STAT 992: Foundation Models for Biomedical Data

Ben Lengerich

Lecture 14: Foundation Models for EHRs

March 23, 2026

Previously: Imaging, Genomics

- Massive unlabeled archives
- Expensive annotation
- Representation reuse
- Distribution shift
- Shortcut learning

These domains have **stable data-generating processes**.

Today: Electronic Health Records (EHRs)

- What is the data-generating process in EHRs?
  - biology
  - patient decisions
  - clinician decisions
  - hospital workflows

No single **stable data-generating process**?

The promise of Real-World Evidence and EHRs

- Massive observational data
- Continuous patient monitoring
- Real-world clinical decisions

Challenges of EHR data:

	Imaging	Genomics	EHRs
Generation	Physics	Biology	**Humans + interventions**
Observation	Passive	Mostly passive	**Active (decisions)**
Measurement	Fixed	Fixed	**Adaptive**

Some terminology

- "Real world" → specific regulatory implications (FDA 2018):
  - Real-world **data** (RWD): data relating to patient health status and/or delivery of health routinely collected from a variety of sources
  - Real-world **evidence** (RWE): clinical evidence regarding the usage and potential benefits/risks of a medical produce derived from analysis of RWD
- Electronic Medical Record (**EMR**): digital version of a patient's paper chart
- Electronic Health Record (**EHR**): Multi-organizational EMR

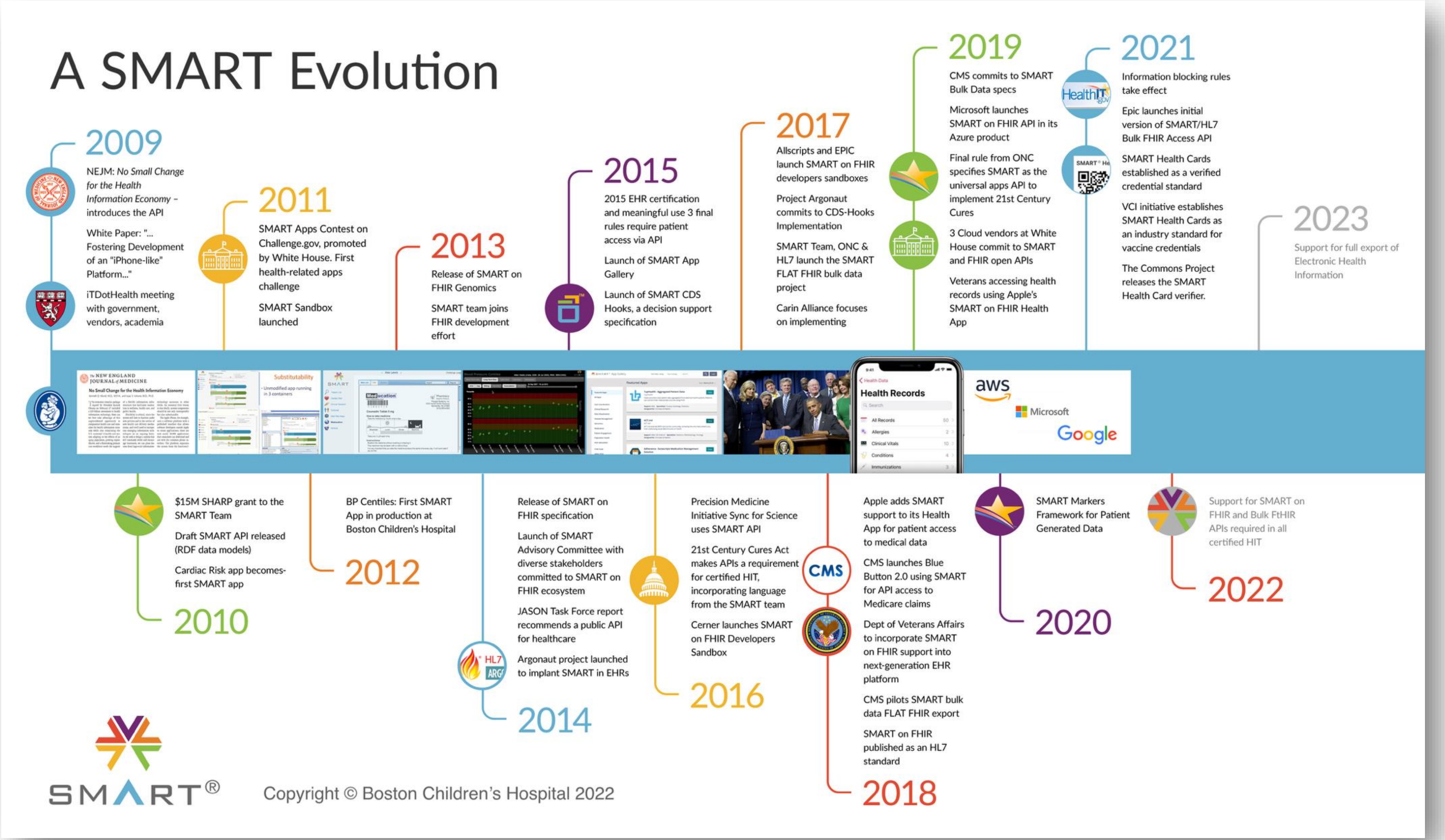
Components of EHRs

- Patient Details:
  - Demographics, background, exposures, genetics
- **Time Series** of encounter details:
  - Clinical notes: unstructured text, frequently recorded for the purpose of handoff between shifts.
  - Lab tests: Sometimes routine, sometimes ordered for a reason.
  - Vitals: BP, Temp, HR, RR…generally frequently updated
  - Treatments: Critical for data analyses. More later.
  - Input/Output: Food/water/waste…
  - Genetics: Rare in real-world datasets but common for targeted studies
  - Billing Details (ICD)
  - Outcomes: mortality, discharge, re-admission

Subjective: ANXIETY STATE NOS 300.00 DEPRESSIVE DISORDER NEC 311 ATRIAL FIBRILLATION 427.31 OLD MYOCARDIAL INFARCT 412 CONGESTIVE HEART FAILURE 428.0 Current outpatient prescriptions ** LOPRESSOR 50 MG PO TABS 1 tab two times a day 60 5	**Objective:** 250.00 DM, CONTROLLED, TYPE II (primary encounter diagnosis) 428.0 CONGESTIVE HEART FAILURE 585.3 KIDNEY DZ,CHRONIC (GFR>30–59) STAGE III 412 OLD MYOCARDIAL INFARCT 715.09 GENERAL OSTEOARTHROSIS 427.31 ATRIAL FIBRILLATION
Assessment: BP 122/68 | Pulse 78 | Temp (Src) 98.1 (Oral) | Resp 22 | Wt 227 lbs Abdomen: abdomen soft, non-tender, obese and no masses or organomegaly Back: No CVA tenderness Extremities: No edema	**Plan:** Continue present medication(s): Referral(s) to: eye Injection(s) ordered: b12 Schedule labs: Labs on return.

Sondhi et al 2012

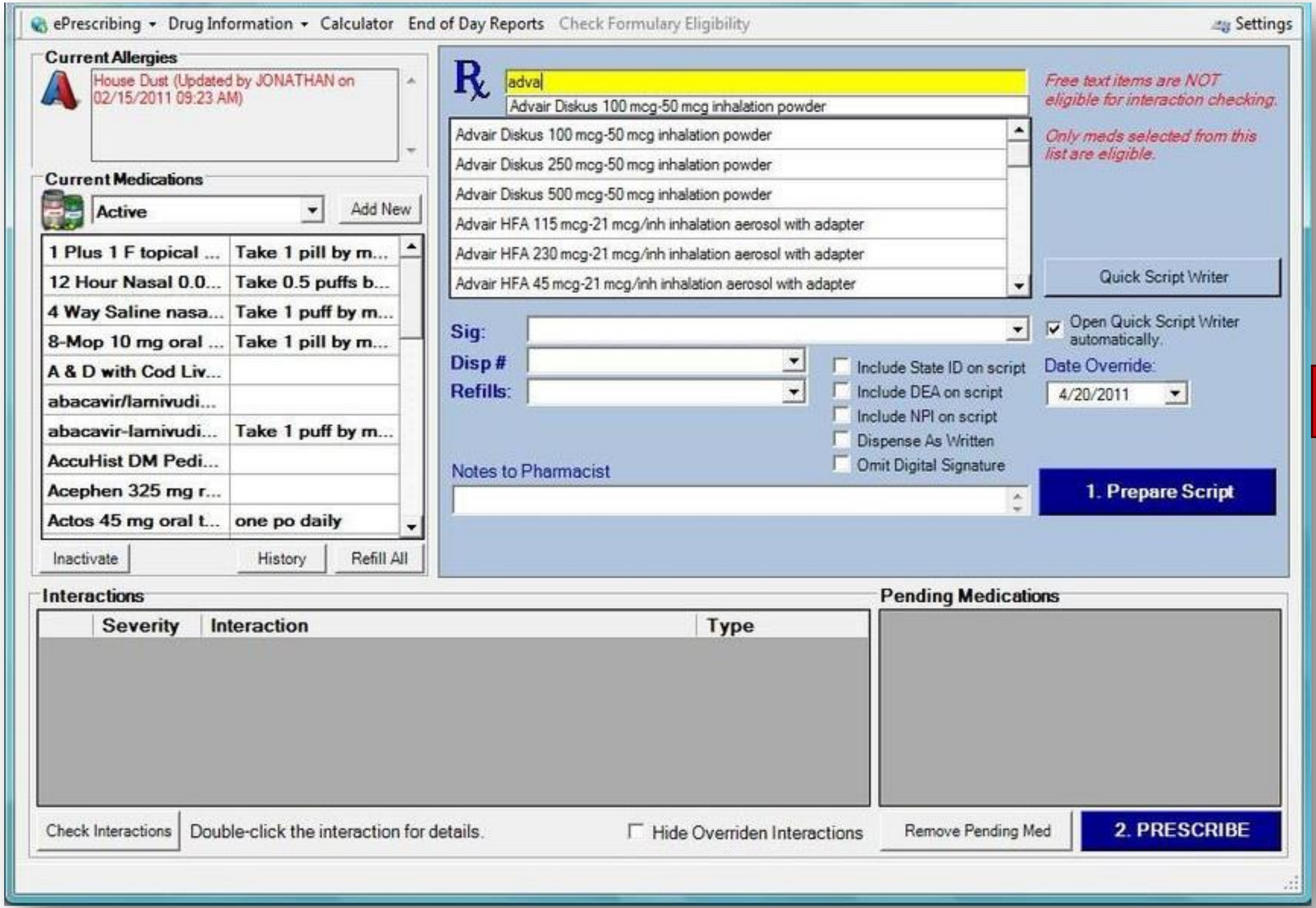
Toward standardization of EHRs



An exciting time: EHRs and foundation models

- Standardized data → opportunities
  - Applying pretrained LLMs for:
    - Streamlining clinical workflow
    - Interpreting clinical notes
  - Building foundation models of EHRs

Applying LLMs for streamlining clinical workflow



This can be improved!

Applying LLMs for interpreting clinical notes

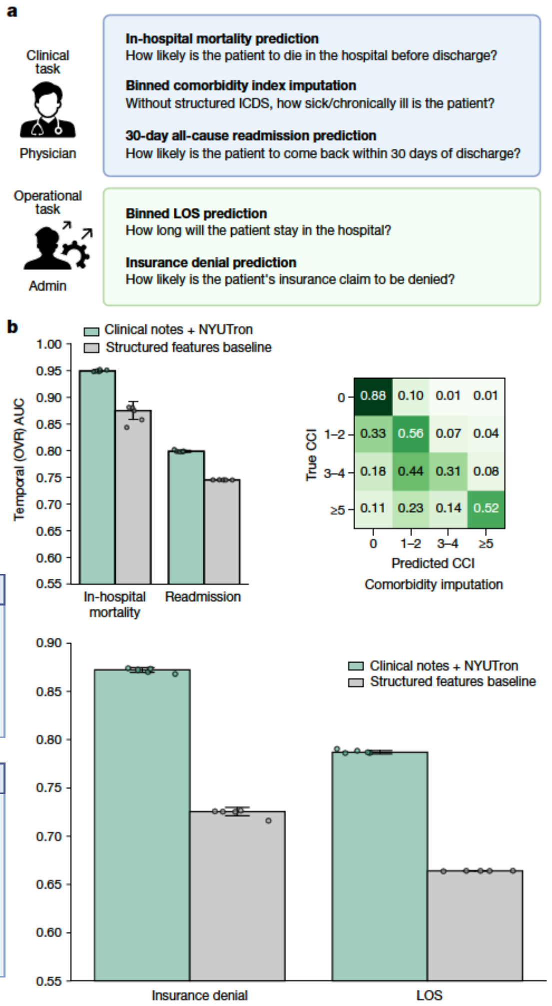
…Improves Predictive Performance

Article

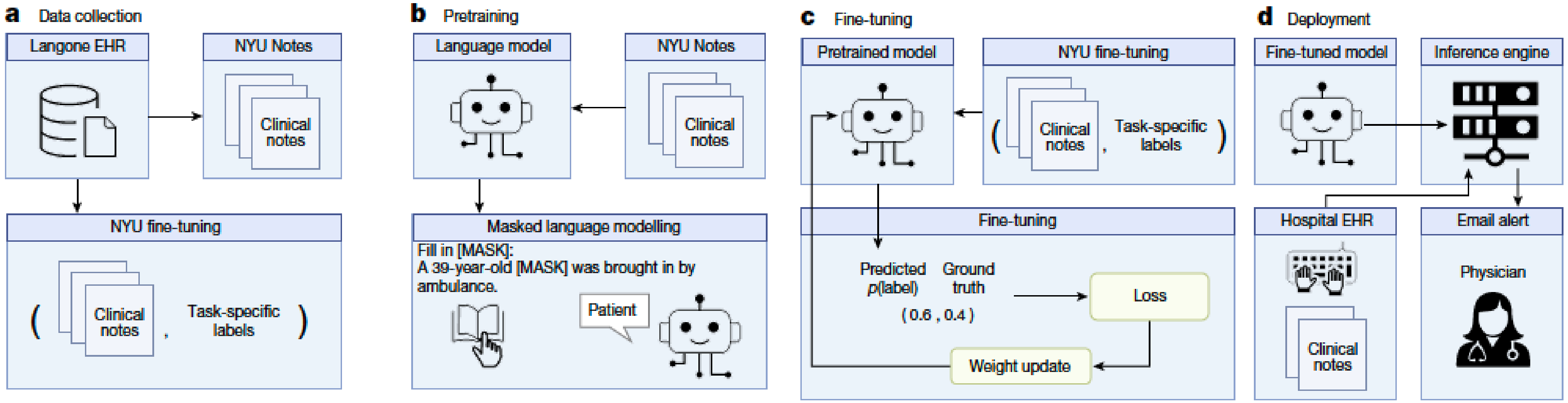
Health system-scale language models are all-purpose prediction engines

https://doi.org/10.1038/s41586-023-06160-y
Received: 14 October 2022
Accepted: 2 May 2023
Published online: 7 June 2023
Open access

Lavender Yao Jiang[1,2], Xujin Chris Liu[1,3], Nima Pour Nejatian[4], Mustafa Nasir-Moin[1], Duo Wang[5], Anas Abidin[4], Kevin Eaton[6], Howard Antony Riina[1], Ilya Laufer[1], Paawan Punjabi[6], Madeline Miceli[6], Nora C. Kim[1], Cordelia Orillac[1], Zane Schnurman[1], Christopher Livia[1], Hannah Weiss[1], David Kurland[1], Sean Neifert[1], Yosef Dastagirzada[1], Douglas Kondziolka[1], Alexander T. M. Cheung[1], Grace Yang[1,2], Ming Cao[1,2], Mona Flores[4], Anthony B. Costa[4], Yindalon Aphinyanaphongs[5,7], Kyunghyun Cho[2,8,9,10] & Eric Karl Oermann[1,2,11]



Massive Pretraining…



But general-purpose LLMs might be better…

Table 5: Performance of different models on multiple choice components of MultiMedQA [SAT+22]. GPT-4 outperforms GPT-3.5 and Flan-PaLM 540B on every dataset except PubMedQA. GPT-4 and GPT-3.5 were prompted with zero-shot direct prompts.

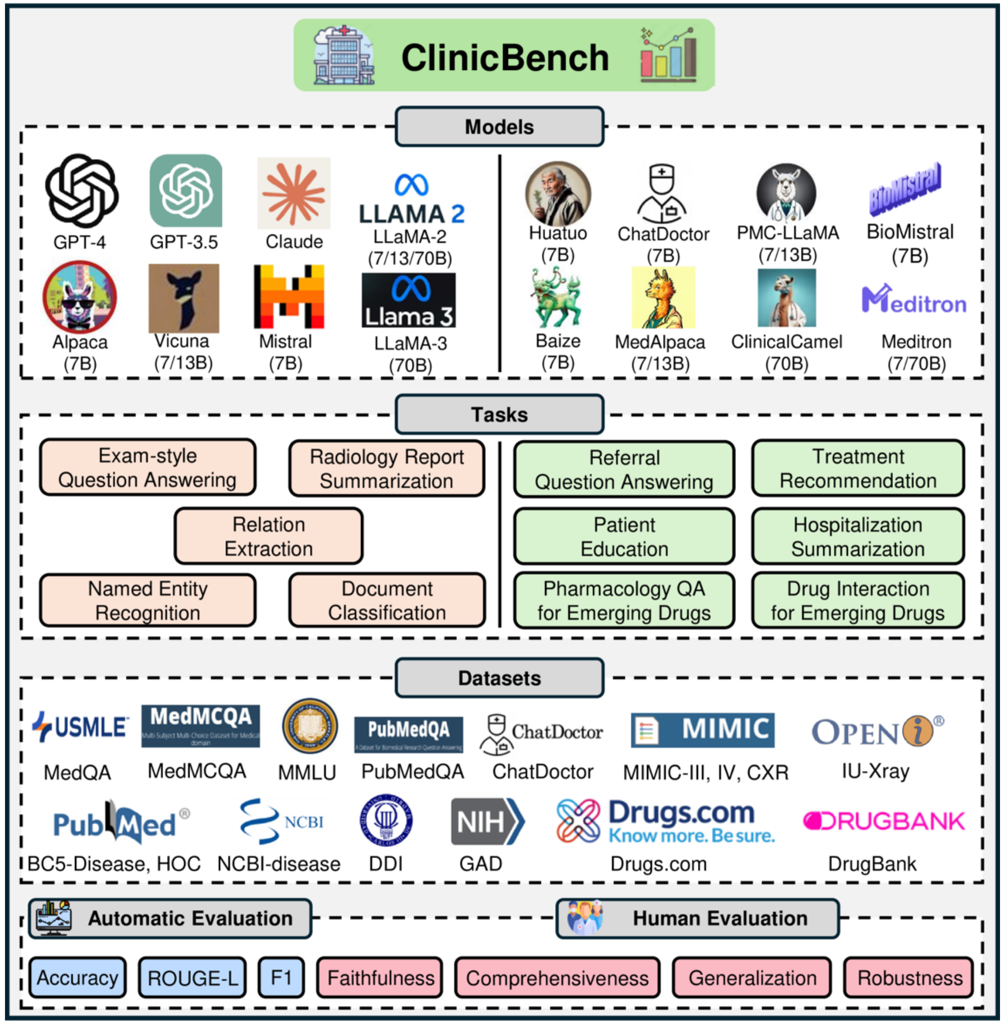
Dataset	GPT-4-base 5 shot / 0 shot	GPT-4 5 shot / 0 shot	GPT-3.5 5 shot / 0 shot	Flan-PaLM 540B* few shot
MedQA				
Mainland China	**78.63** / 74.34	75.31 / 71.07	44.89 / 40.31	–
Taiwan	**87.47** / 85.14	84.57 / 82.17	53.72 / 50.60	–
US (5-option)	**82.25** / 81.38	78.63 / 74.71	47.05 / 44.62	–
US (4-option)	**86.10** / 84.45	81.38 / 78.87	53.57 / 50.82	60.3**
PubMedQA				
Reasoning Required	77.40 / **80.40**	74.40 / 75.20	60.20 / 71.60	79.0
MedMCQA				
Dev	**73.66** / 73.42	72.36 / 69.52	51.02 / 50.08	56.5
MMLU				
Clinical Knowledge	**88.68** / 86.79	86.42 / 86.04	68.68 / 69.81	77.0
Medical Genetics	**97.00** / 94.00	92.00 / 91.00	68.00 / 70.00	70.0
Anatomy	82.96 / **85.19**	80.00 / 80.00	60.74 / 56.30	65.2
Professional Medicine	92.65 / **93.75**	**93.75** / 93.01	69.85 / 70.22	83.8
College Biology	**97.22** / 95.83	93.75 / 95.14	72.92 / 72.22	87.5
College Medicine	**80.92** / 80.35	76.30 / 76.88	63.58 / 61.27	69.9

* Sourced directly from [SAT+22]. We use Flan-PaLM 540B few-shot results as the most directly comparable setting to our experimental setup. The number of few shot prompts used by Flan-PaLM 540B varies per dataset (between 3 and 5).

** We note that [SAT+22] reports a preliminary performance of 67.2% here with Med-PaLM, a prompt-tuned variant of Flan-PaLM 540B, using an ensemble of chain-of-thought, few-shot prompts.

[Nori et al 2023]

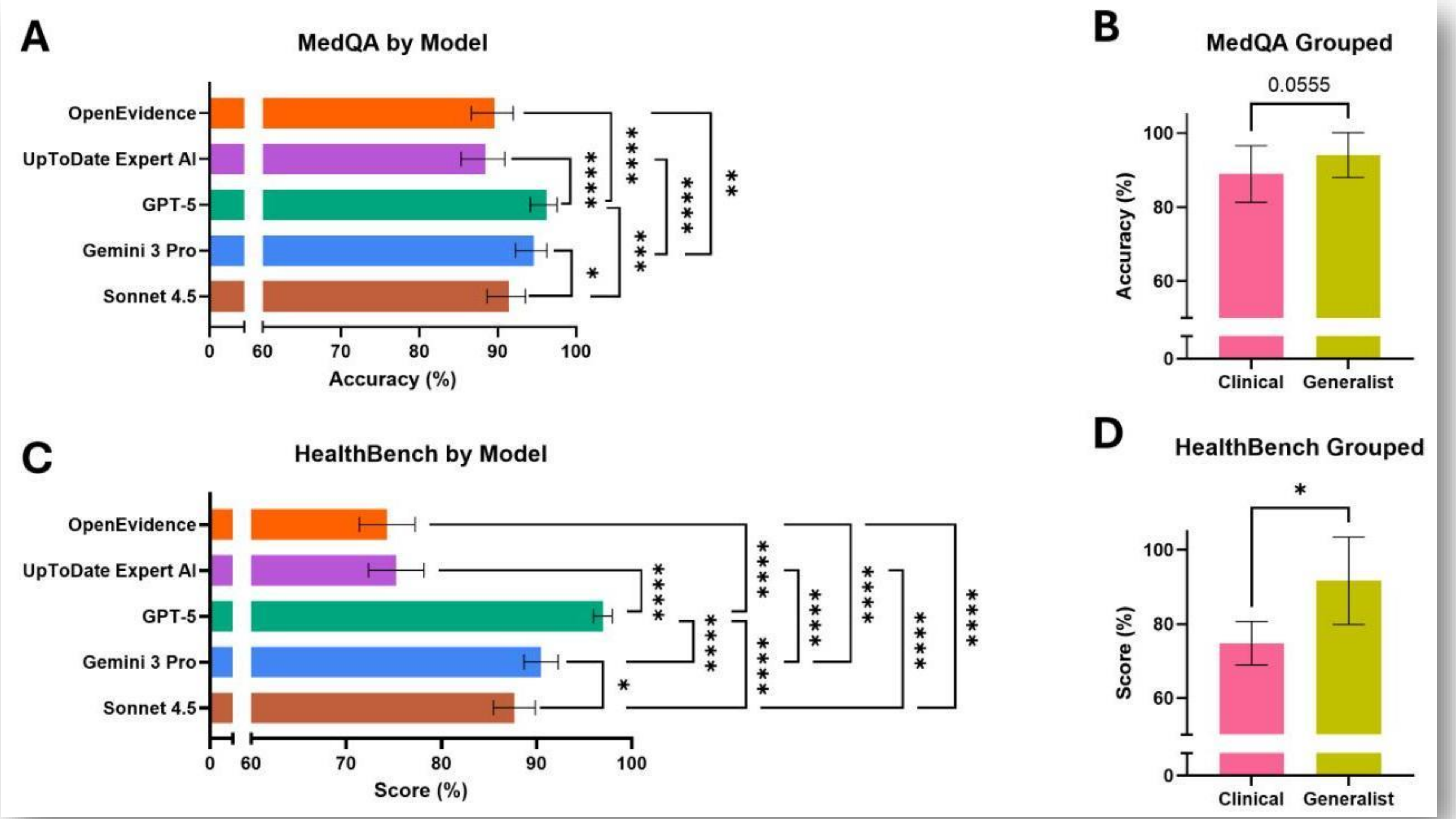
...or not?



Types	Methods	Hospi. Sum. F	Hospi. Sum. C	Hospi. Sum. P	Hospi. Sum. S	Patient Edu. F	Patient Edu. C	Patient Edu. P	Patient Edu. S
General Large Language Models	Alpaca	18.0	43.0	48.0	24.0	11.0	19.0	18.0	20.0
	Vicuna-7B	25.0	46.0	56.0	31.0	14.0	26.0	22.0	27.0
	LLaMA-2-7B	41.0	51.0	62.0	36.0	50.0	45.0	59.0	39.0
	Mistral	59.0	58.0	70.0	56.0	54.0	48.0	76.0	44.0
	Vicuna-13B	46.0	53.0	65.0	43.0	42.0	33.0	40.0	32.0
	LLaMA-2-13B	52.0	62.0	67.0	49.0	55.0	58.0	60.0	41.0
	LLaMA-2-70B	65.0	70.0	73.0	63.0	60.0	66.0	71.0	51.0
	LLaMA-3-70B	73.0	**81.0**	**85.0**	78.0	69.0	**75.0**	**83.0**	**77.0**
Medical Large Language Models	Baize-Healthcare	30.0	20.0	41.0	47.0	17.0	16.0	28.0	36.0
	MedAlpaca-7B	37.0	32.0	33.0	52.0	19.0	20.0	15.0	31.0
	Meditron-7B	63.0	55.0	58.0	64.0	57.0	50.0	47.0	59.0
	BioMistral	68.0	47.0	44.0	73.0	66.0	46.0	49.0	62.0
	PMC-LLaMA-13B	45.0	39.0	30.0	53.0	35.0	21.0	13.0	34.0
	MedAlpaca-13B	49.0	40.0	42.0	61.0	38.0	23.0	27.0	37.0
	ClinicalCamel	75.0	59.0	61.0	69.0	64.0	55.0	50.0	56.0
	Meditron-70B	**79.0**	72.0	54.0	**82.0**	**71.0**	60.0	67.0	74.0

[ClinicBench, 2024]

But probably general-purpose are better.



[Viswanaht 2025]

Building Foundation Models of EHRs

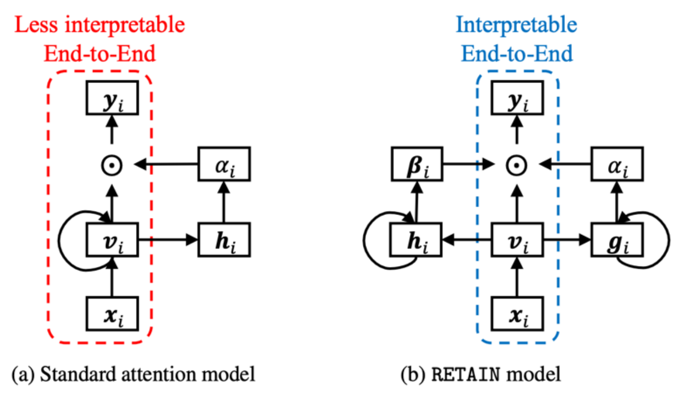
Early Machine Learning for EHR

- Before deep learning: patient → feature engineering → model
- Examples:
  - counts of diagnoses
  - average labs
  - medication indicators
- Models:
  - logistic regression
  - random forests
  - gradient boosting

Big limitation: static representations

Sequence Models for EHRs

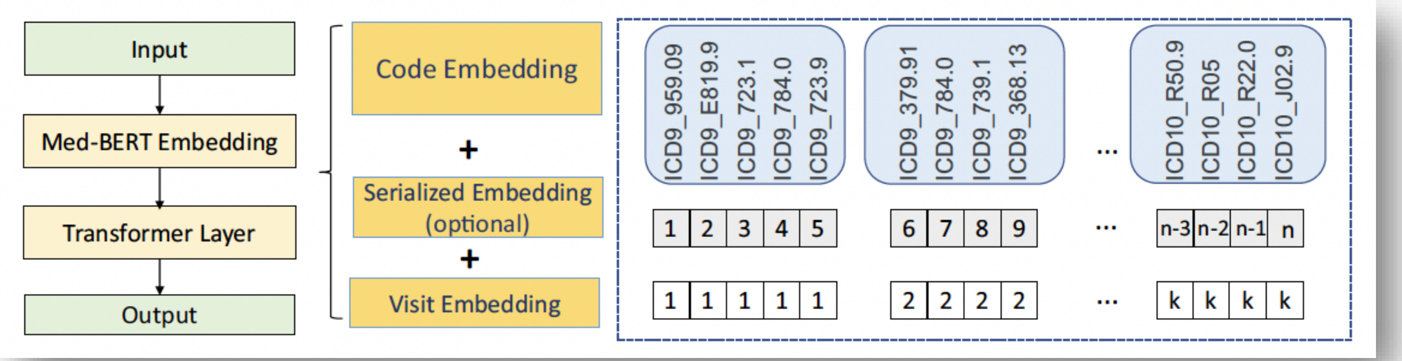
- Idea: Model patient history as a sequence of visits.
- Architecture: visit embeddings → RNN → prediction
- Examples:
  - DeepCare [2016]
    - LSTM
  - RETAIN [2017]
    - Key Idea: Attention-generation mechanism doesn't need to interpretable, but hidden state does



Limitation: bad at long histories, limited representation learning

Transformers for Patient Trajectories

- Idea: Treat medical events like tokens.
- Tokens → embeddings → transformer → patient rep.
- Examples:
  - MedBERT [2021]
  - ComET [2025]
    - Supervised via patient deterioriation over time



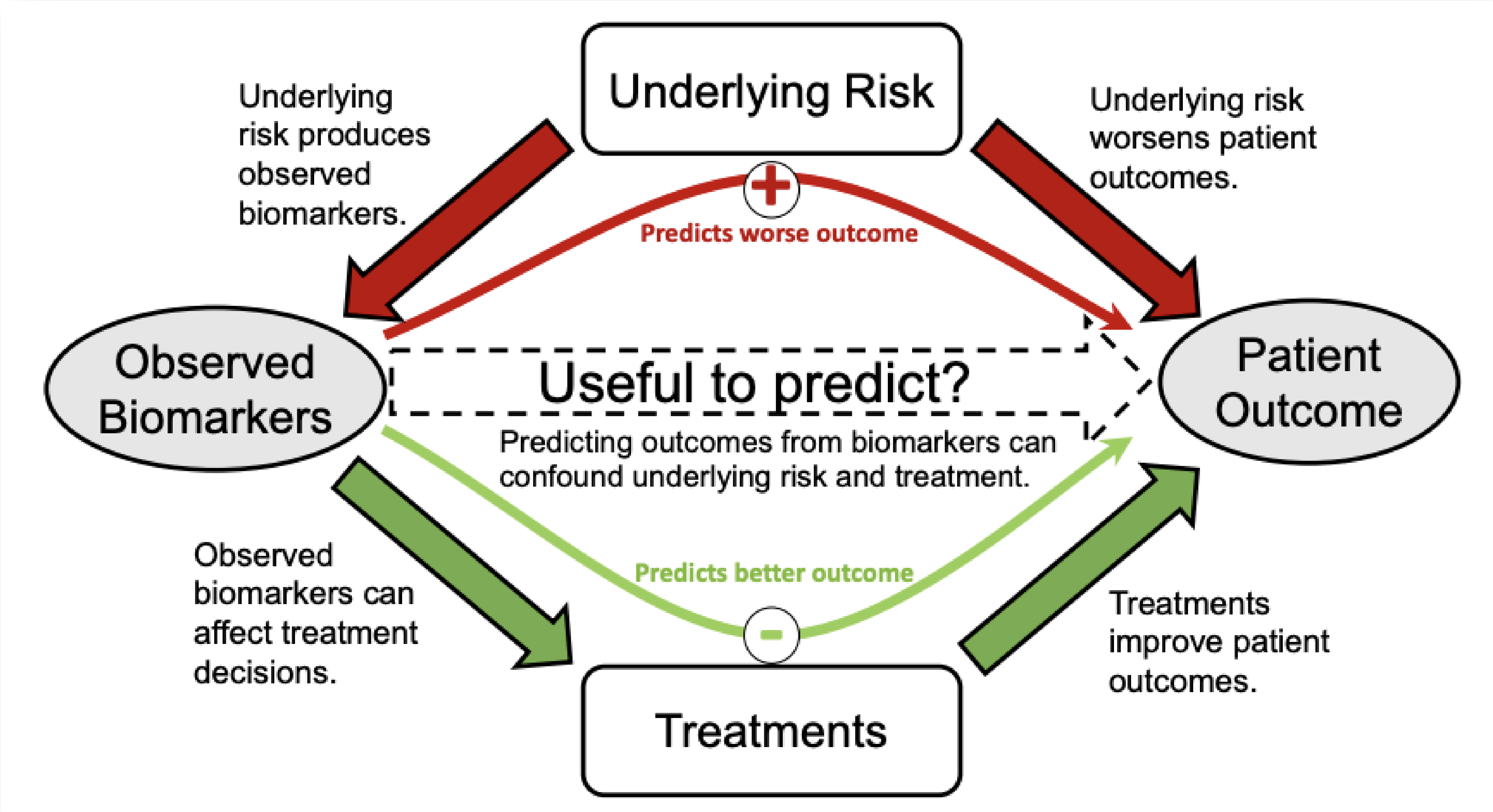
Should we do causal inference with foundation models?

What EHRs don’t capture

What’s not in EHRs?

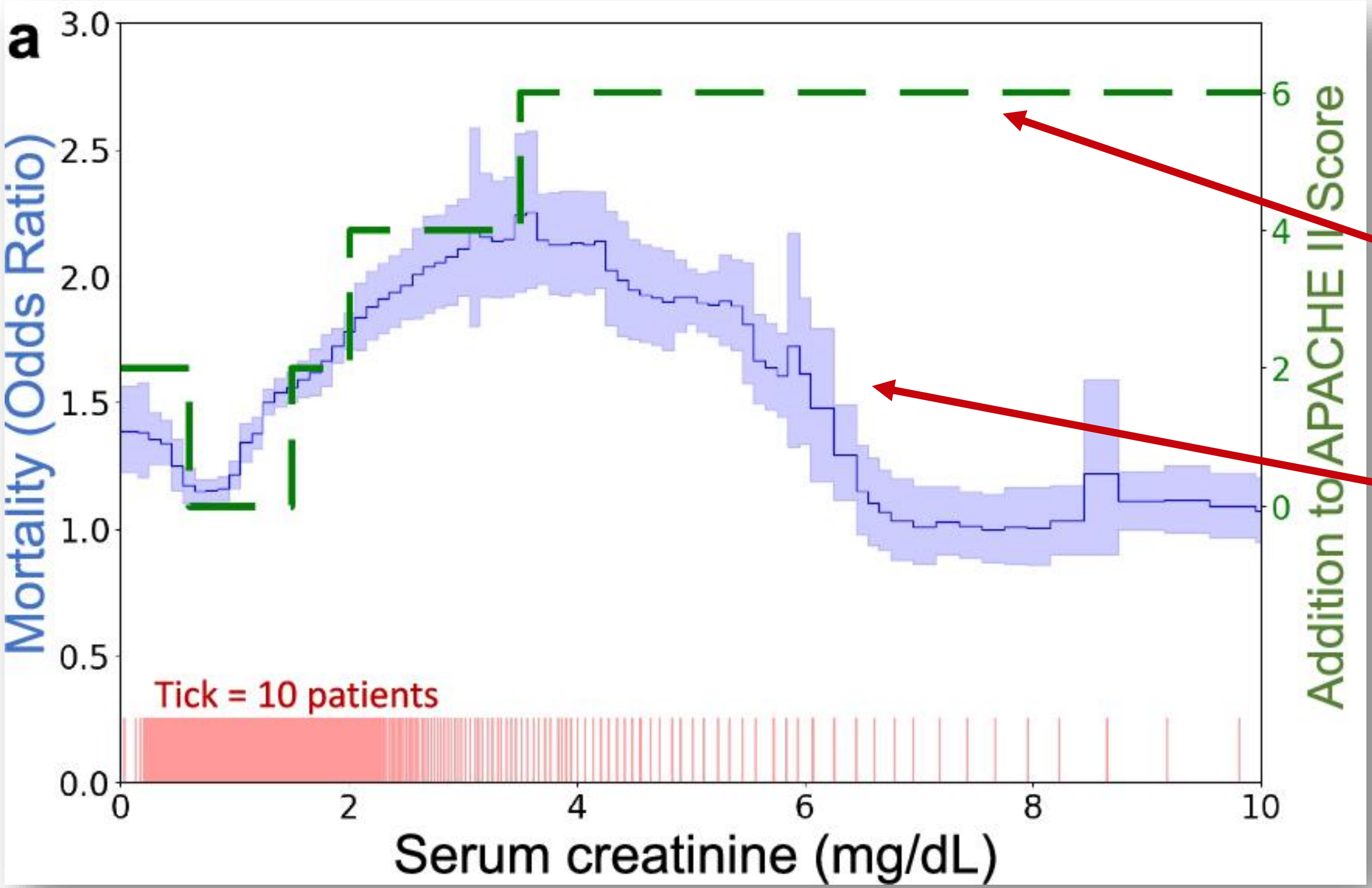
- Mental judgements: patient urgency, treatment urgency, etc.
- Socioeconomics: accessibility of treatment
- Reasons for measuring features: Why are we ordering a lab test now?
- Medical errors: Did we forget to replace the IV drip on time?
- Provider biases: Does every doctor make the same decisions?
- Billing biases: Is the purpose of a billing code diagnosis or payment?
- Small factors that add up: How much sleep did the patient get last night? When did the nurse turn them over in the bed last? Did the patient’s family give them a bottle of water and change our I/O calculations?
- **Randomness**: clinical decision rules/protocols

RWD comes from complex human behaviors



[Lengerich et al 2025]

Complexity + interventions → repeated confounding

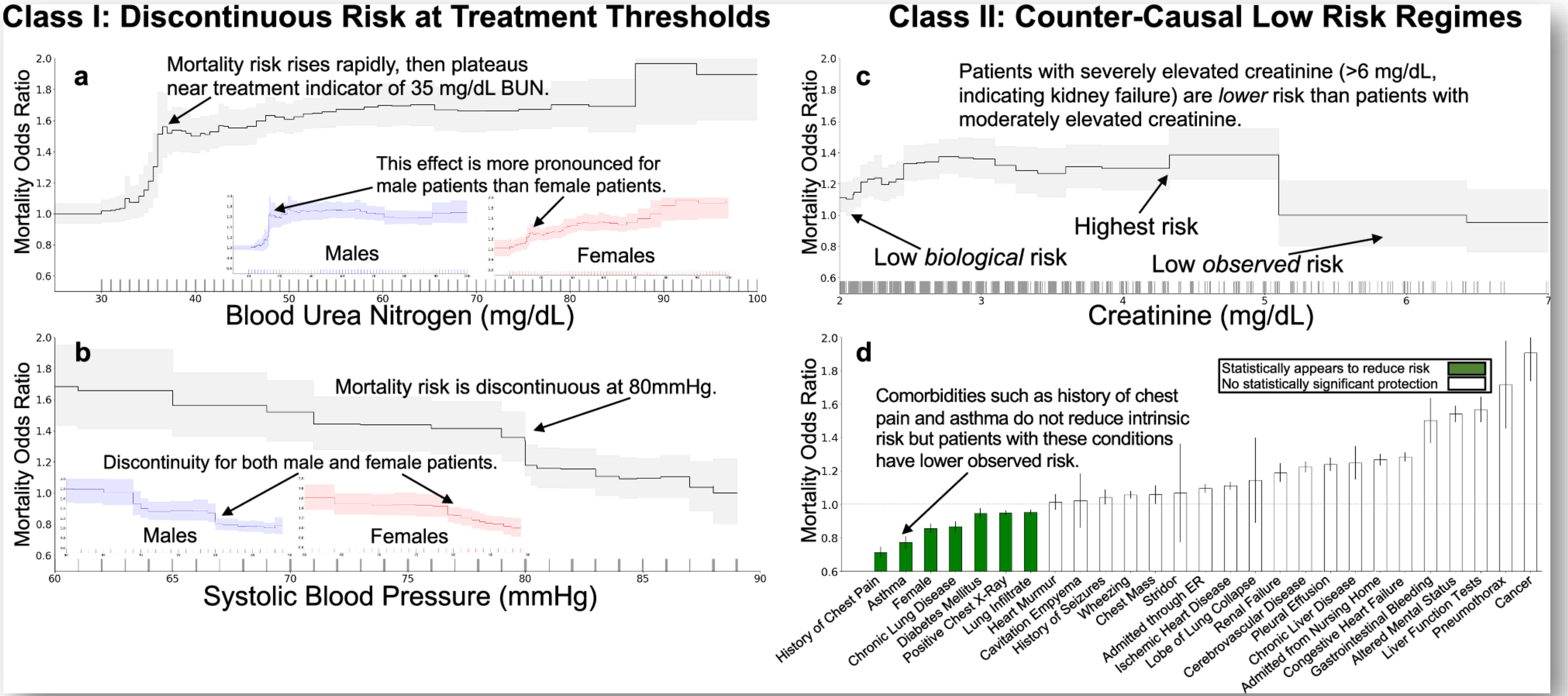


Biological risk goes up

but real-world risk goes down

[Lengerich et al 2025]

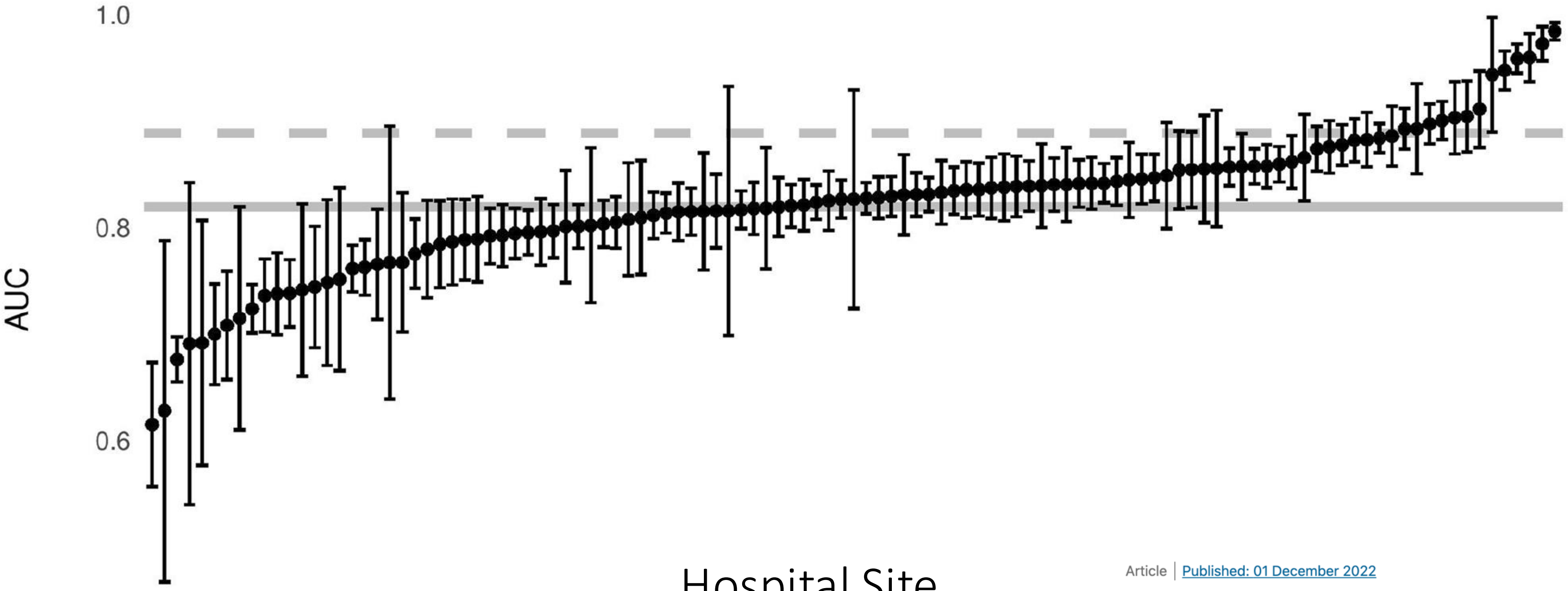
Complexity + interventions → repeated confounding



[Lengerich et al 2025]

Hidden causality → Failure to transport

- Logistic regression model of **acute kidney injury** (AKI)



Hospital Site

Article | Published: 01 December 2022

Generalizability of an acute kidney injury prediction model across health systems

Jie Cao, Xiaosong Zhang, Vahakn Shahinian, Huiying Yin, Diane Steffick, Rajiv Saran, Susan Crowley, Michael Mathis, Girish N. Nadkarni, Michael Heung & Karandeep Singh

Nature Machine Intelligence **4**, 1121–1129 (2022) | Cite this article

More famous failure to transport

- Sepsis: Life-threatening condition that occurs when the body's response to infection damages its own tissues.
  - Want system to alert clinicians when patients are at high risk of sepsis.
- Epic Sepsis Model (ESM): Logistic regression from 500k patient encounters
  - Reported performance: 0.76-0.83
  - External performance: 0.63 0.73 0.83

Michigan

June 21, 2021

External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients

Andrew Wong, MD[1]; Erkin Otles, MEng[2,3]; John P. Donnelly, PhD[4]; et al

Colorado

Accuracy of the Epic Sepsis Prediction Model in a Regional Health System

Tellen D. Bennett, MD, MS[1,2], Seth Russell, MS[2], James King, MIDS[2], Lisa Schilling, MD,MSPH[2,3], Chan Voong, MUSA[2], Nancy Rogers, BA,PMP[4], Bonnie Adrian, PhD,RN[4,5], Nicholas Bruce, PhD[6], Debashis Ghosh, PhD[2,7]

[1]Pediatric Critical Care, University of Colorado School of Medicine, Aurora, CO; [2]CU Data Science to Patient Value (D2V), Anschutz Medical Campus, Aurora, CO; [3]General Internal Medicine, University of Colorado School of Medicine, Aurora, CO; [4]Clinical Informatics, University of Colorado Health, Aurora, CO; [5]University of Colorado College of Nursing, Aurora, CO; [6]Epic Corporation, Verona, WI; [7]Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO;

South Carolina

Crit Care Explor. 2023 Jul; 5(7): e0941. Published online 2023 Jun 30.
doi: 10.1097/CCE.0000000000000941

PMCID: PMC10317482 | PMID: 37405252

Epic Sepsis Model Inpatient Predictive Analytic Tool: A Validation Study

John Cull, MD, Robert Brevetta, DO, Jeff Gerac, MD, Shanu Kothari, MD, and Dawn Blackhurst, DrPH

Open problems for EHR foundation models

- Generalization (Transport)
- Representation Learning
- Scaling Laws: Do larger medical models improve performance like LLMs?
- Federated training
- Multi-modal integration

The big picture: EHR foundation models

- Imaging invariances
  - physics
  - anatomy
- Genomics invariances
  - evolutionary constraints
  - regulatory grammar

Questions?

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON