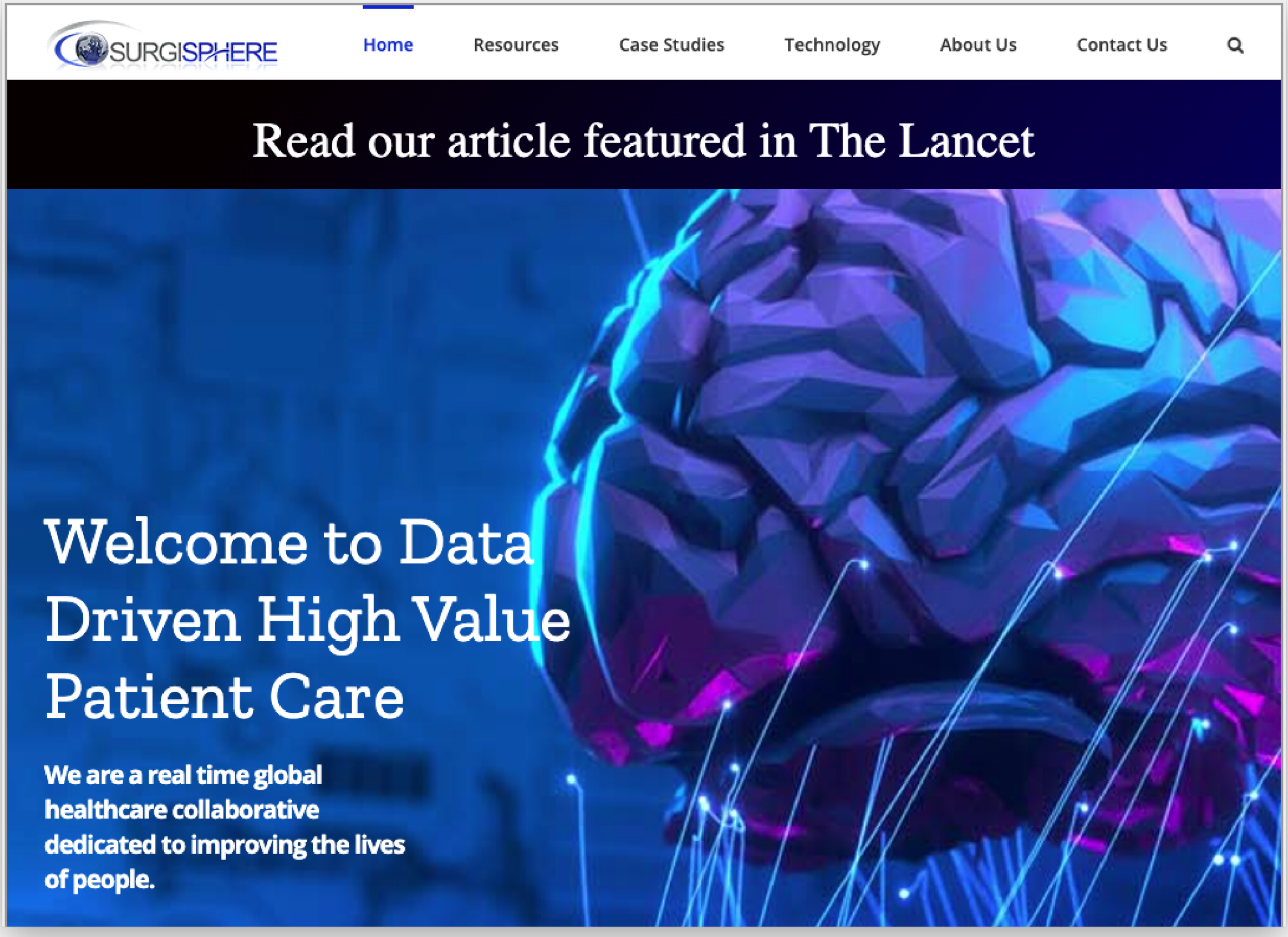
What did the algorithms learn?

> "were tested on the same data they were trained on...learnt to identify duplicate patients
>
> "learnt to use the text font to make predictions".
>
> "learned to predict risk from a person’s position"

Surgisphere

- Led by Sapan Desai, Vascular Surgeon (MD/PhD Chicago)
- 100,000 patients across 671 hospitals
- Papers in NEJM and The Lancet
- Huge global impact

Surgsiphere website from archive.today (May 2020)



AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Provenance of data questioned

- Papers retracted!
- Huge waste of time
- Caused confusion and distrust

The New York Times

New Covid Shot Questions | Symptoms and Treatment | Free Tests Are Back | New Vaccines Are Coming

Two Huge Covid-19 Studies Are Retracted After Scientists Sound Alarms

The reports, published in two leading journals, were retracted after authors could not verify an enormous database of medical records.

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

FINANCIAL TIMES

Pharmaceuticals sector + Add to myFT

WHO says it did not see Surgisphere data that halted virus drug trial

Authors retract Lancet and NEJM studies after company declines to transfer data sets

Home / Archive / October 2020 / Features

The Surgisphere Scandal: What Went Wrong?

The high-profile retractions of two COVID-19 studies stunned the scientific community earlier this year and prompted calls for reviews of how science is conducted, published, and acted upon. The warning signs had been there all along.

This ~~was a turning point~~

This *should have been a turning point*

Home > BMC Medicine > Article

Evidence of unreliable data and poor data provenance in clinical prediction model research and clinical practice

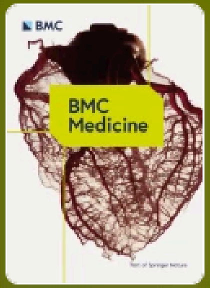
Research | Open access | Published: 04 June 2026

Cite this article

You have full access to this open access article

Download PDF

Save article



BMC Medicine

Aims and scope →

Submit manuscript →

Alexander D Gibson, Nicole M White, Gary S Collins & Adrian G Barnett

Abstract

Background

Clinical prediction models are often created using large routinely collected datasets. It is essential that prediction models are developed with appropriate data and methods and transparently reported to ensure that decisions are based on reliable predictions. Kaggle is a popular competition and data repository website where users learn and apply analysis skills on a range of datasets.

Methods

We identified two large, publicly available Kaggle datasets, on stroke and diabetes, that lack clear data provenance, but are widely used in clinical prediction models in peer reviewed

ASSOCI
LEARNI

These failures were not just “bad data”

They are infrastructural:

- provenance — where did the data come from?
- cohort logic — who was included, and why?
- leakage control — what information was available when?
- evaluation — what artifacts did the model learn?
- governance — who was allowed to use the data, and how?

No one interrogated how the data came to exist and whether it was the right data for the problem.

The data-generating process

Byproduct of care vs. designed for research

Byproduct of care

- Generated while clinicians treat, document, and bill.
- Abundant, messy, biased by workflow.
- Most clinical data lives here.

Designed for research

- Generated by a protocol to answer a defined question.
- Rarer, cleaner *for its purpose* — still biased.
- Benchmarks, registries, trials.

We need to ask:

- what process created the data, and
- what did that process make visible or invisible?

Data is not a neutral input.

It is generated by a process, for a purpose — and that shapes what you can credibly ask.

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Patient timeline

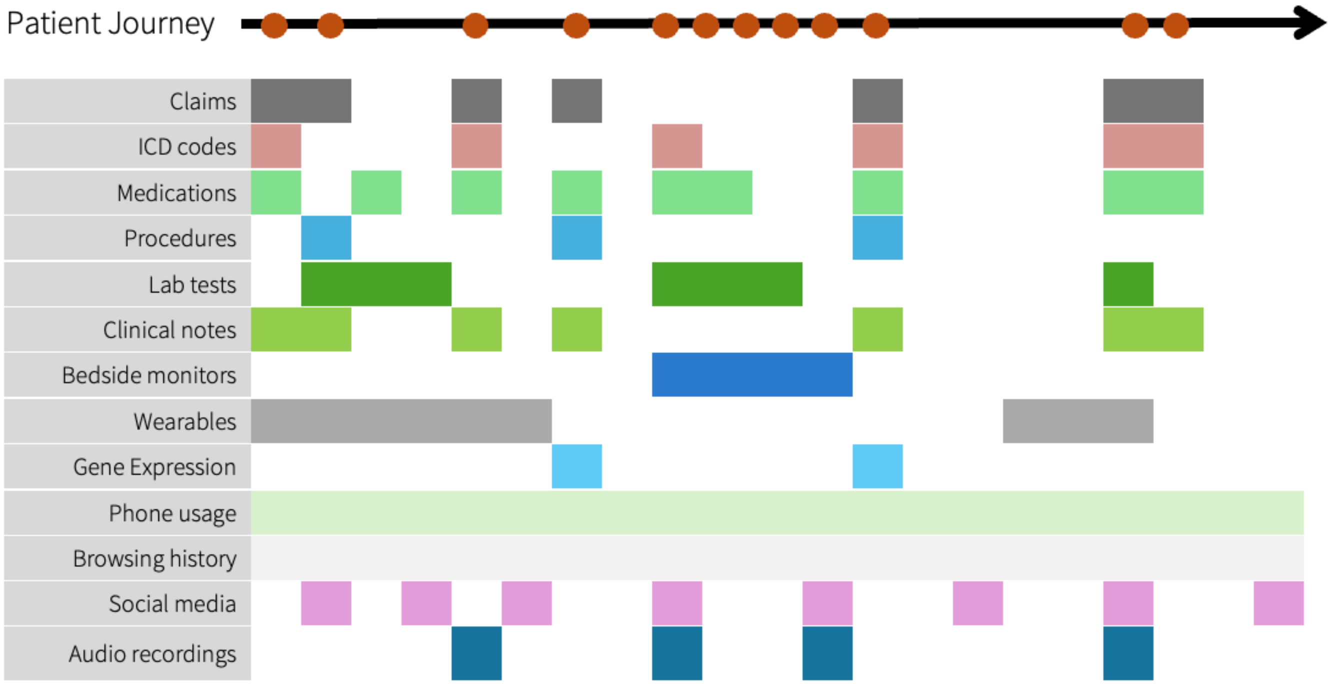
Often we think about isolated sets of data. ICU data; OR data; wearables data.

Helpful to remind ourselves that we are dealing with a sequence of events

Times of health and sickness.

Typically we are viewing just snapshots of the journey.

Credit: Emily Alsentzer.



AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Case study — how a clinical dataset comes to exist

MIMIC as an example:

- Deidentified critical-care data from patients admitted to a Boston Hospital
- **Built** through extraction, structuring, and a substantial de-identification effort.
- A "ready-to-use" dataset represents **significant invisible work** — and the choices in that work shape the data.

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Critical care

Infusions

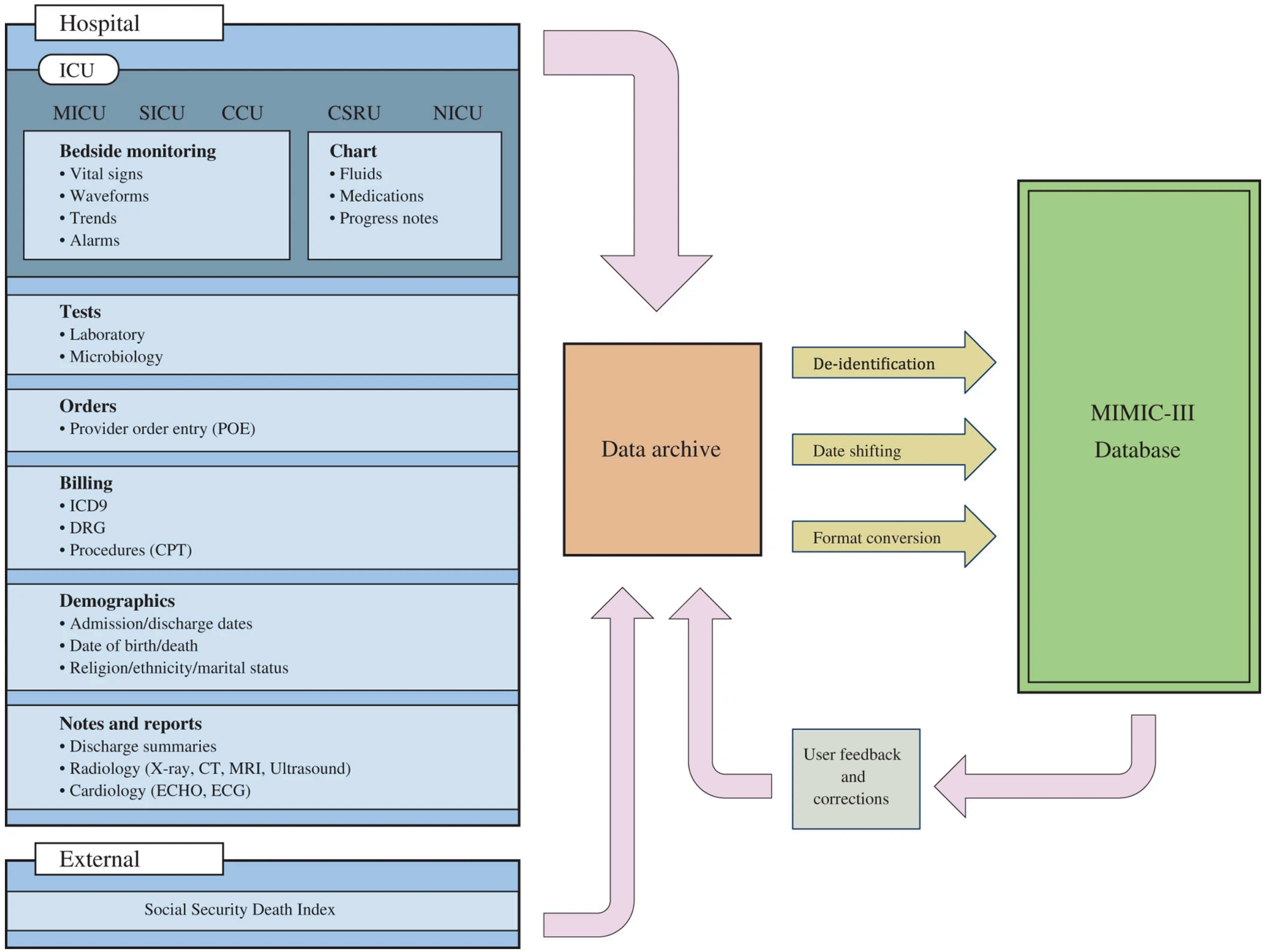
Intermittent treatment

Organ support

Waveforms

Observations, discussion, documentation

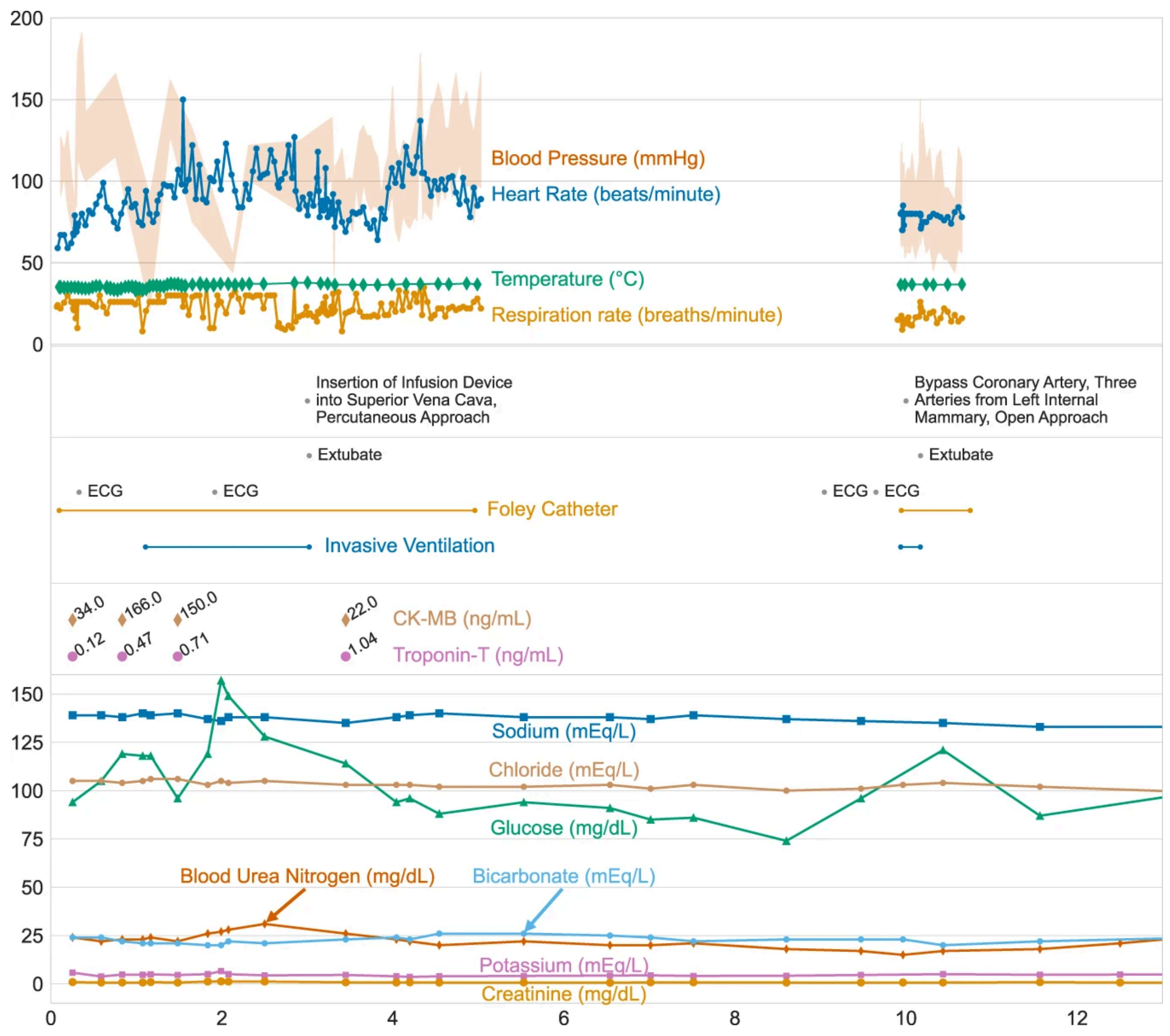
MIMIC


Hospital
ICU
MICU
SICU
CCU
CSRU
NICU
Bedside monitoring
• Vital signs
• Waveforms
• Trends
• Alarms
Chart
• Fluids
• Medications
• Progress notes
Tests
• Laboratory
• Microbiology
Orders
• Provider order entry (POE)
Billing
• ICD9
• DRG
• Procedures (CPT)
Demographics
• Admission/discharge dates
• Date of birth/death
• Religion/ethnicity/marital status
Notes and reports
• Discharge summaries
• Radiology (X-ray, CT, MRI, Ultrasound)
• Cardiology (ECHO, ECG)
External
Social Security Death Index
Data archive
De-identification
Date shifting
Format conversion
MIMIC-III
Database
User feedback
and
corrections

hadm_id=28503629

Vital signs captured only during ICU stay

Target monitoring increases frequency of measurements



AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Real-life data

reddit | r/MachineLearning | Find a community, post, or user | LOG IN

captkrob 2 points · 1 year ago

If you want an idea of a "real world" disease dataset, try looking at the MIMIC-III database (https://mimic.physionet.org/).

It's probably significantly more work to request the data and set it up in an easy format for ML than you had ever planned on, but this is the kind of thing actual data scientists and informaticians working on this kind of problem are used to dealing with. The dataset is tremendously rich, but also incredibly noisy. In other words, much like medical practice.

Share Save

Cleaning should visible and reusable

Encourage data processing pipelines, cohort selection to be shared in a public code repository.

Journal of the American Medical Informatics Association, 25(1), 2018, 32–39
doi: 10.1093/jamia/ocx084
Advance Access Publication Date: 27 September 2017
Research and Applications

Research and Applications

The MIMIC Code Repository: enabling reproducibility in critical care research

Alistair EW Johnson,[1,*] David J Stone,[2] Leo A Celi,[1,3] and Tom J Pollard[1]

[1]Massachusetts Institute of Technology, Cambridge, MA, USA, [2]University of Virginia School of Medicine, Charlottesville, VA, USA, and [3]Beth Israel Deaconess Medical Center, Boston, MA, USA

*Corresponding Author: Alistair EW Johnson, Massachusetts Institute of Technology, E25-505, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. E-mail: aewj@mit.edu.

Received 25 May 2017; Revised 11 July 2017; Editorial Decision 21 July 2017; Accepted 27 July 2017

ABSTRACT

Objective: Lack of reproducibility in medical studies is a barrier to the generation of a robust knowledge base to support clinical decision-making. In this paper we outline the Medical Information Mart for Intensive Care (MIMIC) Code Repository, a centralized code base for generating reproducible studies on an openly available critical care dataset.

Materials and Methods: Code is provided to load the data into a relational structure, create extractions of the data, and reproduce entire analysis plans including research studies.

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Care data are selected by policy

- Missingness is **MNAR by construction** — a value is absent because of a *decision* (not ordered ⇒ not worried).
- In EHR data, “controls” often mean patients who were not tested, treated, coded, or escalated, not a random sample of true negatives.
- Conditioning on a measurement (e.g. "lab was drawn") leads to **sampling bias**.

> The data encodes the *policy* that produced it — and policies are not exchangeable across sites or time.

Each modality captures measurement *decisions*

MODALITY	THE MEASUREMENT IS ITSELF A DECISION…	…WHICH BREAKS
Labs	ordered when a clinician is concerned	condition on "drawn" → sampling bias
Vitals	sampled faster when acuity is high	frequency is a severity proxy, not noise
Notes	dictated selectively, templated	presence ≠ incidence; documentation drift
Dx / billing codes	coded for reimbursement	prevalence tracks coding incentives
Imaging	protocol- and referral-gated	case-mix bias; site-specific acquisition

Design choices influence downstream use

PhysioNet

Share About FAQs Explore

Database | Credentialed Access

MIMIC-CXR Database

Alistair Johnson , Tom Pollard , Roger Mark , Seth Berkowitz , Steven Horng

Views Over Time

Month	Unique Viewers
June 2026	94
May 2026	129
April 2026	162
March 2026	167
February 2026	120
January 2026	163

PhysioNet

Share About FAQs Exp

Database | Credentialed Access

MIMIC-CXR-JPG - chest radiographs

Alistair Johnson , Matthew Lungren , Yifan Peng , Zhiyong Lu , Roger Mark ,

Views Over Time

Month	Unique Viewers
June 2026	330
May 2026	458
April 2026	555
March 2026	508
February 2026	418
January 2026	412

AHLI HEALTH AI SUMMER CAMP 2026

GETTING THE DATA

De-identification, governance, and access

AHLI

ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

De-identification and governance

- Data is shared under Health Insurance Portability and Accountability Act of 1996 (Safe Harbor's 18 identifiers, expert determination).

> "A major goal of the Privacy Rule is to assure that individuals' health information is properly protected while allowing the flow of health information needed to provide and promote high quality health care and to protect the public's health and well being." .. "The Rule strikes a balance that permits important uses of information, while protecting the privacy of people who seek care and healing."

> "There are no restrictions on the use or disclosure of de-identified health information. De-identified health information neither identifies nor provides a reasonable basis to identify an individual."

AHLI HEALTH AI SUMMER CAMP 2026

1. Manual review (time-consuming)

- Scan infrequent values in string fields (e.g. group lab tests, review tail)
- Scrape news articles

TICKETS SCHEDULE

Injured woman hit by ball released from hospital

By Alec Shirkey
July 11, 2015

BOSTON -- Fenway Park became the scene of another scary situation during the Red Sox's **5-1 loss** to the Yankees on Friday night when a woman was injured by a foul ball.

A Beth Israel Deaconess spokesman said the hurt woman was treated and released from the hospital on Saturday.

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

2. Pattern matching (predictable, fragile)

- Known identifiers, e.g. match variations on:
  - Patient name and location
  - Dates
  - "VIP list"
- Document structure:
  - Fields like "Name: XXX"
- PHI structure:
  - Dates (YYYY/MM/DD)
  - Email (username@domain.tld)

3. Machine learning (unpredictable, flexible)

Deidentification of free-text medical records using pre-trained bidirectional transformers

Alistair E. W. Johnson
aewj@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

Lucas Bulgarelli
lucas1@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

Tom J. Pollard
tpollard@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

ABSTRACT

The ability of caregivers and investigators to share patient data is fundamental to many areas of clinical practice and biomedical research. Prior to sharing, it is often necessary to remove identifiers such as names, contact details, and dates in order to protect patient privacy. Deidentification, the process of removing identifiers, is challenging, however. High-quality annotated data for developing models is scarce; many target identifiers are highly heterogenous (for example, there are uncountable variations of patient names); and in practice anything less than perfect sensitivity may be considered a failure. As a result, patient data is often withheld when sharing would be beneficial, and identifiable patient data is often

1 INTRODUCTION

The advent of large, open access text corpuses and the resurgence of neural networks has driven advances in state-of-the-art model performance in natural language processing [10, 27]. Barriers to sharing clinical text, however, have stifled progress in the medical domain. An unintended consequence is that research has become hyperfocused on the few datasets that are readily accessible. MIMIC-III, one of the only public sources of electronic health record data, for example, has been referred to as "one of the most (over)analyzed clinical datasets" [12, 29]. This paucity of data is to the detriment of important issues including bias, generalizability, and reproducibility [5].

https://dl.acm.org/doi/pdf/10.1145/3368555.3384455

AHLI HEALTH AI SUMMER CAMP 2026

Deidentification is not bias-free

In the Name of Fairness: Assessing the Bias in Clinical Record De-identification

Yuxin Xiao*
yuxin102@mit.edu
Massachusetts Institute of Technology
USA

Shulammite Lim*
shulim@mit.edu
Massachusetts Institute of Technology
USA

Tom Joseph Pollard
tpollard@mit.edu
Massachusetts Institute of Technology
USA

Marzyeh Ghassemi
mghassem@mit.edu
Massachusetts Institute of Technology
USA

ABSTRACT

Data sharing is crucial for open science and reproducible research, but the legal sharing of clinical data requires the removal of protected health information from electronic health records. This process, known as de-identification, is often achieved through the use of machine learning algorithms by many commercial and open-source systems. While these systems have shown compelling results on average, the variation in their performance across different demographic groups has not been thoroughly examined. In this work, we investigate the bias of de-identification systems on names in clinical notes via a large-scale empirical analysis. To achieve this, we create 16 name sets that vary along four demographic dimensions: gender, race, name popularity, and the decade of popularity. We insert these names into 100 manually curated clinical

ACM Reference Format:
Yuxin Xiao, Shulammite Lim, Tom Joseph Pollard, and Marzyeh Ghassemi. 2023. In the Name of Fairness: Assessing the Bias in Clinical Record De-identification. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3593013.3593982

1 INTRODUCTION

The increased availability of clinical datasets [51, 73, 74] plays a significant role in the recent advancements in machine learning (ML)-aided healthcare systems [15, 39, 107, 118]. In order to share clinical trial data legally, stakeholders must adhere to the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provisions by masking 18 types of protected health information (PHI). If done

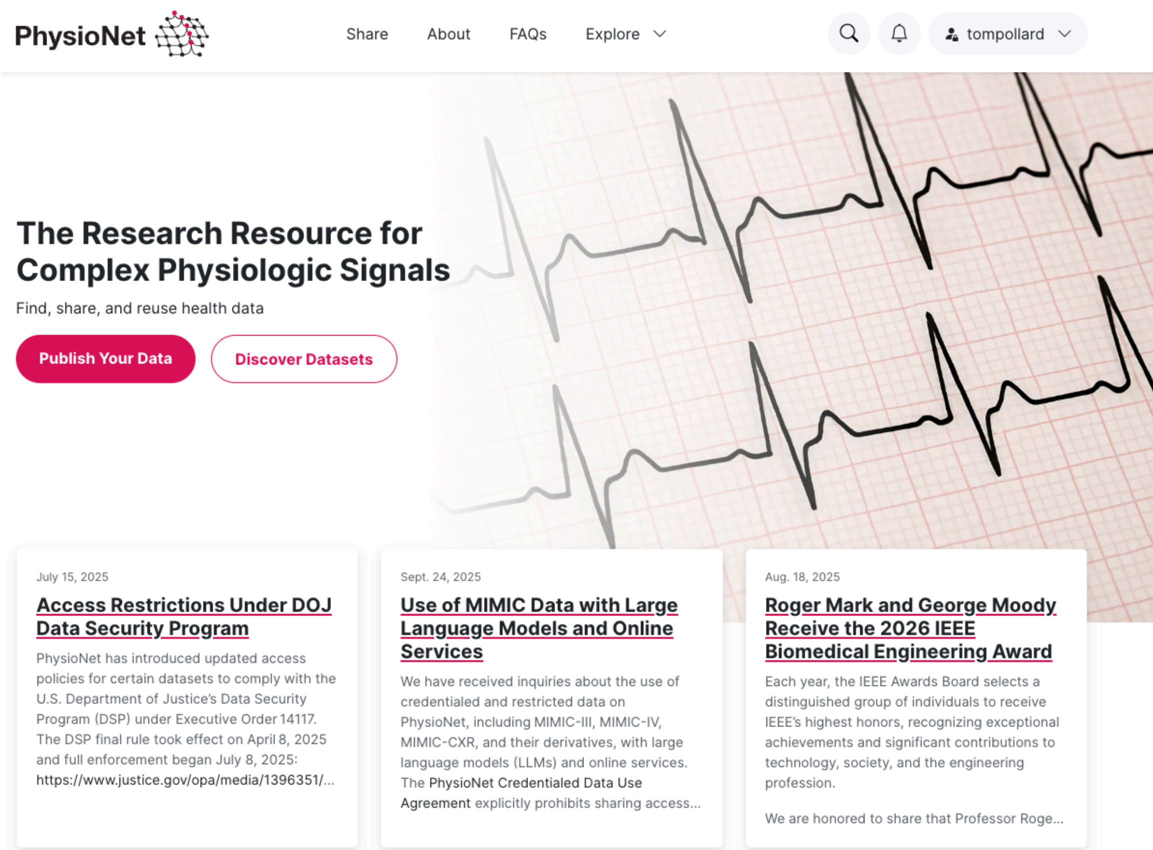
AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Access mechanics

1. **Credentialing**: human-subjects training, approved identity.
2. **Data use agreement**: sign and comply with the DUA.

Realistic timelines: weeks to months.

Start early — before you need the data.



Increasing value of health data makes it a target

Even well governed resources face downstream misuse.

FINANCIAL TIMES

HOME WORLD UK COMPANIES TECH MARKETS CLIMATE OPINION LEX STOCK PICKING GAME LIFE & ARTS HOW TO SPEND IT

UK Biobank Ltd. + Add to myFT

Medical data of 500,000 people in UK for sale on Chinese site

Alibaba 'swiftly removed' listings for UK Biobank data, institution says

Save

UK Biobank stores genetic data collected over more than 15 years © Dave Guttridge/UK Biobank

Michael Peel and **Chris Smyth**

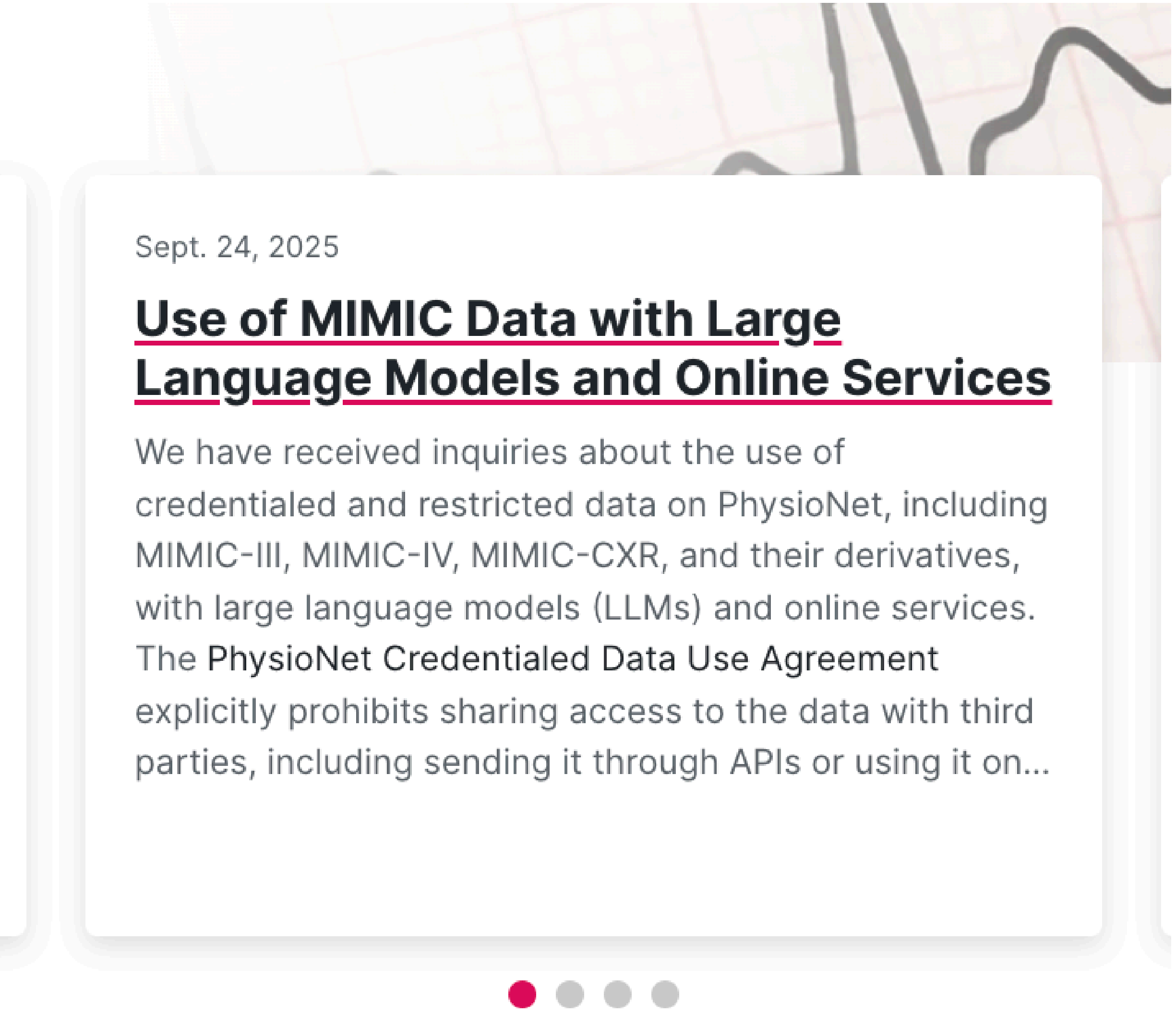
Published APR 23 2026

168

LLMS

Researchers want to send data to external services (e.g. LLMs) for their research.

This involves sharing with a 3rd party (not allowed under our Data Use Agreement)



Sept. 24, 2025

Use of MIMIC Data with Large Language Models and Online Services

We have received inquiries about the use of credentialed and restricted data on PhysioNet, including MIMIC-III, MIMIC-IV, MIMIC-CXR, and their derivatives, with large language models (LLMs) and online services. The PhysioNet Credentialed Data Use Agreement explicitly prohibits sharing access to the data with third parties, including sending it through APIs or using it on...

Membership attacks

High capacity models can memorize data. Membership attacks become a real threat.

External data: insurance claim: age, weekend accident, injuries

Released clinical data: de-identified admission: same age, same timing, same injuries

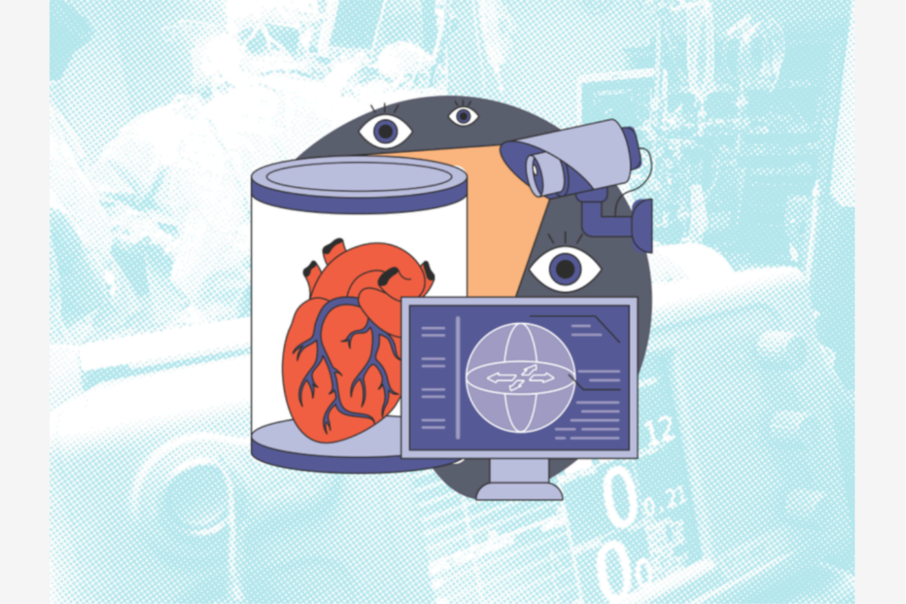
New information: clinical note shows suspected intoxication

MIT scientists investigate memorization risk in the age of clinical AI

New research demonstrates how AI models can be tested to ensure they don't cause harm by revealing anonymized patient health data.

Alex Ouyang | Abdul Latif Jameel Clinic for Machine Learning in Health
January 5, 2026

PRESS INQUIRIES



MIT scientists are developing tests to ensure that AI models aren't memorizing sensitive patient information.

Image: Alex Ouyang/MIT Jameel Clinic, with Adobe Stock

Regulation is changing

Department of Justice recently introduced the Data Security Program

Limits sharing of deidentified health data with "Countries of Concern"

Preventing Access to U.S. Sensitive Personal Data and Government-Related Data by Countries of Concern or Covered Persons

A Rule by the Justice Department on 01/08/2025

PDF
Document Details
Document Dates
Table of Contents
Related Documents
Public Comments
Regulations.gov Data
Sharing

PUBLISHED DOCUMENT: 2024-31486 (90 FR 1636)

AGENCY:

National Security Division, Department of Justice.

SUMMARY:

The Department of Justice is issuing a final rule to implement Executive Order 14117 of February 28, 2024 (Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern), by prohibiting and restricting certain data transactions with certain countries or persons.

DATES:

This rule has been classified as meeting the criteria under 5 U.S.C. 804(2) and is effective April 8, 2025. However, at the conclusion of the Congressional review, if the effective date has been changed, the Department of Justice will publish a document in the **Federal Register** to establish the actual date of effectiveness or to terminate the rule. The incorporation by reference of certain material listed in this rule is approved by the Director of the Federal Register as of April 8, 2025.

Public data enables (weak?) research

Public data enables (weak?) research

Study explored trends in publications using public datasets

“excess of 12,000 papers in 2025”

“measures to control access to open data are required”


Journal of Clinical Epidemiology
Volume 193, May 2026, 112203

Methodological Aspects of Research Integrity and Culture Series

Quantifying new threats to health and biomedical literature integrity from rapidly scaled publications and problematic research

Matt Spick [a], Anthony Onoja [a], Charlie Harrison [b], Stefan Stender [c], Jennifer Byrne [d e], Nophar Geifman [a]

Highlights

Key findings

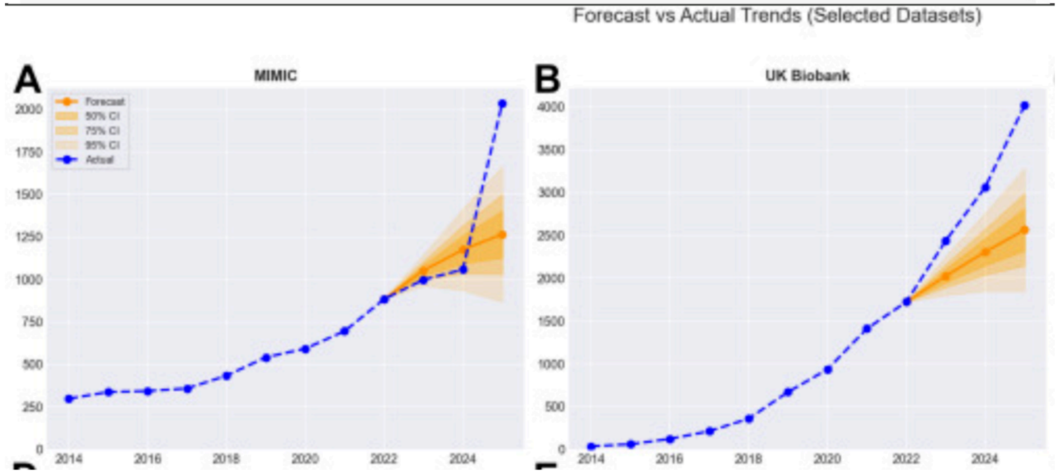
- We estimate an annual excess of 12 thousand papers exploiting Open Science in 2025

What this adds to what is known?

- Growth in AI-assisted low-quality paper production using Open Data has exploded - AI papers are introducing false discoveries and misleading claims to the literature

What is the implication and what should change now?

- Measures to deter formulaic submissions and control access to Open Data are required



ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Public datasets can create oversaturation

Healthcare, Machine Learning

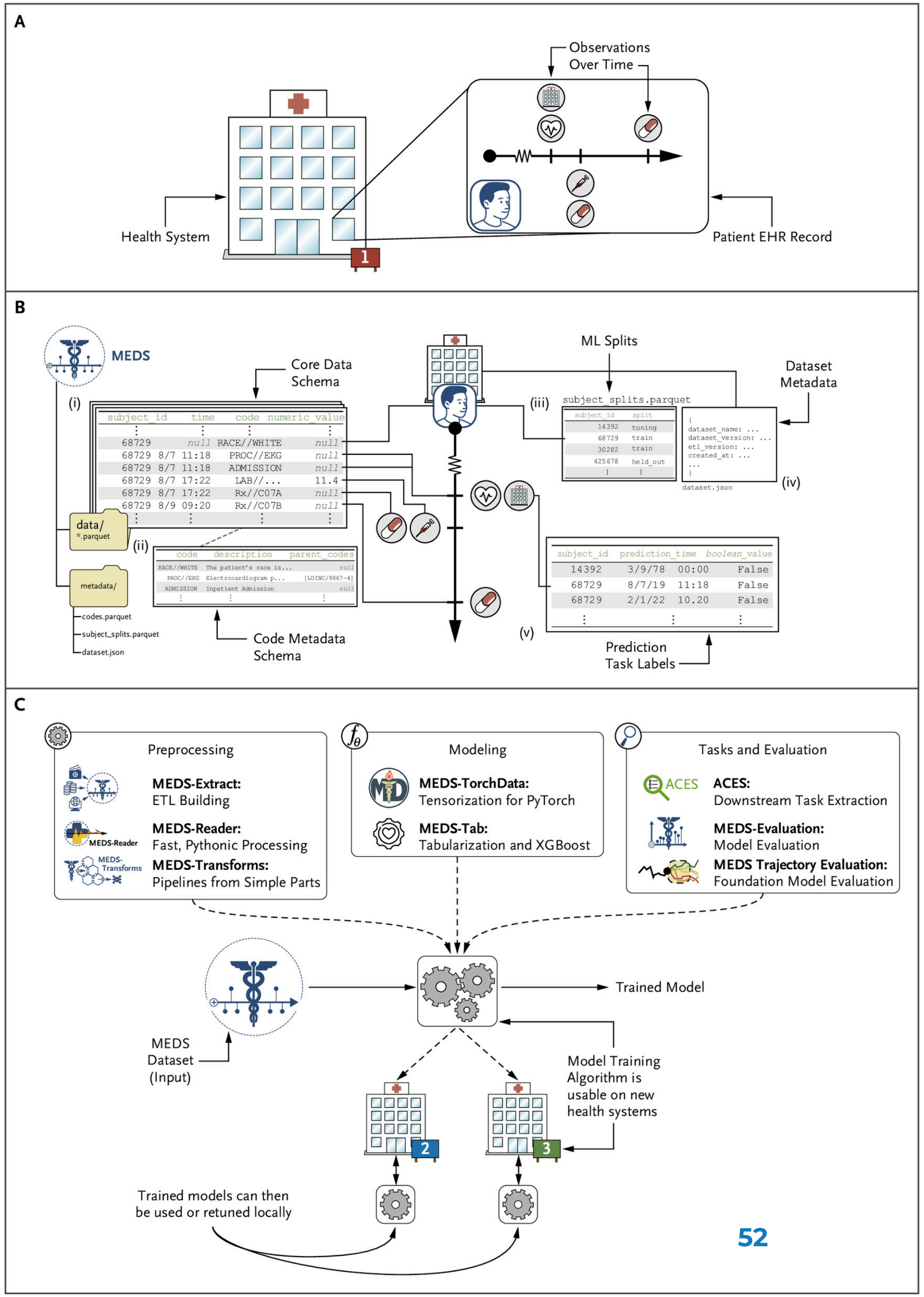
The Shaky Foundations of Foundation Models in Healthcare

Scholars detail the current state of large language models in healthcare and advocate for better evaluation frameworks.

Feb 27, 2023 |
Michael Wornow, Yizhe Xu, Birju Patel, Rahul Thapa, Ethan Steinberg, Scott Fleming, Jason Fries, Nigam Shah

Standards solve syntax; representation is still open

- A common schema — **MEDS**, events as `(patient, time, code, value)` — makes data *portable*.
- **How to represent** irregular, multimodal, multi-scale event streams is an **open research problem** (e.g. tokenization, time encoding, and concept vocabularies).
- Standardization reduces busywork, allowing you to focus on the modeling question (Day 4).



AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

AHLI HEALTH AI SUMMER CAMP 2026

READING BETWEEN THE LINES

What the data hides

AHLI

ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Provenance-driven bias

- **Time:** coding, testing, and treatment change; the model learns the era.
- **Cohort definitions:** cases and controls encode how patients were found.
- **Selection:** the dataset contains who entered the system, not everyone at risk.
- **Documentation:** codes and notes reflect workflow, incentives, and physiology.

> Ask: is this signal clinical, or is it a trace of how the data were produced?

When a label is a billing artifact

A diagnosis code looks like ground truth. It isn't.

- A code may appear because it **justifies reimbursement**, not because it's the primary clinical concern.
- It may be **absent** for a real condition that wasn't billed.
- Prevalence in the data tracks **coding incentives** as much as biology.

Train on that label, and you partly learn the hospital's billing rules.

↩ Day 1: label noise

→ Day 3: estimands

Measurements are not neutral

JAMA Internal Medicine
Search All
Enter Search Term

This Issue | Views 25,689 | Citations 59 | Altmetric 2398

Download PDF | More | Cite This | Permissions

Original Investigation
FREE

July 11, 2022

Assessment of Racial and Ethnic Differences in Oxygen Supplementation Among Patients in the Intensive Care Unit

Eric Raphael Gottlieb, MD, MS[1,2,3]; Jennifer Ziegler, MD, MSc[4]; Katharine Morley, MD, MPH[2,5]; et al

Author Affiliations | Article Information

JAMA Intern Med. 2022;182(8):849-858. doi:10.1001/jamainternmed.2022.2587

Editorial Comment
Related Articles

Key Points

Question Are there differences in supplemental oxygen administration among patients of different races and ethnicities associated with pulse oximeter performance discrepancies?

Findings In this cohort study of 3069 patients in the intensive care unit, Asian, Black, and

AHLI HEALTH AI SUMMER CAMP 2026

X-rays encode race

THE LANCET
Digital Health

ARTICLES | VOLUME 4, ISSUE 6, E406-E414, JUNE 2022 Download Full Issue

AI recognition of patient race in medical imaging: a modelling study

Judy Wawira Gichoya, MD • Imon Banerjee, PhD • Ananth Reddy Bhimireddy, MS • John L Burns, MS • Leo Anthony Celi, MD • Li-Ching Chen, BS • Ramon Correa, BS • Natalie Dullerud, MS • Marzyeh Ghassemi, PhD • Shih-Cheng Huang • Po-Chih Kuo, PhD • Matthew P Lungren, MD • Lyle J Palmer, PhD • Brandon J Price, MD • Saptarshi Purkayastha, PhD • Ayis T Pyrros, MD • Lauren Oakden-Rayner, MD • Chima Okechukwu, MS • Laleh Seyyed-Kalantari, PhD • Hari Trivedi, MD • Ryan Wang, BS • Zachary Zaiman • Haoran Zhang, MS • Show less

Open Access • Published: May 11, 2022 • DOI: https://doi.org/10.1016/S2589-7500(22)00063-2 •
Check for updates

Summary

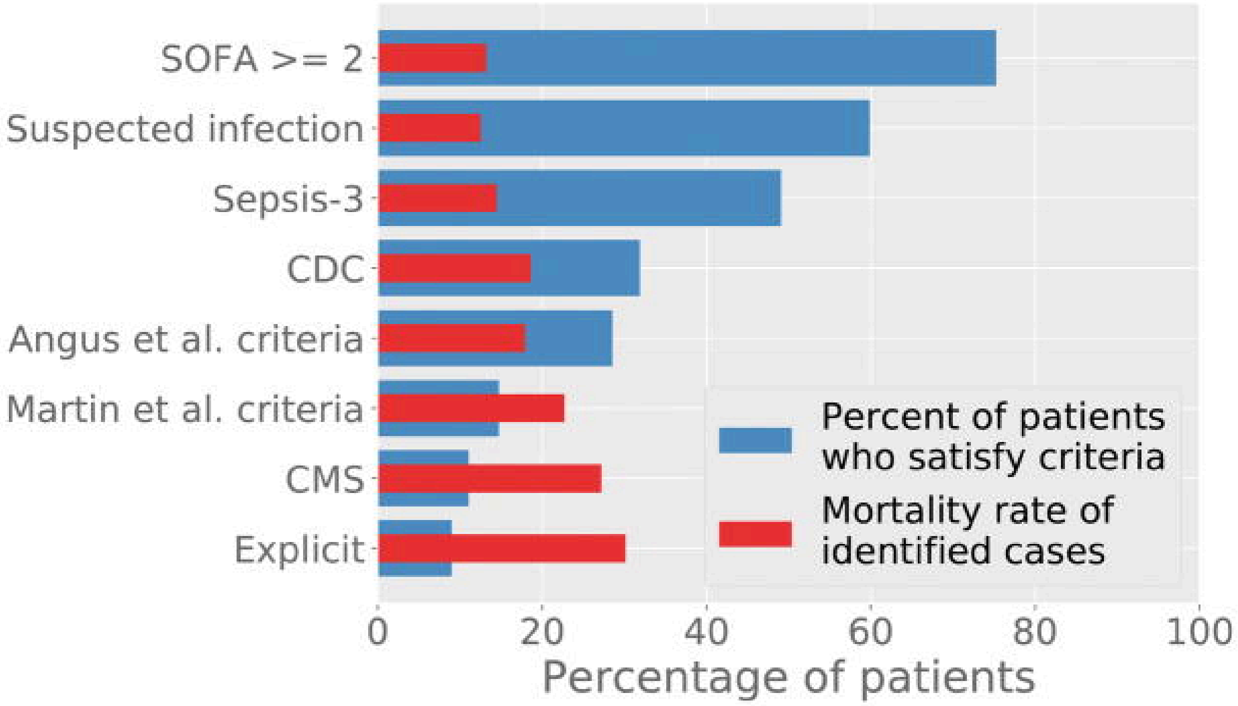
Background

Previous studies in medical imaging have shown disparate abilities of artificial intelligence (AI) to detect a person's race, yet there is no known correlation for race on medical imaging that would be obvious to human experts when interpreting the images. We aimed

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Cohort definitions are important

A comparative analysis of sepsis identification methods in an electronic database. Crit Care Med. 2018 Apr;46(4):494–499. doi: 10.1097/CCM.0000000000002965



Different systems use different clocks

Temporal relationships between different streams may be highly informative.

Different clocks on different systems obscure these relationships.

Datasets are often a poor representation of clinical reality

Many public datasets are poor representations of clinical reality.

Promote research that isn't deployable.

JAMA Network Open

Home | Issues | Multimedia

Home | JAMA Network Open | Vol. 8, No. 12

Original Investigation | Health Informatics

Diagnostic Codes in AI Prediction Models and Label Leakage of Same-Admission Clinical Outcomes

Bashar Ramadan, MBBS[1]; Ming-Chieh Liu, MS[1]; Michael C. Burkhart, PhD[1]; et al

Author Affiliations | Article Information

RELATED ARTICLES FIGURES SUPPLEMENTAL CONTENT

Key Points

Question Are *International Classification of Diseases* (*ICD*) diagnostic codes, which are only finalized after hospital discharge, associated with inflated performance of artificial intelligence (AI) health care prediction models?

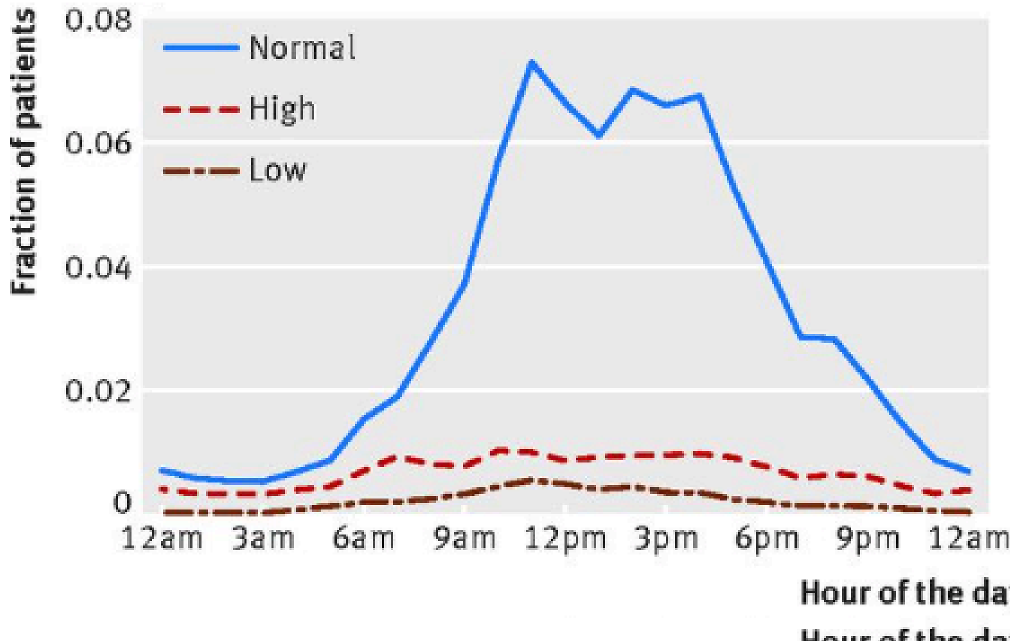
Findings In this prognostic study of 180 640 patients, 40.2% of published AI models trained to predict same-admission outcomes used *ICD* codes as features. Prediction models for inpatient mortality trained on *ICD* codes predicted in-hospital mortality with high accuracy, with the most important codes (eg, brain death, encounter for palliative care) not available in time for clinically useful mortality prediction.

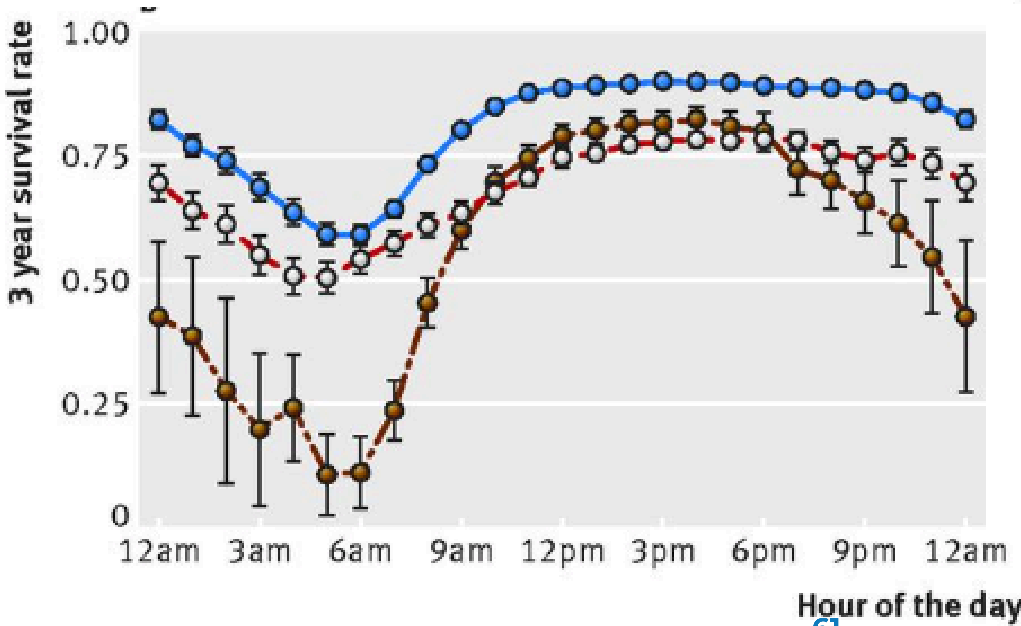
Timing of laboratory tests

> "the timing of when laboratory tests were ordered was more accurate than the test results in predicting survival in 118 of 174 tests

> patients with normal white blood cell count values taken at 4 am have lower survival (85.4%) than patients with an abnormal measurement at 4 pm

Biases in electronic health record data due to processes within the healthcare system. BMJ 2018;361:k1479





ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Dataset shift

Occurs when the distribution of data changes between training and deployment

Leads to deterioration of model performance.

https://www.nejm.org/doi/full/10.1056/NEJMc2104626


The New England Journal of Medicine NEJM Evidence NEJM AI NEJM Catalyst NEJM Journal Watch

The NEW ENGLAND JOURNAL of MEDICINE

CURRENT ISSUE SPECIALTIES TOPICS

CORRESPONDENCE

The Clinician and Dataset Shift in Artificial Intelligence

Published July 14, 2021 | N Engl J Med 2021;385:283-286 | DOI: 10.1056/NEJMc2104626 | VOL. 385 NO. 3

TO THE EDITOR:

Artificial intelligence (AI) systems are now regularly being used in medical settings,[1] although regulatory oversight is inconsistent and undeveloped.[2,3] Safe deployment of clinical AI requires informed clinician-users, who are generally responsible for identifying and reporting emerging problems. Clinicians may also serve as administrators in governing the use of clinical AI. A natural question follows: are clinicians adequately prepared to identify circumstances in which AI systems fail to perform their intended function reliably?

A major driver of AI system malfunction is known as “dataset shift.”[4,5] Most clinical AI systems today use machine learning, algorithms that leverage statistical methods to learn key

Changes in ICD codes

Examined insurance claims for >18m people in the US from 2010 to 2017

Transition from ICD-9 to ICD-10 led to instantaneous +/- of >20% in the prevalence of many diagnostic categories.

https://jamanetwork.com/journals/jamanetworkopen/full

Original Investigation | Health Policy

Diagnostic Category Prevalence in 3 Classification Systems Across the Transition to the *International Classification of Diseases, Tenth Revision, Clinical Modification*

Randall P. Ellis, PhD[1]; Heather E. Hsu, MD, MPH[2]; Chenlu Song, MA[1]; et al

Author Affiliations | Article Information

RELATED ARTICLES FIGURES SUPPLEMENTAL CONTENT

Key Points

Question Was the transition from *International Classification of Diseases, Ninth Revision, Clinical Modification* (*ICD-9-CM*) to the *Tenth Revision* (*ICD-10-CM*) in October 2015 associated with changes in diagnostic category prevalence when diagnoses are grouped by classification system?

Findings This interrupted time series analysis and cross-sectional study examined insurance claims for more than 18 million privately insured adults and children in the US from 2010 to 2017 and found instantaneous increases or decreases of 20% or more associated with the *ICD-10-CM* transition for nearly 1 in 6 (16%) diagnostic categories

Who is missing entirely

Perhaps the most significant bias is in the patients who **never enter the data.**

- Under-served populations under-appear — or appear only at crisis.
- Missingness is **informative**: absence reflects access, not health.
- A model trained on who showed up encodes who **didn't.**

→ Days 3 & 5: equity

→ data fusion (next)

Frontier — "more data" can make it worse

- Combining sites is **data fusion**, not addition: each source carries its **own selection mechanism.**
- Pooling without modeling those mechanisms can **compound** bias — and *worsen* transport to a new site.

> The instinct "just get more data" is right only if you know what each source selected for.

↩ Day 1: more sites ≠ more general

→ Day 5: transport

Data answers the question it was generated for

- Spending and utilization data reflect the **health system**, not just health.
- There is almost always a **mismatch** between what the data *measures* and what you *want to know*.
- Name that gap explicitly — before it silently becomes your result.

AHLI HEALTH AI SUMMER CAMP 2026

BACK TO YOUR OWN PROJECTS

Interrogating your own dataset

AHLI

ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Four questions to ask of your data:

1. **Where did it come from** — what process generated it, for what purpose?
2. **Who is missing** — who never enters, who is under-measured?
3. **What does it silently encode** — coding, workflow, system incentives?
4. **Does it support your Day 1 problem** — or quietly change it?

The gap check

> Does the data that *exists* actually support the problem you posed yesterday?

- If **yes** — say how, specifically.
- If **no** — you must **revise the problem, the data plan, or both.**

A problem statement without a data spec is still a wish.

Your data specification

Interrogate the data your project depends on.

Today's deliverable: Data specification sheet.
Consider **what data exists**, the **generating process**, your **access plan**, and the **anticipated provenance-driven bias.**

Next:

- 11:00am: small group, interrogate your data
- 1:30pm: Sage Bionetworks discuss data beyond the EHR