Foundation Models for EHR Data

BIODS 271: Foundation Models for Healthcare

March 6, 2024

Jason Fries, PhD Research Scientist, Shah Lab
Center for Biomedical Informatics Research

Stanford MEDICINE | Clinical Excellence Research Center

About Me

Background

- Computer Science, PhD (University of Iowa)
- Stanford Postdoc (Scott Delp & Chris Re)
- Now Research Scientist (Nigam Shah)

Research Interests

- Methods for Building & Evaluating Foundation Models for Healthcare
- Data-Centric AI
- Human-AI Collaboration

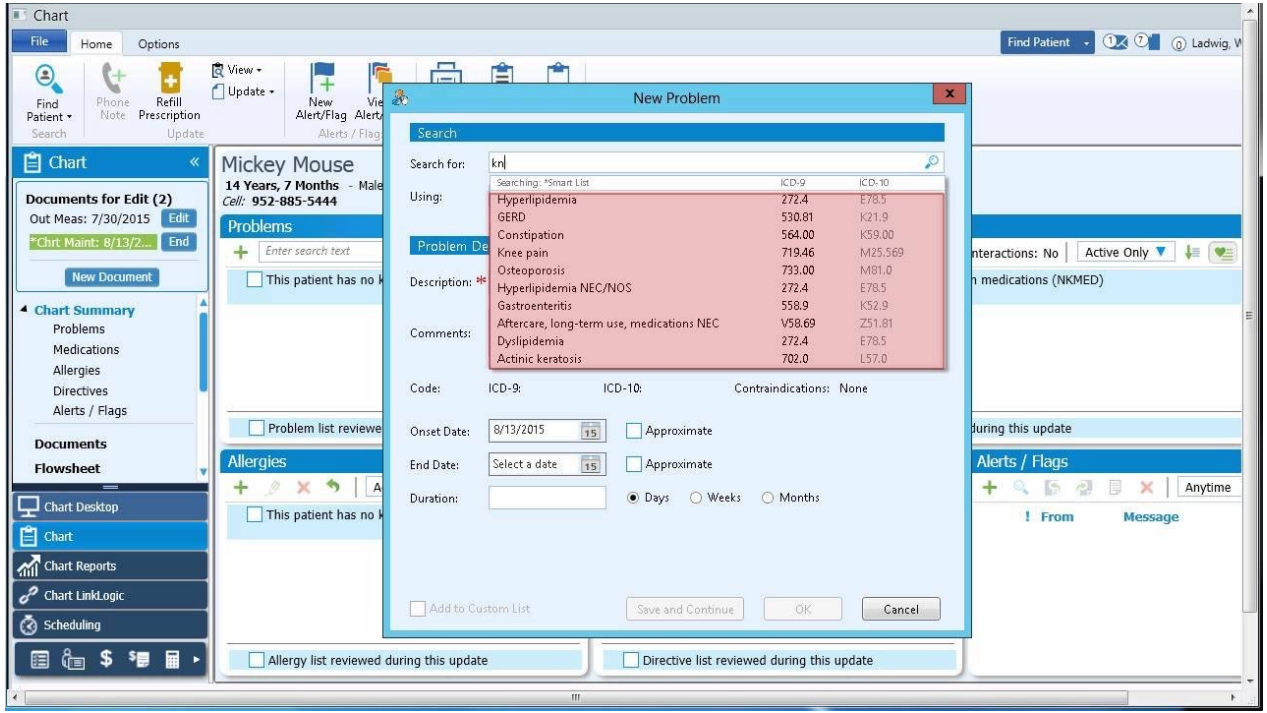
Outline

- **Overview: EHR Data & Tasks**
  - Electronic Health Record Data
  - Common Tasks
- **Modeling**
  - Problem Representation
  - Architectures & Pretraining Objectives
- **Evaluation**
- **Opportunities**

Thanks to Michael Wornow, Ethan Steinberg, Nigam Shah for slides!

Overview: EHR Data & Tasks

Electronic Health Records (EHR)



Healthcare Worker View

- GUI-based
- Data portal for a patients
- Focus on a single patient at a time

Epic Cerner

Common EHR Data Modalities

Labs | Vitals | Medication List

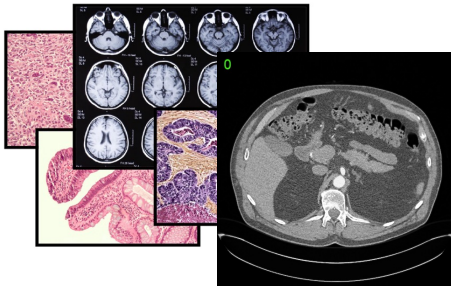
Notes | Past Medical History

Problem List | Social History

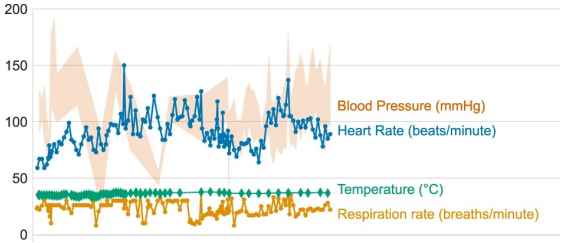
...

Care Plan | Treatment Plan

Tabular Data

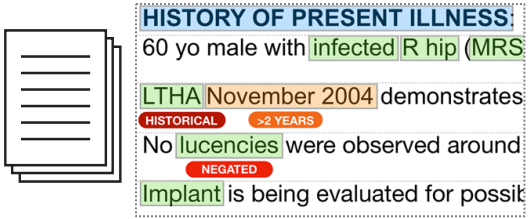


Imaging



Johnson et al. 2023. MIMIC-IV

Timeseries



Notes / Text

Audio /Conversations

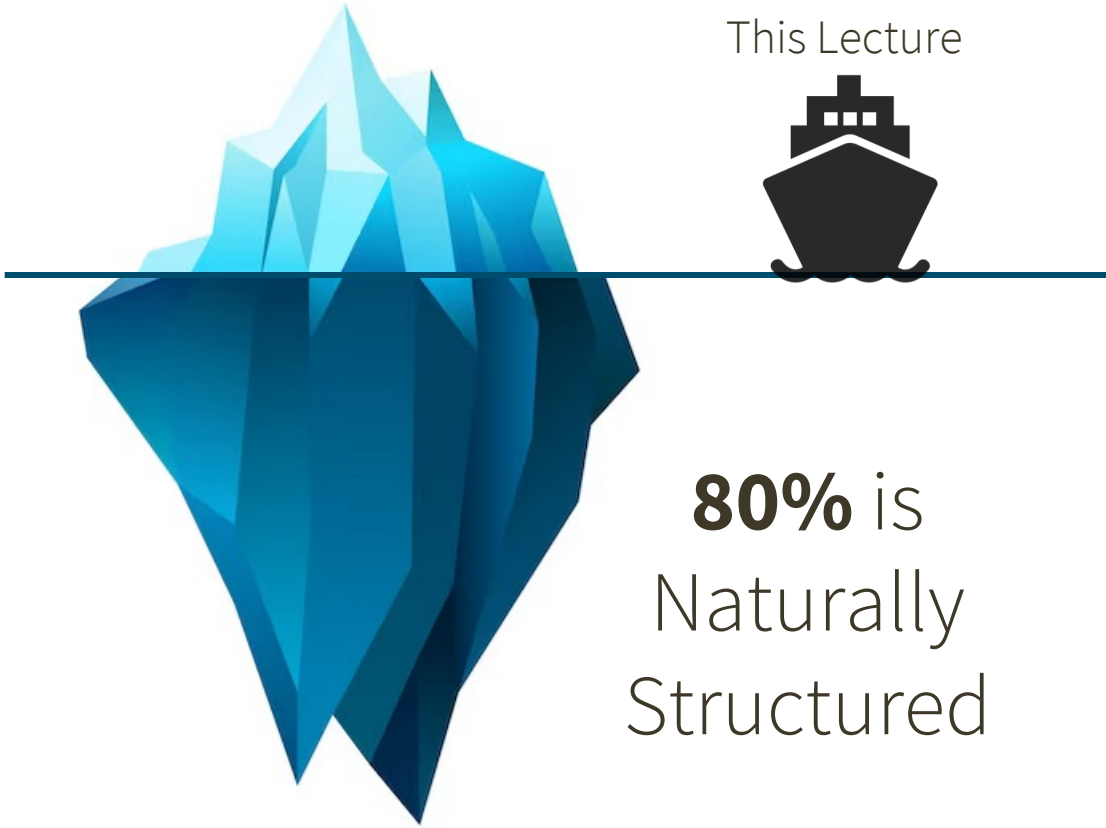
Video **Genomics**

...Near Future

STRUCTURED DATA

"NATURALLY STRUCTURED" DATA

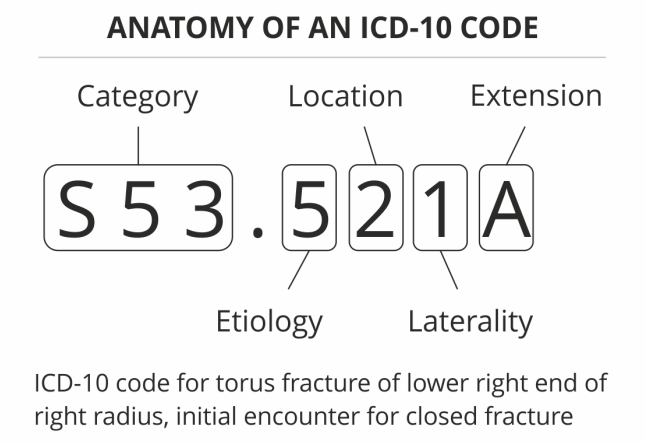
Common EHR Data Modalities



Hospitals Generate **~50 Petabytes** of Data **Per Year**

80% is Naturally Structured

Structured Data: Medical Vocabularies

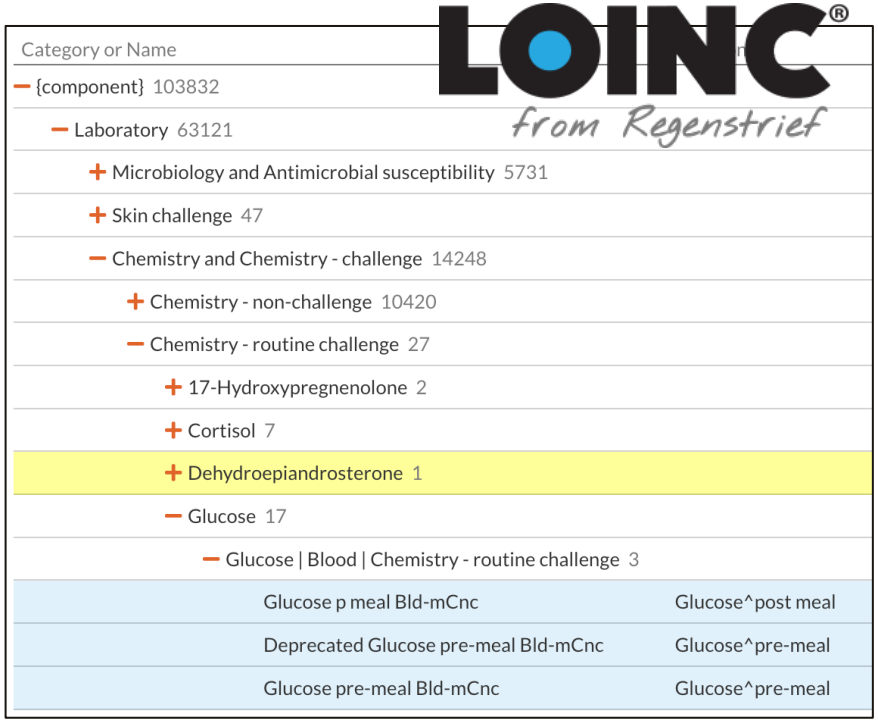


ICD-10 code for torus fracture of lower right end of right radius, initial encounter for closed fracture

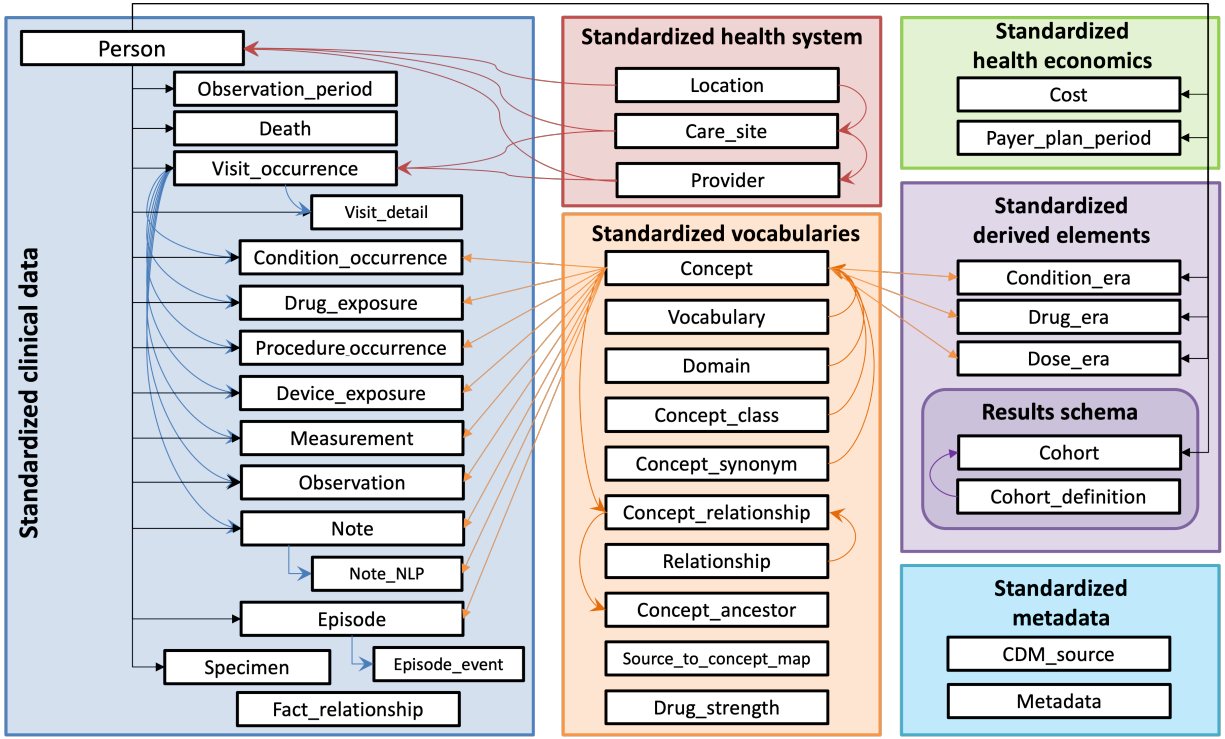
https://blogs.halodoc.io/

- Controlled Vocabularies
- **Knowledge Graphs**

$code_i \in$ Vocabulary



Electronic Health Records (EHR)

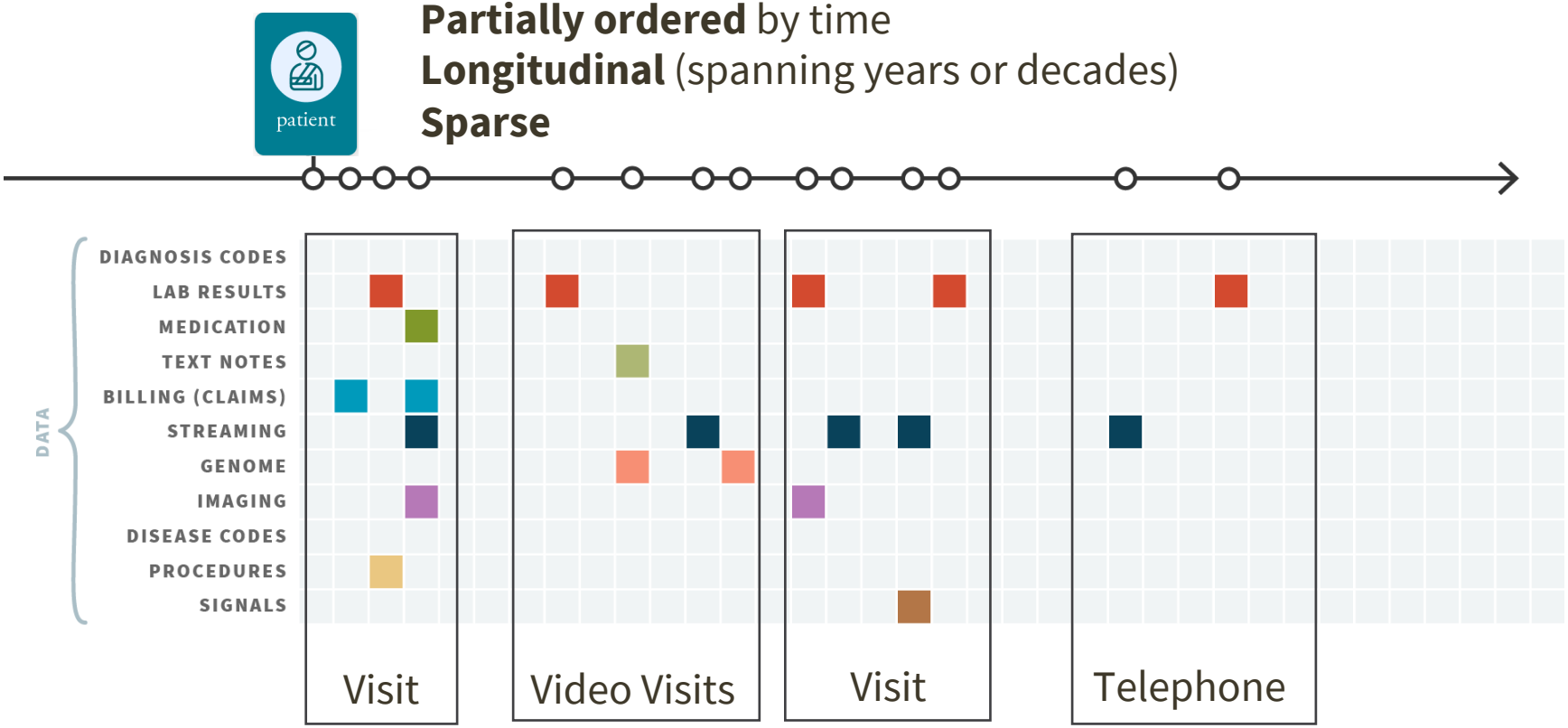


Data Scientist View

- Relational databases
- Some data model (Epic, OMOP, i2b2)
- Apply functions to all patients
- **Often transformed in un-inspectable ways**

Conceptualize EHR Data as a Multimodal Event Stream

Heterogenous (multiple types of data)
Partially ordered by time
Longitudinal (spanning years or decades)
Sparse



How Can AI for Healthcare?

Atherosclerotic cardiovascular disease risk assessment: An American Society for Preventive Cardiology clinical practice statement

Nathan D. Wong [a,*], Matthew J. Budoff [b], Keith Ferdinand [c], Ian M. Graham [d], Erin D. Michos [e], Tina Reddy [c], Michael D. Shapiro [f], Peter P. Toth [e,g]

nature medicine

Article

https://doi.org/10.1038/s41591-023-02332-5

A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories

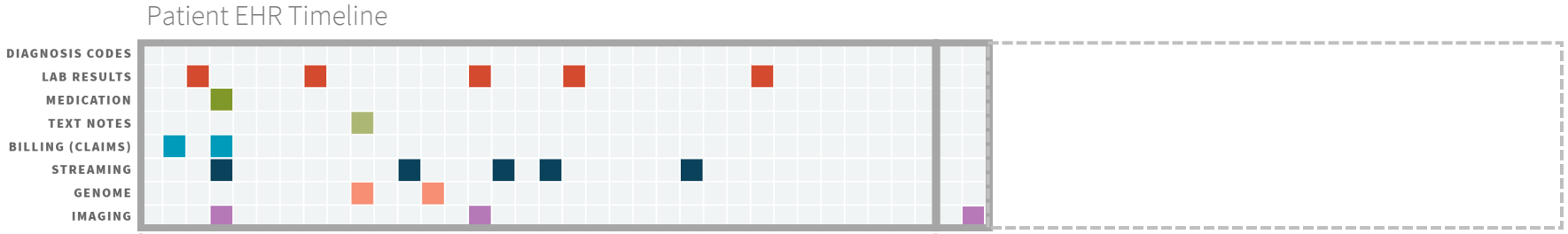
How Can AI for Healthcare?

- **Improved patient outcomes**
  - Treatment selection
  - Disease diagnosis (e.g. early detection of cancer)
  - Risk stratification (e.g. mortality, cancer progression)
  - Abnormal test result prediction (e.g. lab values)

- **More efficient hospital operations**
  - Predictions for quality metrics (e.g. 30-day readmission likelihood)
  - Resource allocation (e.g. anticipating ICU transfers)
  - Billing (e.g. identify mis-coding of patient records)

- **Research**
  - Causal inference (e.g. drug trials and observational studies)
  - Identify off-label drug benefits

AI to Enhance Medical Decision Making



Example ML Applications

What Occurred in the Past?

- Chart Summarization
- Cohort Construction
- Training Data Construction

What is Occurring Now?

- Identify blood clots in lung CT scans
- Identify cancerous cells in pathology slides

Predict Future Risks & Intervention Benefits

- Will patient develop nephritis?
- Will patient develop chronic pulmonary hypertension?

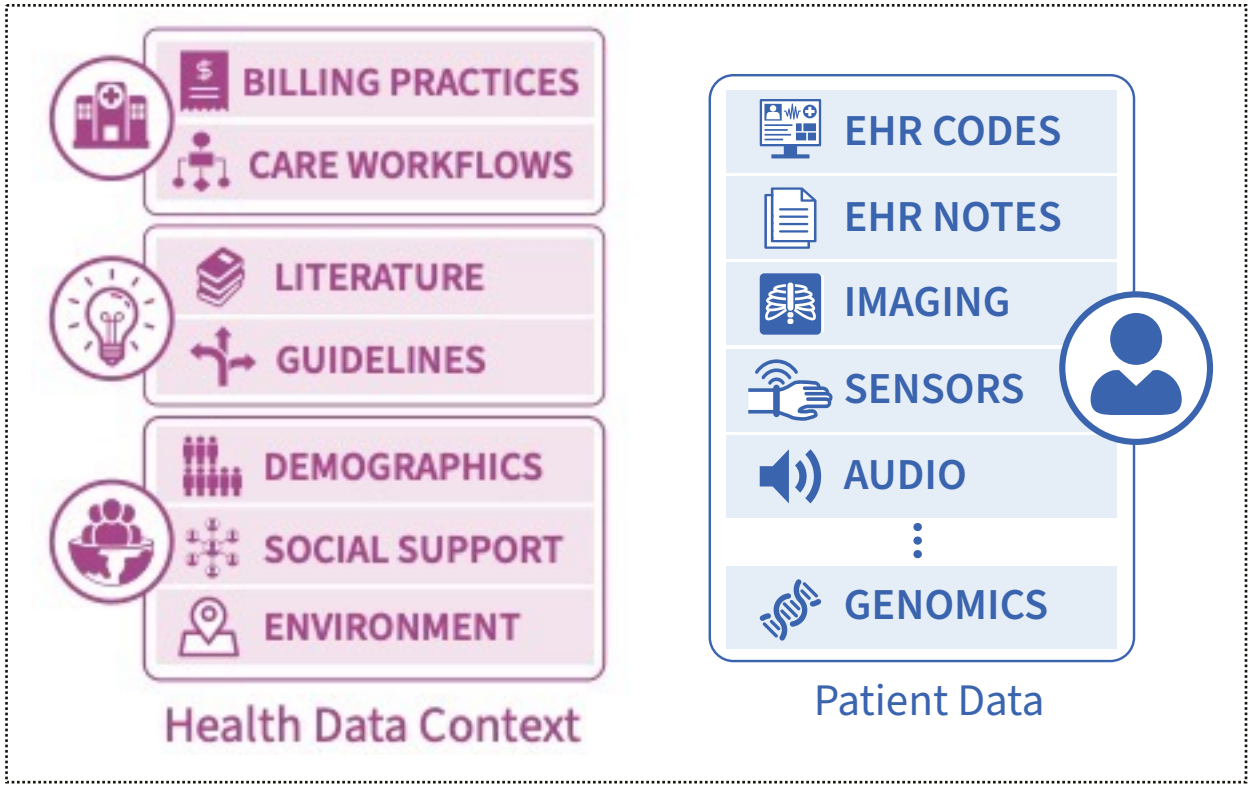
Whether to Treat | **How to Treat**

subject to

Policy | Capacity to Act

Intervention Properties

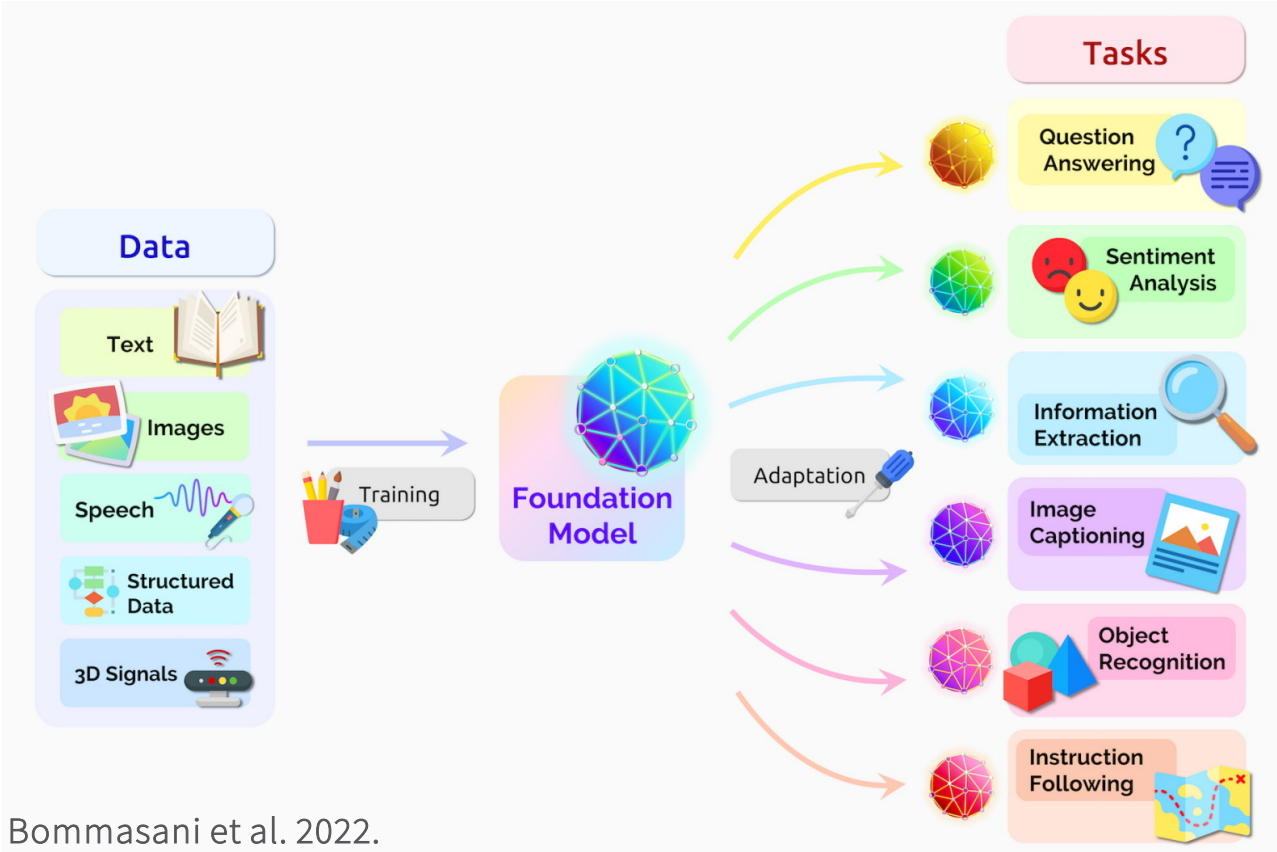
EHR Data Generation Context



***Inspectability* into the data generation process** is key to mitigating “data cascades” in AI *

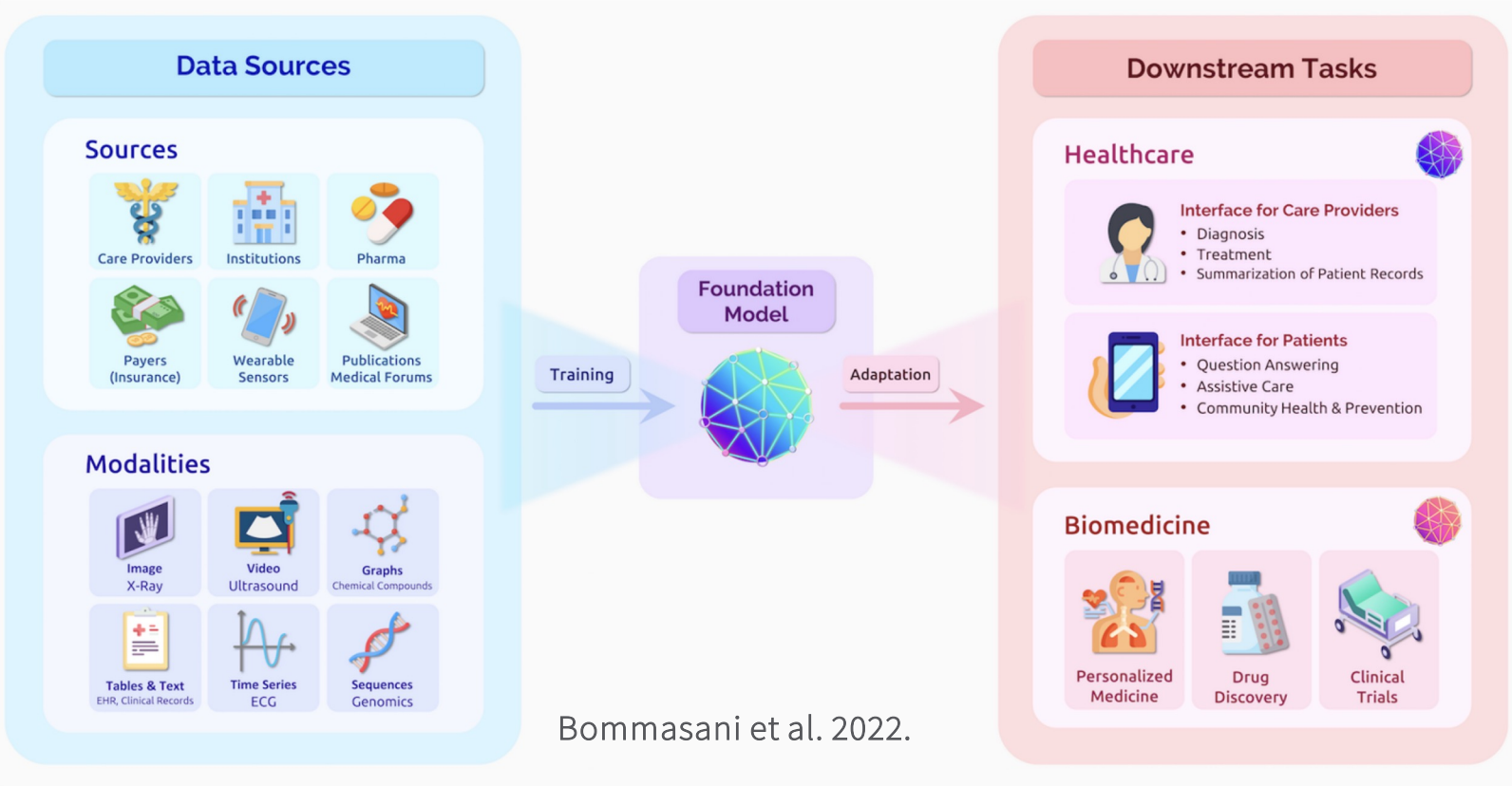
* "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI (Sambasivan et al. 2021)

Foundation Models and AI's "Industrial Age"



Bommasani et al. 2022.

Foundation Models and AI's "Industrial Age"



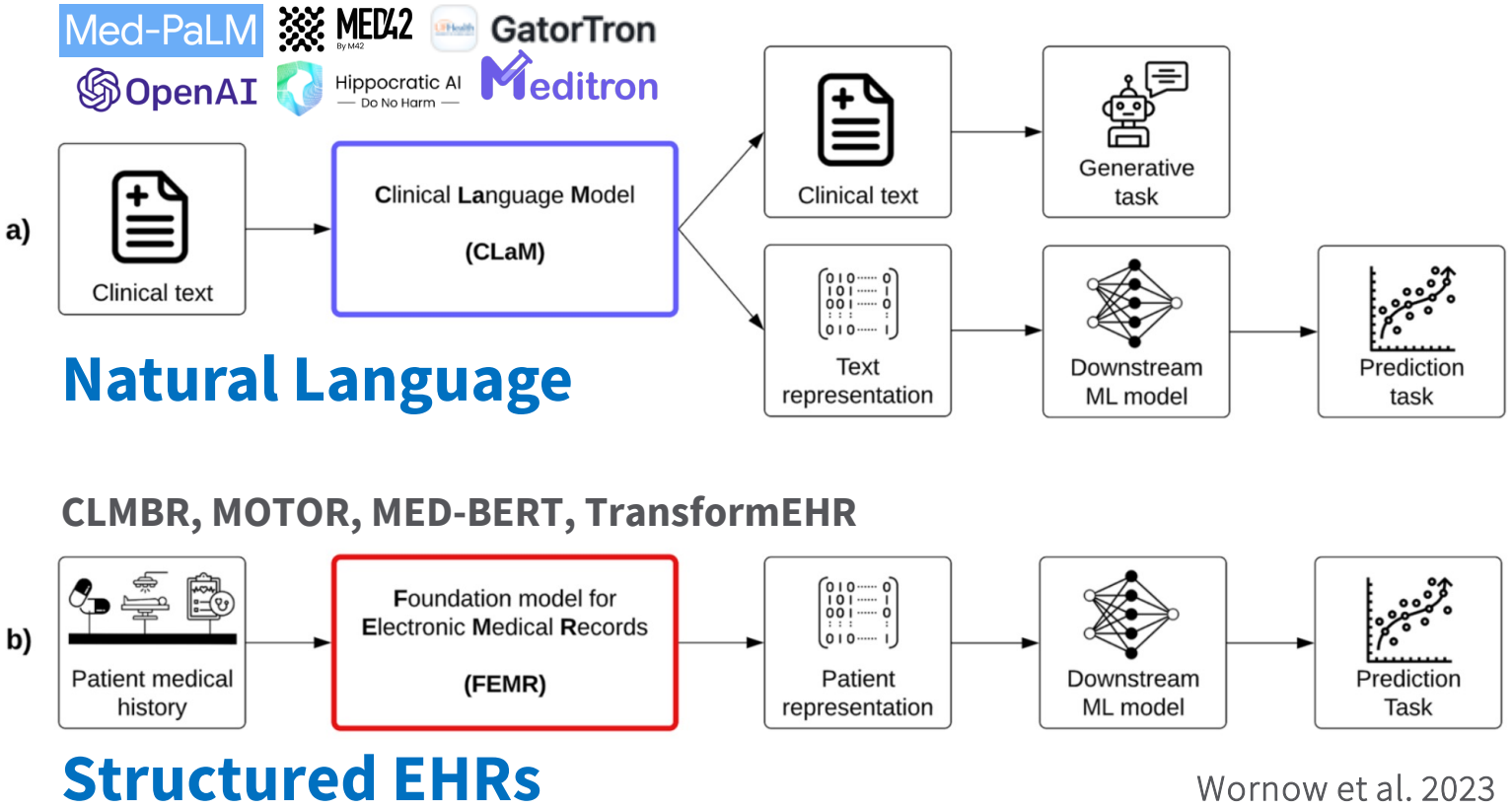
Bommasani et al. 2022.

Modeling: Architectures & Pretraining

Outline

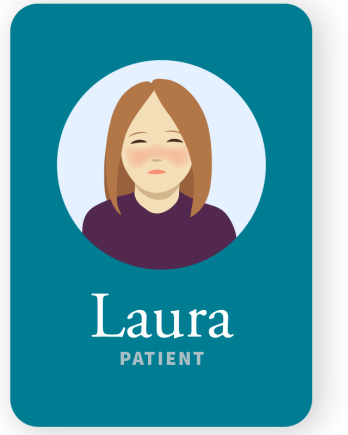
- **Classes of EHR Foundation Models**
- **Transfer Learning**
  - Representational Choices
- **Self-Supervised Pretraining**
  - Masked Language Modeling
  - GPT/Autoregressive
  - Time-to-Event

Two Classes of EHR Foundation Models



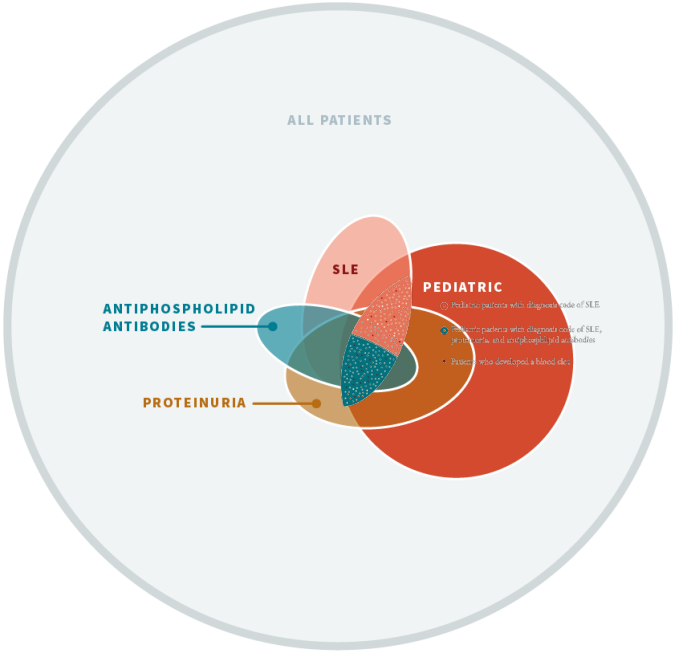
Wornow et al. 2023

Classic Approach to Building and Patient Model

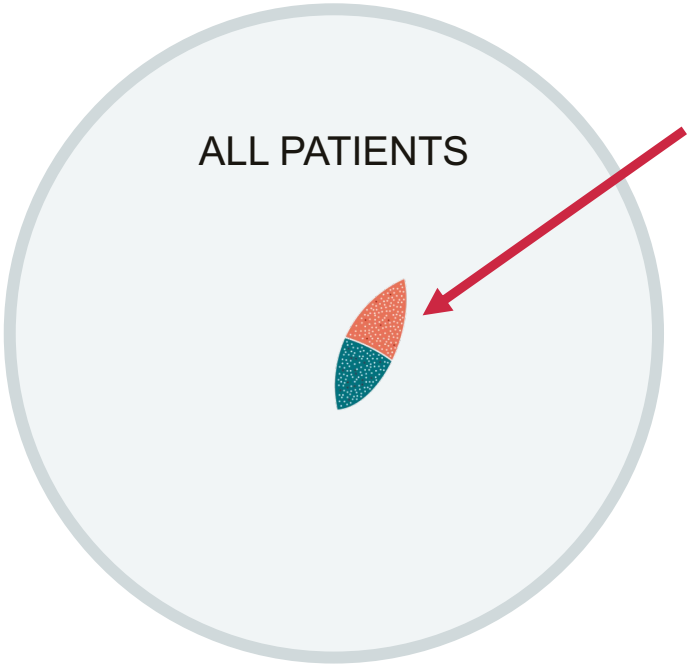


MEET LAURA

A teenager with systemic lupus erythematosus (SLE), proteinuria, pancreatitis and positive for antiphospholipid antibodies



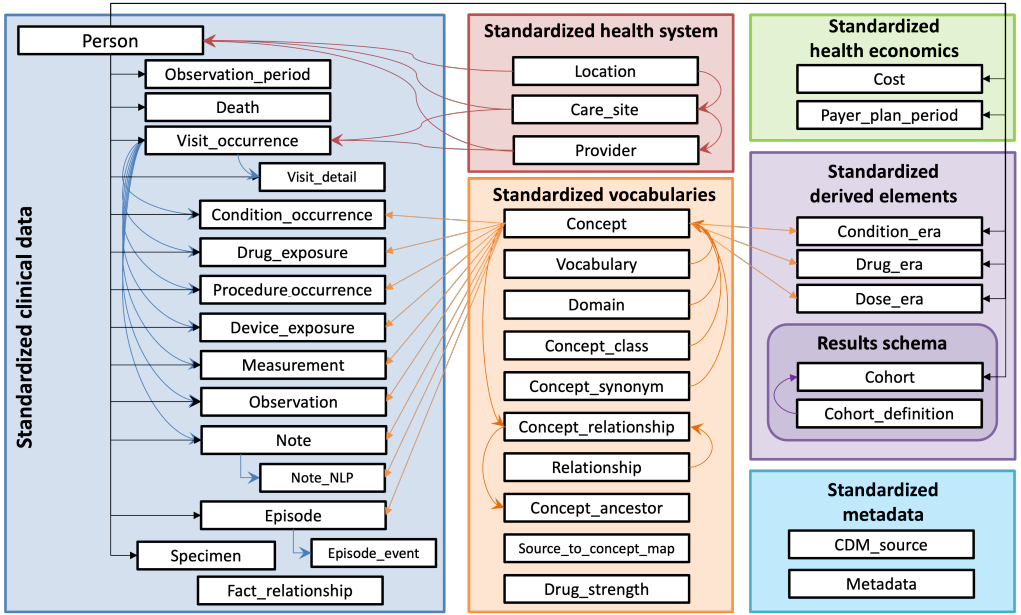
Classic Approaches Often Fail Due to Limited Data



Classic medical ML approach is to train on a small **target cohort**

…but this doesn't take advantage of the structure present in the entire patient population

Initial Challenges in Transfer Learning with EHR



Relational/tabular data **has been difficult for transfer learning**

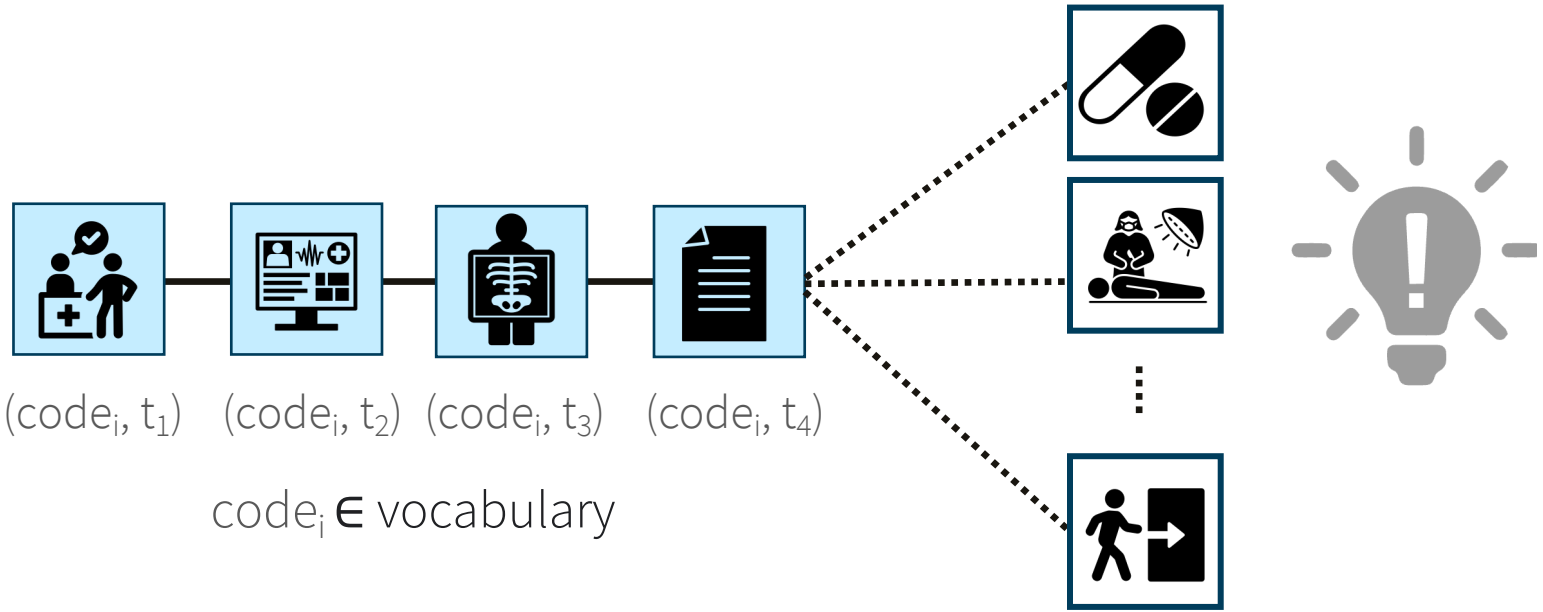
Insufficient Scale of Training Data

~195k Patients

(Rajkomar et al. 2018)

Model EHR Data as Sequences (*Event Streams**)

Hypothesis: Accurately ***generating*** future health states captures many use cases of medical AI

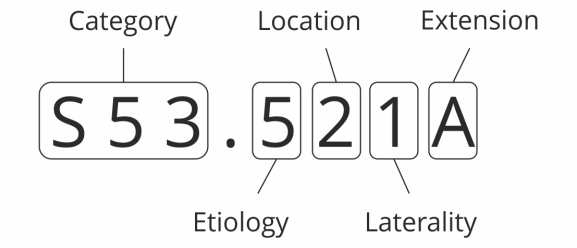


* McDermott et al. 2023.

More Like NLP Now, but Key Differences!

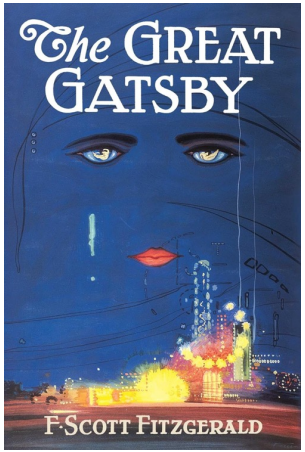
Tokenization / Vocabulary

	NLP	EHR
Vocabulary Size	50k	**250k+**
Subwords	Yes	**No**
Tokens Semantics	Flat	***Hierarchical, Complex Dependencies***



Sequence Properties

	NLP	EHR
Vocabulary Size	32k	**250k+**
Ordering	Total	***Partial***
Time Intervals	None	***Discontinuous***
Sampling Fidelity	All	***Sparse/Errors***



50% Patients
>= 68k tokens

Self-Supervised Pretraining Objectives

BERT-Style (Masked Language Modeling)

BEHRT (Li et al. 2020)
MedBERT (Rasmy et al. 2021)
ClaimPT (Zeng et al. 2022)

GPT-Style (Autoregressive)

CLMBR (Steinberg et al. 2020)
TransformEHR (Yang et al. 2023)

Time-to-Event

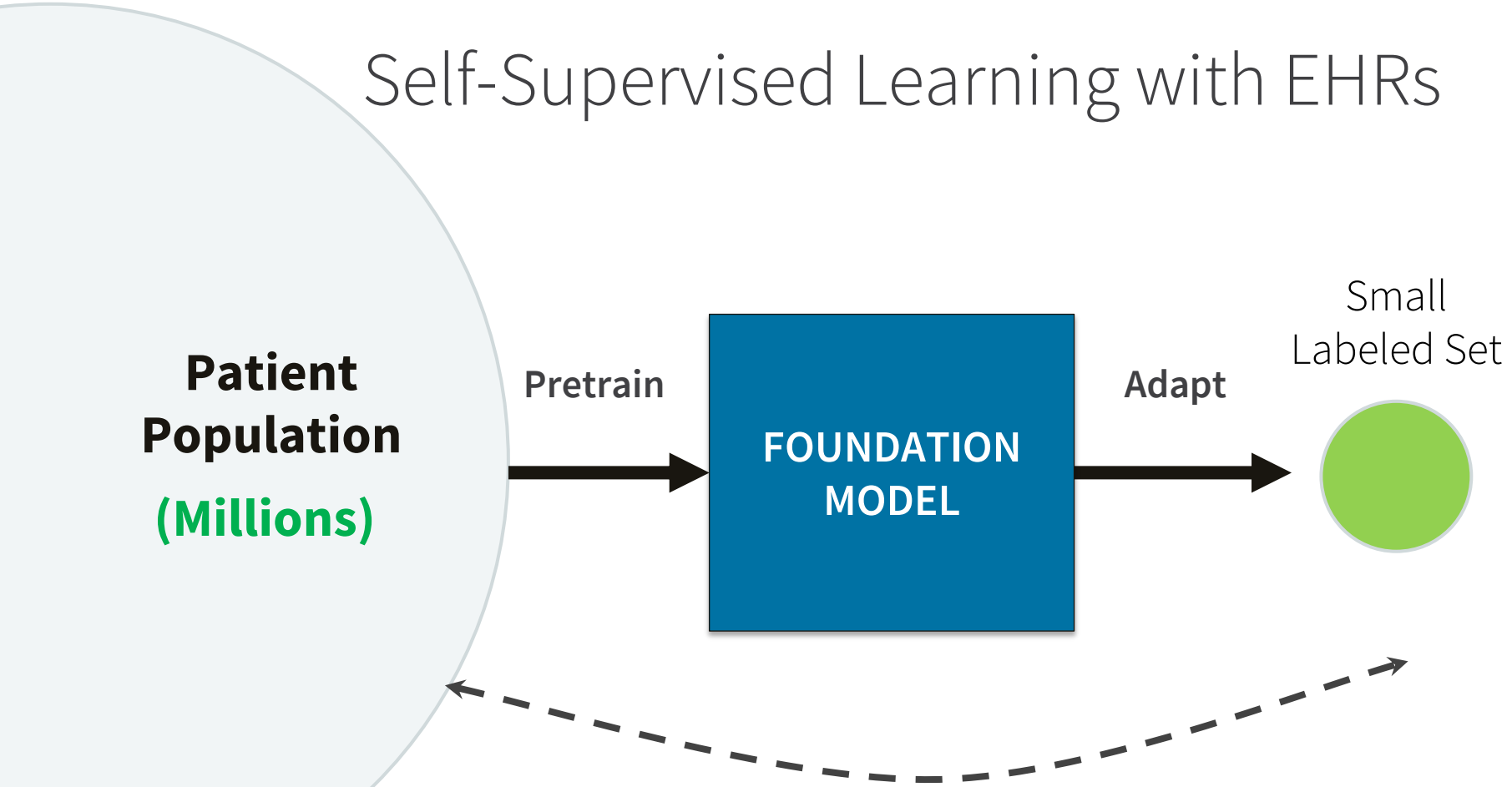
MOTOR (Steinberg et al. 2024)

Graph-Based

GraphCare (Jiang et al. 2024)

Won't talk about these

Self-Supervised Learning with EHRs



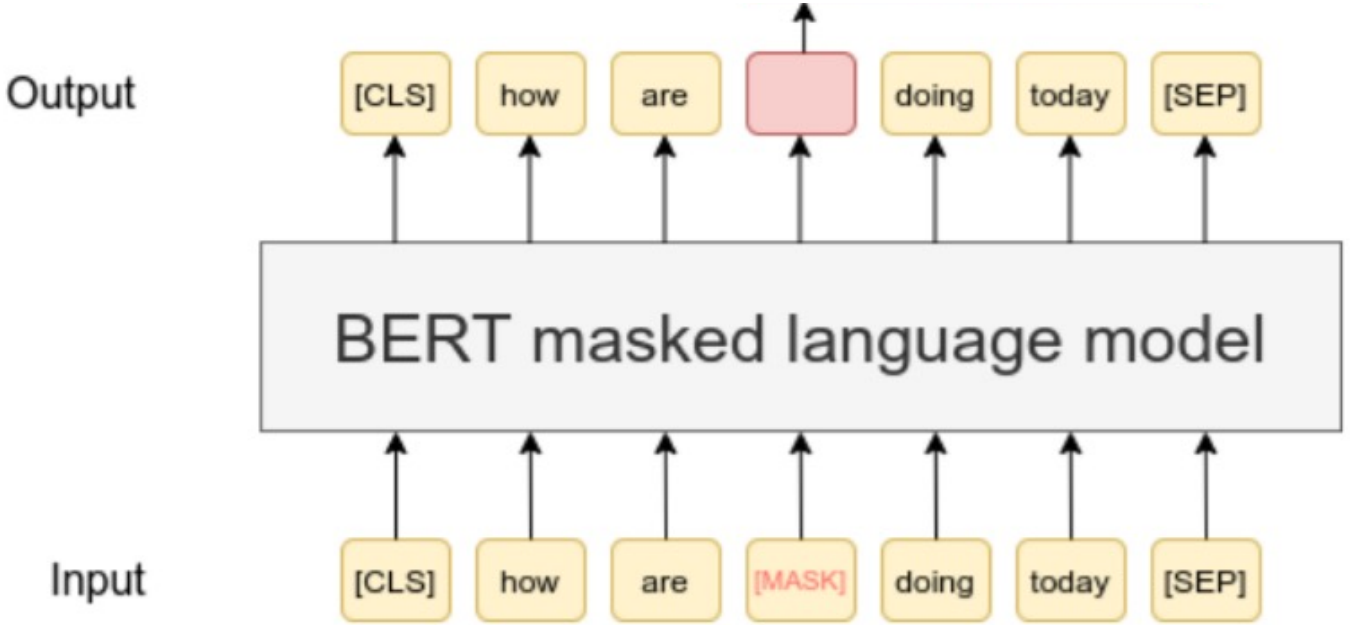
Transfer Learning: *Assumes Shared Structure*

BERT-Style (Masked Language Modeling)

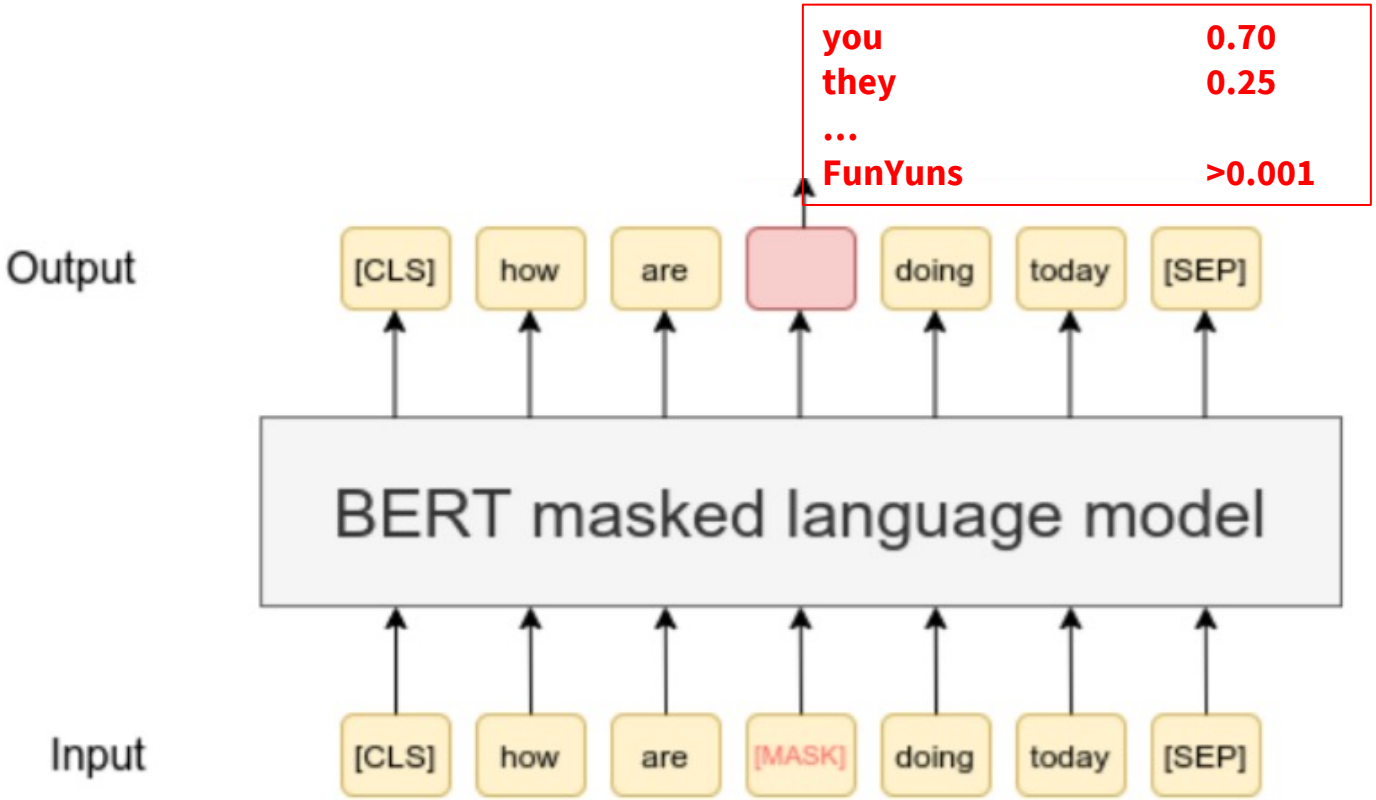
BEHRT (Li et al. 2020)
MedBERT (Rasmy et al. 2021)
ClaimPT (Zeng et al. 2022)

Corruption-based (Masking) Pretraining Objective

- **Mask tokens (15%)**
- **Train Model to Predict [MASK]'ed tokens**



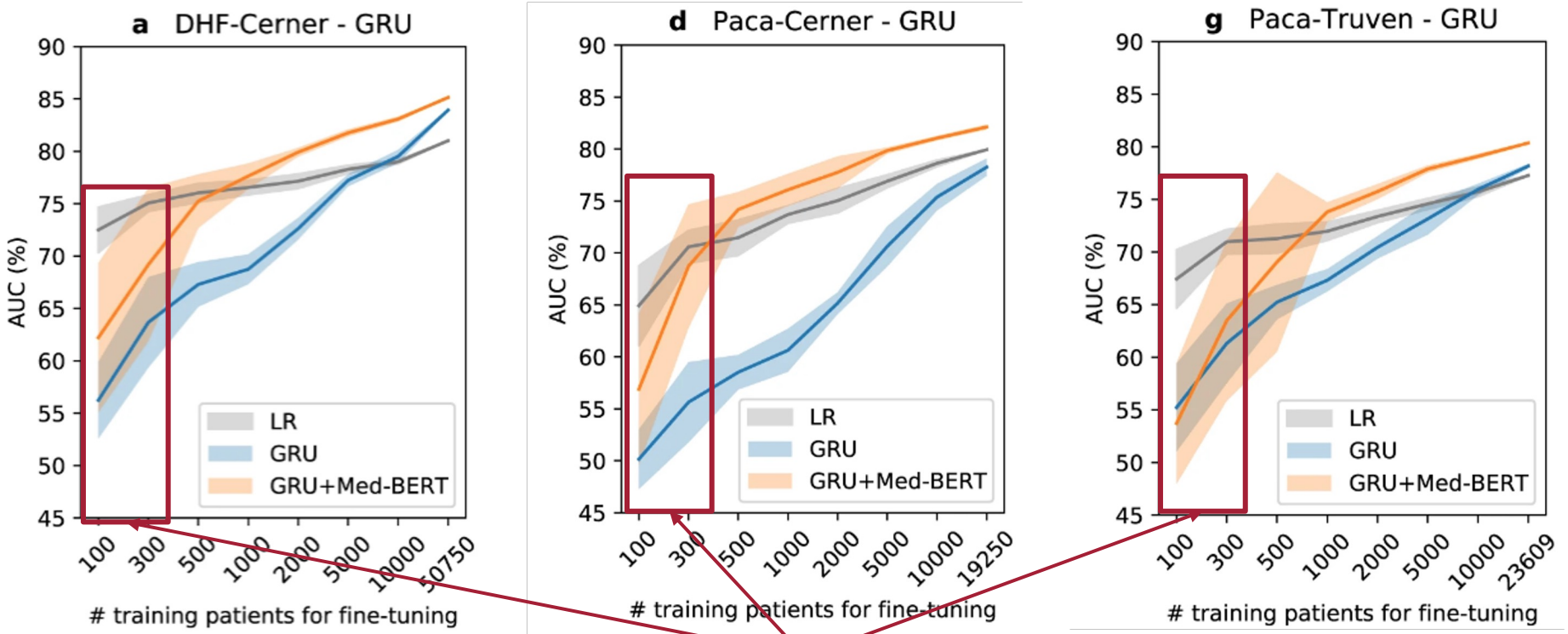
Corruption-based (Masking) Pretraining Objective



BERT-based Architecture (BEHRT)

Li et al. 2020

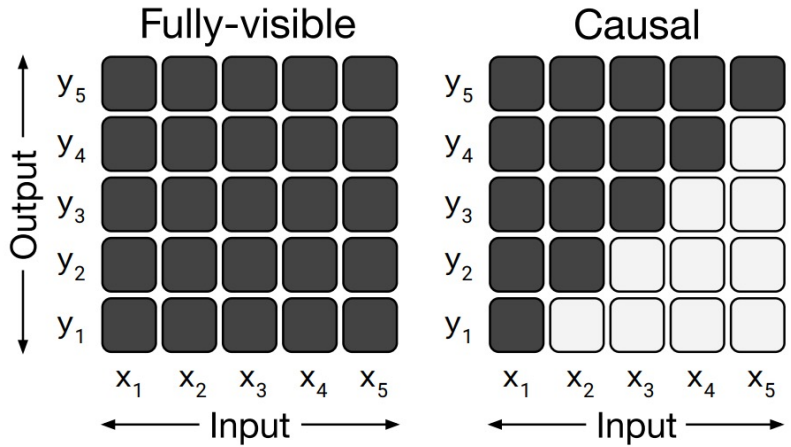
Better performance than baselines (MedBERT)



But few-shot performance isn't great…

Rasmy et al. 2021

Other Disadvantages



Raffel et al. 2019

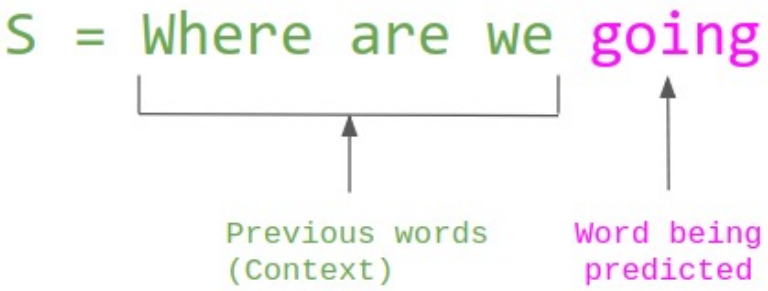
Masked Language Modeling uses **bidirectional attention**. Good for summarizing a sequence, but **not generating the next event/token**

Rasmy et al. 2021

GPT-Style (Autoregressive)

CLMBR (Steinberg et al. 2020)
TransformEHR (Yang et al. 2023)

Self-Supervised Pretraining in Natural Language

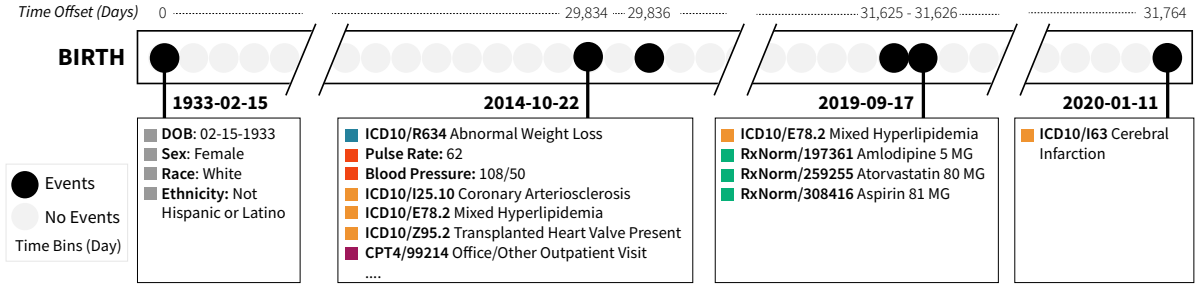


P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

$$P_{(w_1,w_2,\ldots,w_n)} = p(w_1)p(w_2|w_1)p(w_3|w_1,w_2)...p(w_n|w_1,w_2,..,w_{n-1})$$

$$= \prod_{i=1}^{n} p(w_i|w_1,...,w_{i-1})$$

Self-Supervised Pretraining in Event Streams



S = ICD10/R634 ICD10/E78.2 … CPT4/99214 **ICD10/I63**

Patient History

Code being predicted

$code_i \in$ vocabulary

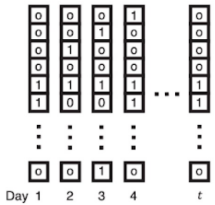
$$\mathbf{P(S) = P(c_1) \times P(c_2 \mid c_1) \times \ldots \times P(c_N \mid c_{<N-1})}$$

$$P(S) = \prod_{i=1}^{N} P(c_i \mid c_{<i})$$

CLMBR: Autoregressive Generative Pretraining

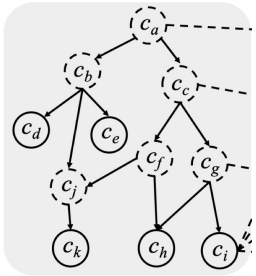
Key Intuitions: Treat codes as words in a symbol vocabulary and use knowledge graphs to better model rare codes

Patient Timeline



Days are represented as a set of <= M codes

Knowledge Graph



Represent codes as paths to the root of a medical ontology

Decoder-only Model

RNN or Transformer Backbone

Autoregressive Objective

$$d = \{c_i\}_{i=1}^{m} \qquad X = (d_1, ..., d_t)$$

Days are a set of unordered codes

Patients are ordered sequences of days

$$P(d_i|d_{<i}) = \prod_{c_j \in d_i} P(c_j|d_{<i}) \prod_{c_j \notin d_i} (1 - P(c_j|d_{<i}))$$

Model days by assuming codes are independent
Use hierarchical info to improve speed & estimation

P(Patient) $$P(X) = \prod_{i}^{t} P(d_i|d_{<i})$$

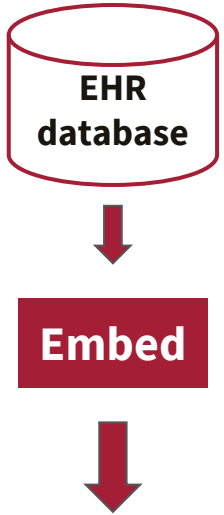
GPT-based Architecture (CLMBR)

2.5 million records

Steinberg et al. 2021

GPT-based Architecture (CLMBR)

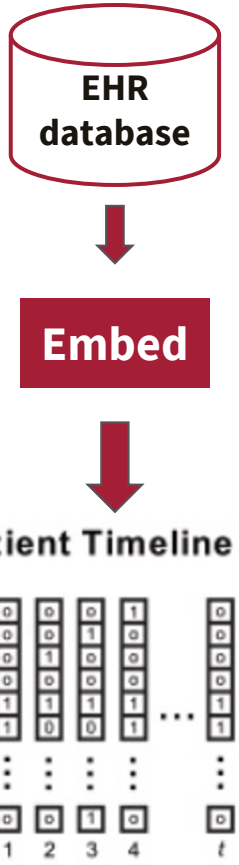
2.5 million records



Steinberg et al. 2021

GPT-based Architecture (CLMBR)

2.5 million records

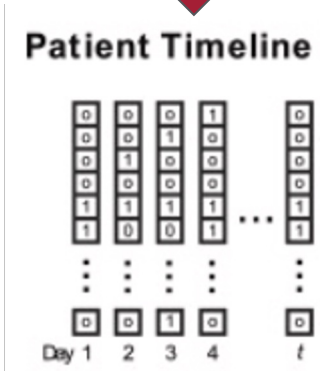


Steinberg et al. 2021

GPT-based Architecture (CLMBR)

2.5 million records

Embed



CLMBR

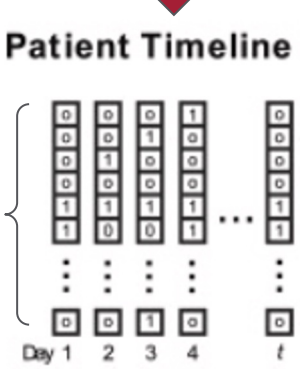
Learning objective:
Next code prediction

Steinberg et al. 2021

GPT-based Architecture (CLMBR)

2.5 million records

Embed



Patient: ICD Code A, ICD Code B, ICD Code C

CLMBR

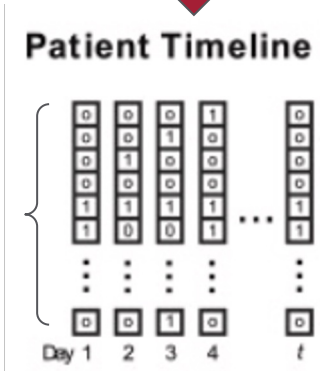
Learning objective:
Next code prediction

Steinberg et al. 2021

GPT-based Architecture (CLMBR)

2.5 million records

Embed



Patient: ICD Code A, ICD Code B, [redacted]

CLMBR

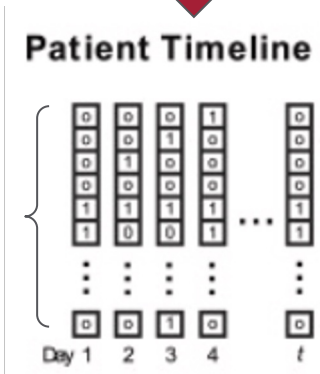
Learning objective:
Next code prediction

Steinberg et al. 2021

GPT-based Architecture (CLMBR)

2.5 million records

Embed



Patient: ICD Code A, ICD Code B [redacted]

What code comes next?

CLMBR: ICD Code C

CLMBR

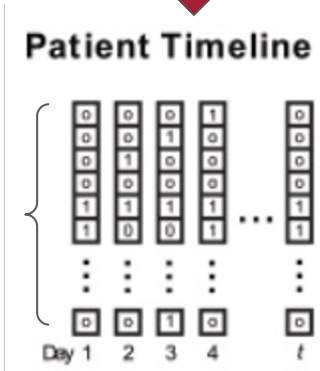
Learning objective:
Next code prediction

Steinberg et al. 2021

GPT-based Architecture (CLMBR)

2.5 million records

Embed



Patient: ICD Code A, ICD Code B, **ICD Code C**

What code comes next?

CLMBR: **ICD Code C**

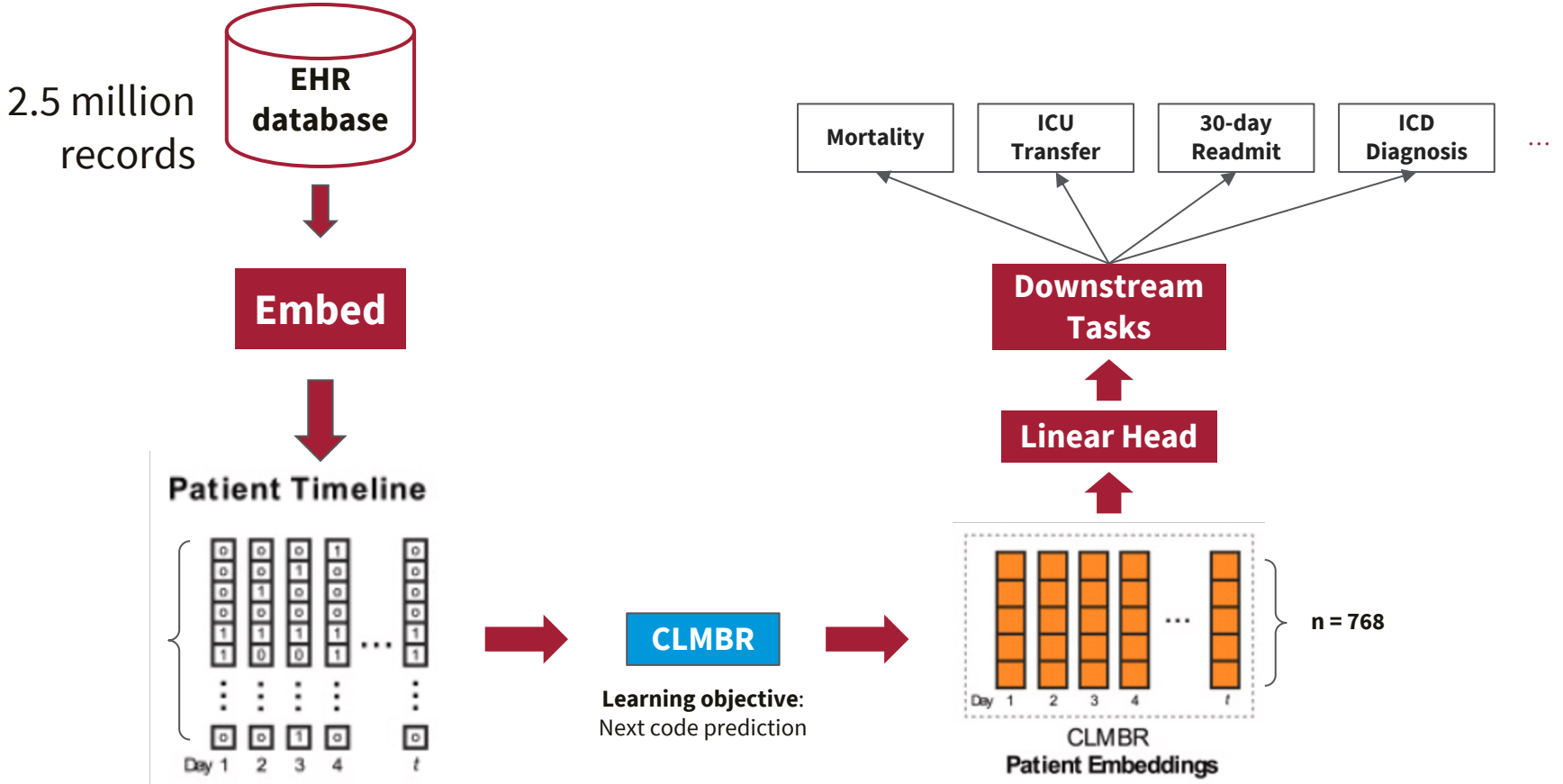
Correct!

CLMBR

Learning objective:
Next code prediction

Steinberg et al. 2021

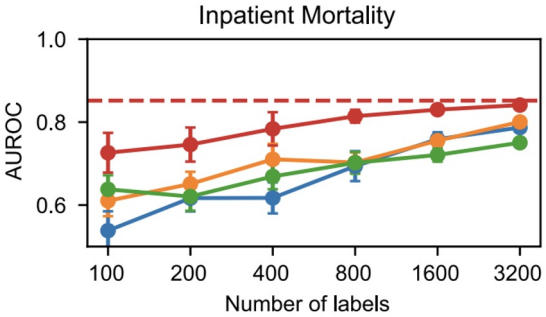
GPT-based Architecture (CLMBR)

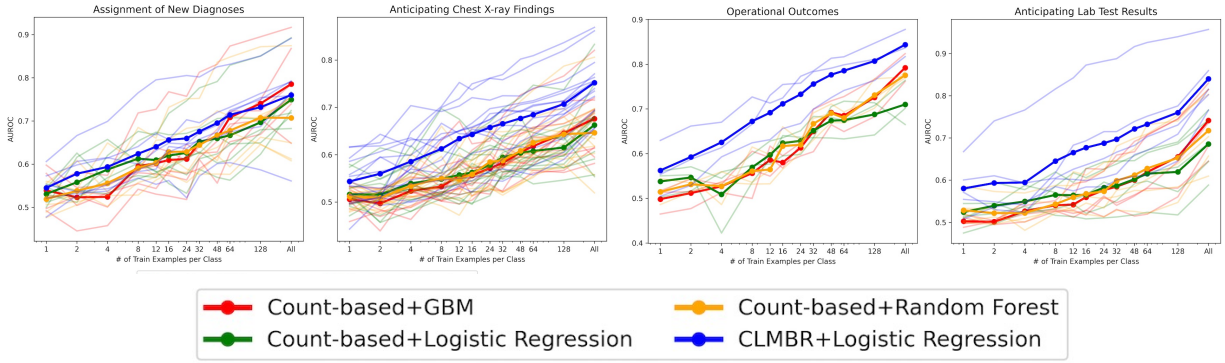


Steinberg et al. 2021

Benefits of Autoregressive EHR Foundation Models

1. Improved Label & Sample Efficiency





+3.5 to 19% AUROC

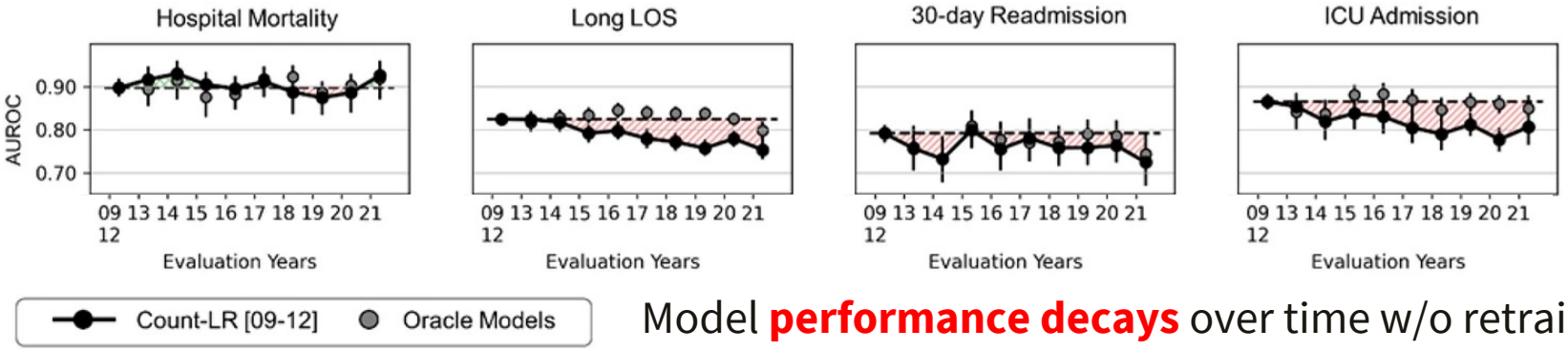
(Steinberg et al. 2020)

Improved Few-Shot Adaptation

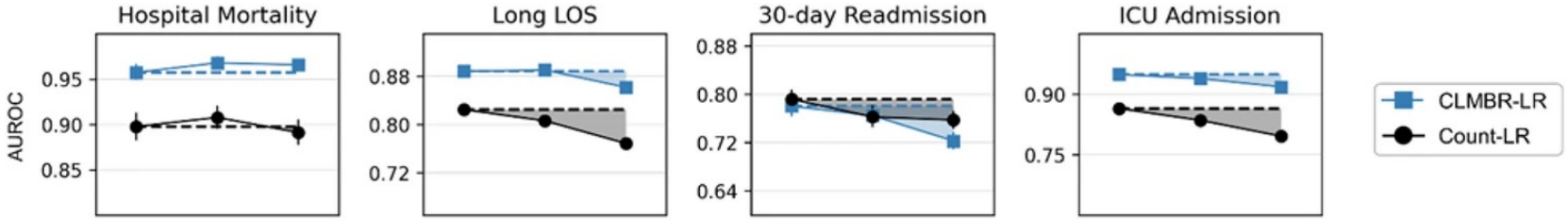
(Wornow et al. 2023)

Benefits of Autoregressive EHR Foundation Models

2. Improved Distributional Robustness



Model **performance decays** over time w/o retraining

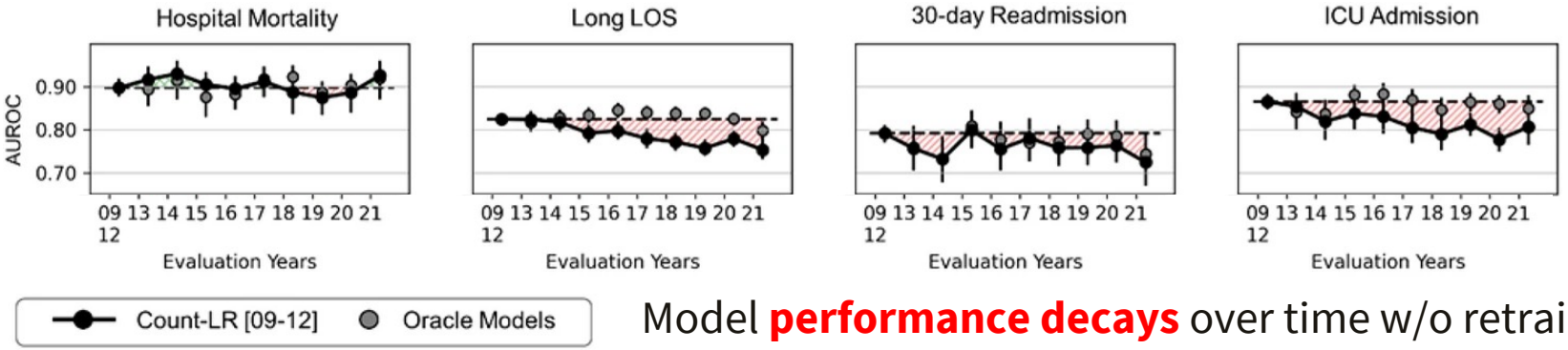


EHR FM's more robust to **temporal distribution shifts**

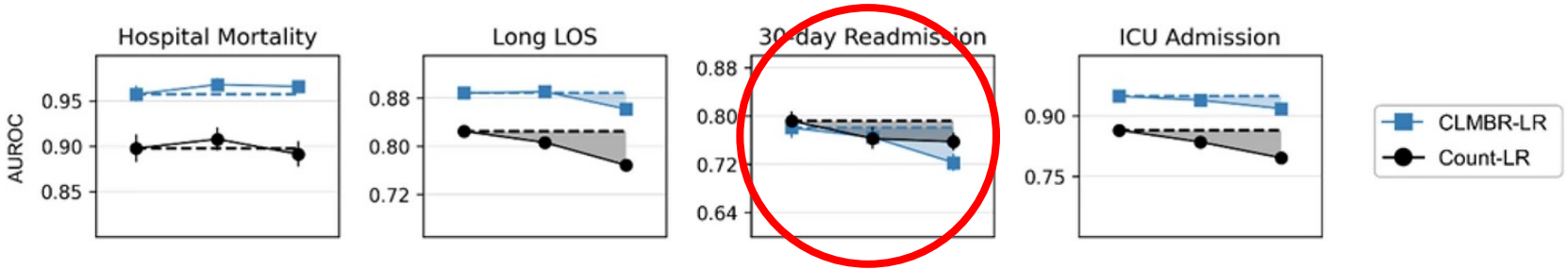
(Guo et al. 2023)

Benefits of Autoregressive EHR Foundation Models

Improved Distributional Robustness



Model **performance decays** over time w/o retraining

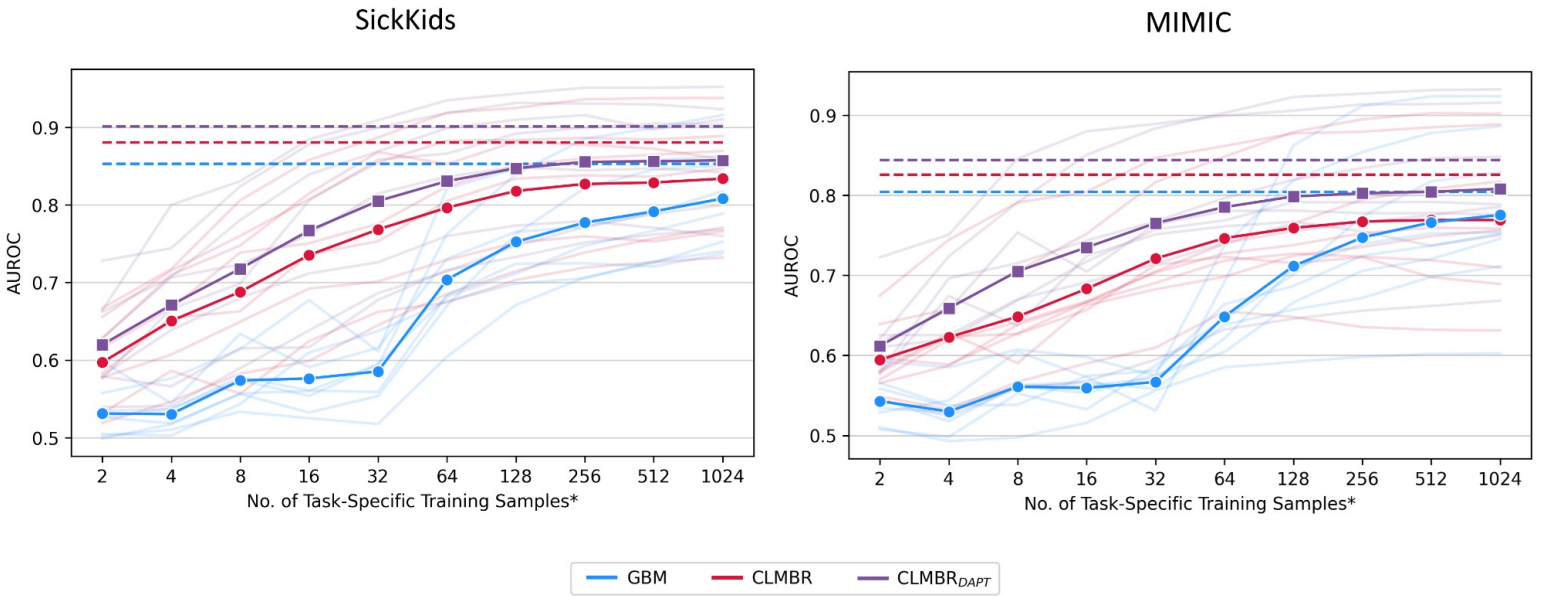


Struggles with longer time horizons

(Guo et al. 2023)

Benefits of Autoregressive EHR Foundation Models

Improved Distributional Robustness

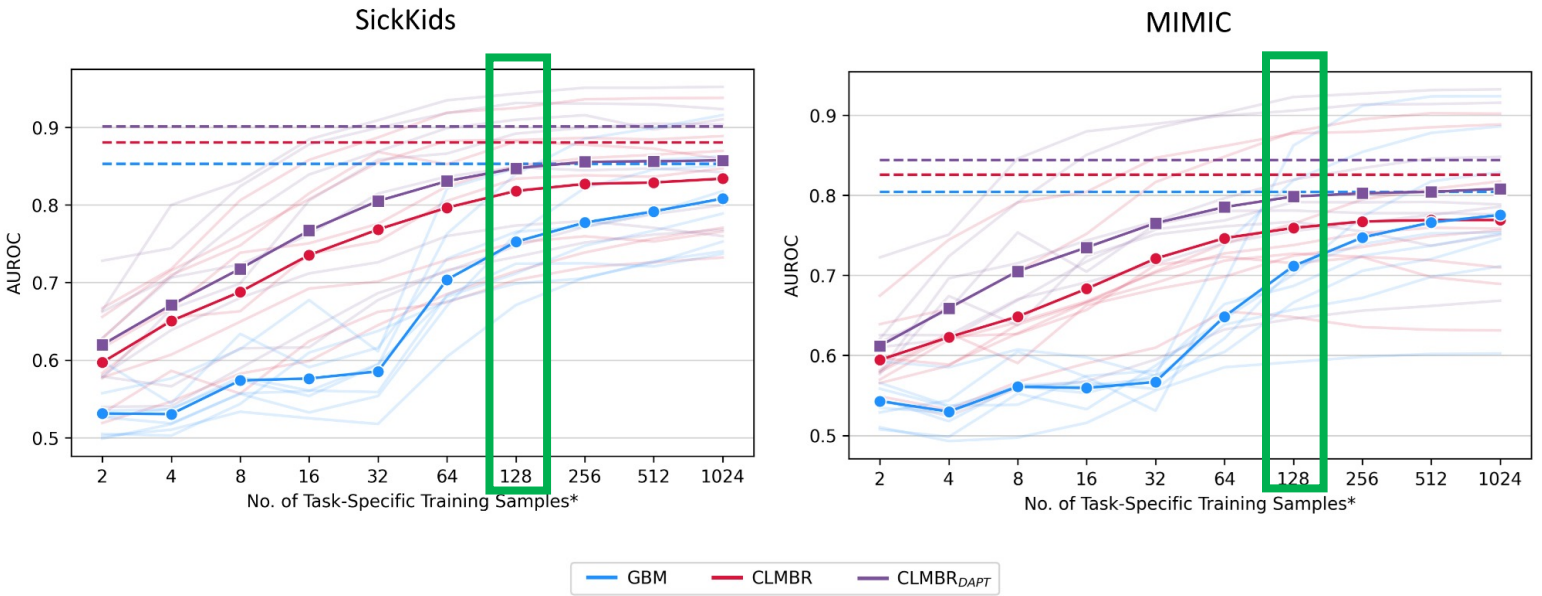


60-90% reduction pretraining data for continued pretraining **across sites**

(Guo et al. 2024)

Benefits of Autoregressive EHR Foundation Models

Improved Distributional Robustness



128 examples performs as well as training on all examples when using gradient boosted models (~2500 examples on avg.)

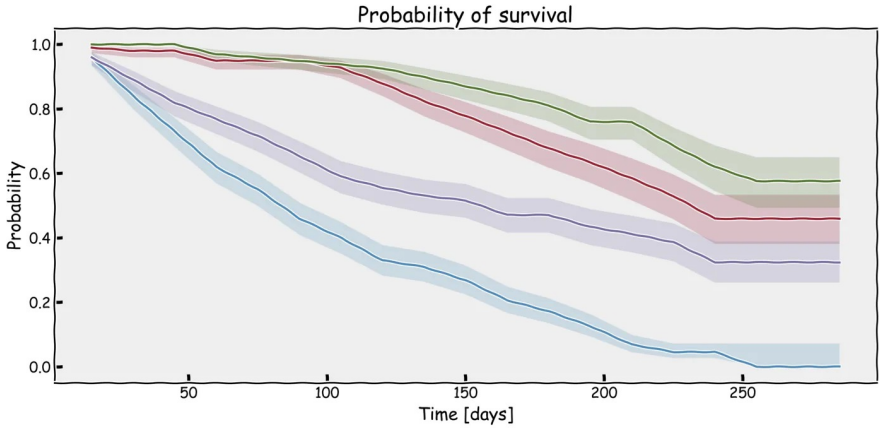
(Guo et al. 2024)

Time-to-Event Modeling

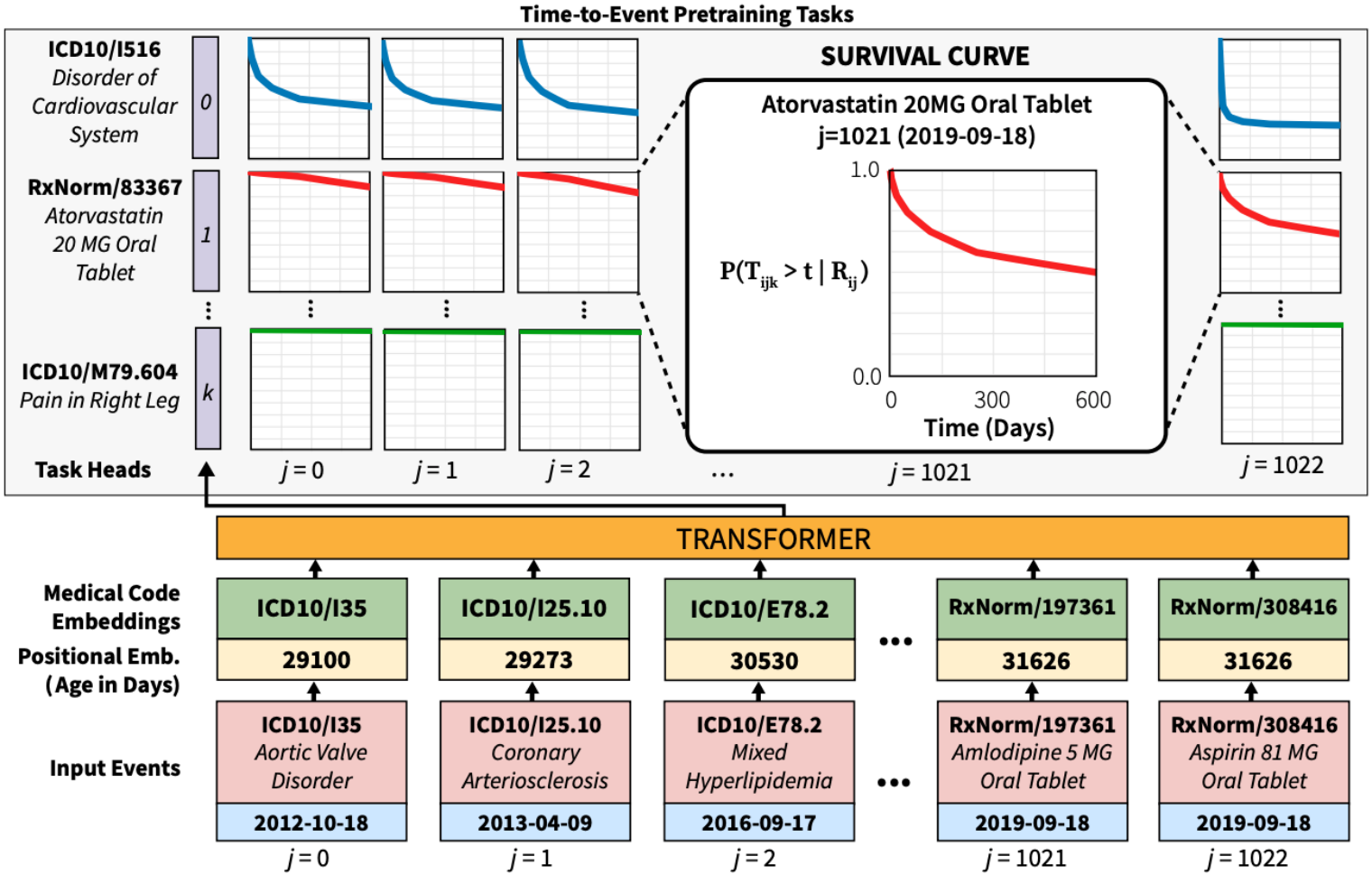
MOTOR: A Time-To-Event Foundation Model For Structured Medical Records

Key Intuition: Predict **if something will happen** and **when it will happen**

- Naturally **handle censoring**
- Model **longer disease trajectories**
- Build a **foundation model for TTE**



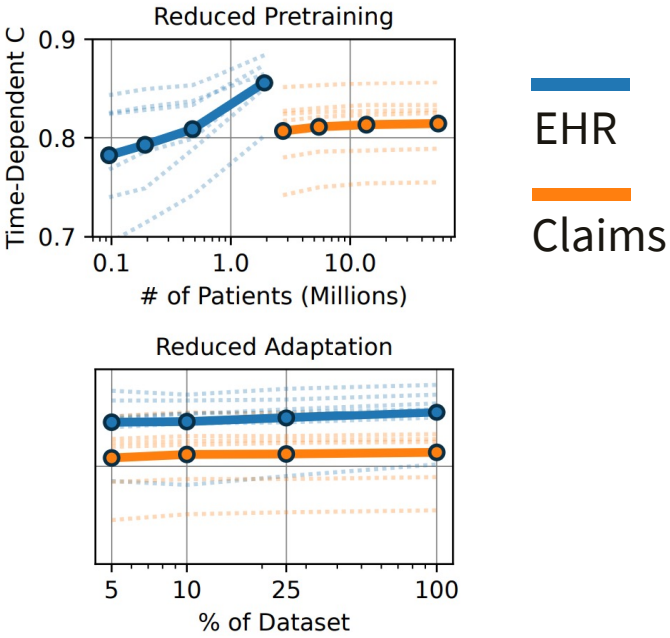
- **8,192** time-to-event tasks
- Up-to **55M patients**
- **19** evaluation tasks


Time-to-Event Pretraining Tasks
ICD10/I516
Disorder of Cardiovascular System
0
RxNorm/83367
Atorvastatin 20 MG Oral Tablet
1
ICD10/M79.604
Pain in Right Leg
k
Task Heads
j = 0
j = 1
j = 2
...
j = 1021
j = 1022
SURVIVAL CURVE
Atorvastatin 20MG Oral Tablet
j=1021 (2019-09-18)
1.0
0.0
0
300
600
Time (Days)
P(T_ijk > t | R_ij)
TRANSFORMER
Medical Code Embeddings
Positional Emb. (Age in Days)
Input Events
ICD10/I35
29100
ICD10/I35
Aortic Valve Disorder
2012-10-18
j = 0
ICD10/I25.10
29273
ICD10/I25.10
Coronary Arteriosclerosis
2013-04-09
j = 1
ICD10/E78.2
30530
ICD10/E78.2
Mixed Hyperlipidemia
2016-09-17
j = 2
RxNorm/197361
31626
RxNorm/197361
Amlodipine 5 MG Oral Tablet
2019-09-18
j = 1021
RxNorm/308416
31626
RxNorm/308416
Aspirin 81 MG Oral Tablet
2019-09-18
j = 1022

Time-to-Event Pretraining

Method	Dataset	Celiac	HA	Lupus	NAFLD	Cancer	Stroke
Cox PH	EHR-OMOP	0.689	0.761	0.770	0.726	0.793	0.779
DeepSurv	-	0.704	0.823	0.790	0.800	0.811	0.830
DSM	-	0.707	0.828	0.784	0.805	0.809	0.835
DeepHit	-	0.695	0.826	0.807	0.805	0.809	0.833
RSF	-	0.729	0.836	0.787	0.802	0.824	0.840
MOTOR-Scratch	-	0.696	0.795	0.803	0.821	0.777	0.831
MOTOR-Probe	-	0.802	0.884	0.850	0.859	0.865	0.874
MOTOR-Finetune	-	**0.802**	**0.887**	**0.863**	**0.864**	**0.865**	**0.875**
Cox PH	MERATIVE	0.538	0.783	0.749	0.799	0.628	0.693
DeepSurv	-	0.719	0.814	0.809	0.828	0.801	0.753
DSM	-	0.725	0.814	0.812	0.833	0.805	0.758
DeepHit	-	0.722	0.815	0.809	0.828	0.802	0.753
RSF	-	0.705	0.810	0.805	0.838	0.798	0.746
MOTOR-Scratch	-	0.737	0.821	0.826	0.850	0.821	0.775
MOTOR-Probe	-	0.755	0.828	0.833	0.856	0.825	0.789
MOTOR-Finetune	-	**0.762**	**0.831**	**0.838**	**0.862**	**0.834**	**0.794**

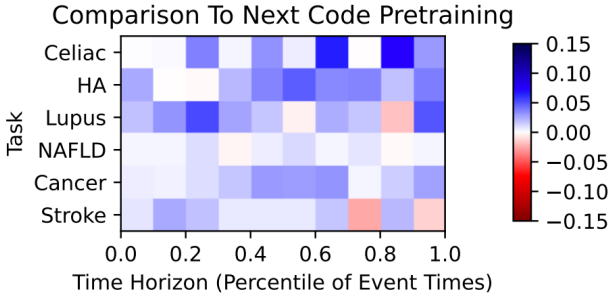
Average 4.6% improvement in C-statistics



Requires up to 95% less data for adaptation

Time-to-Event Pretraining

Objective	Celiac	HA	Lupus	NAFLD	Cancer	Stroke
Next Code	0.774	0.862	0.842	0.860	0.860	0.857
Time-to-Event	**0.802**	**0.887**	**0.863**	**0.864**	**0.865**	**0.875**



Outperforms autoregressive pretraining

Outperforms autoregressive pretraining on longer time horizons

Evaluation: EHR Foundation Models

Reproducibility in Healthcare AI

SCIENCE TRANSLATIONAL MEDICINE | PERSPECTIVE

BIOMEDICAL POLICY

Reproducibility in machine learning for health research: Still a ways to go

Matthew B. A. McDermott[1*†], Shirly Wang[2,3†], Nikki Marinsek[4], Rajesh Ranganath[5], Luca Foschini[4], Marzyeh Ghassemi[2,6,7]

Longstanding Reproducibility Challenges

Medical data are noisy, **replete with errors, biases, missingness**

Most AI is **trained and tested** on **cleaned data**

REVIEW

Global healthcare fairness: We should be sharing more, not less, data

Kenneth P. Seastedt[1*], Patrick Schwab[2], Zach O'Brien[3], Edith Wakida[4], Karen Herrera[5], Portia Grace F. Marcelo[6], Louis Agha-Mir-Salim[7,8], Xavier Borrat Frigola[8,9], Emily Boardman Ndulue[10], Alvin Marcelo[11], Leo Anthony Celi[8,12,13]

PLOS DIGITAL HEALTH

Open & Accessible Model Weights

Sharing pre-trained model

Initially we really hoped to share our models but unfortunately, the pre-trained models are no longer sharable. According to SBMI Data Service Office: "Under the terms of our contracts with data vendors, we are not permitted to share any of the data utilized in our publications, as well as large models derived from those data."

https://github.com/ZhiGroup/Med-BERT

Transfer learning is the primary
value prop of foundation models!

Foundation Models Risk Increasing our Reproducibility Crisis

Thought Leadership on Medical Foundation Models

Healthcare

How Foundation Models Can Advance AI in Healthcare

This new class of models may lead to more affordable, easily adaptable health AI.

Dec 15, 2022 |
Jason Fries, Ethan Steinberg, Scott Fleming, Michael Wornow, Yizhe Xu, Keith Morse, Dev Dash, Nigam Shah

Review Article | Open access | Published: 29 July 2023

The shaky foundations of large language models and foundation models for electronic health records

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries & Nigam H. Shah

npj Digital Medicine **6**, Article number: 135 (2023) | Cite this article

Enriching the Axes of Evaluation

Enabling Open Science

- Improve transparency of evaluation with
  - **More open datasets + models**
  - **Shared recipes for training and evaluating models**
- Better healthcare benchmarks
  - **Measure properties beyond accuracy**
  - **Align with real user (healthcare workers) needs**

Many benchmarks for EHRs exist…

Benchmark	ICU/ED Visits	Non-ICU/ED Visits	# of Tasks	Few Shot Eval	Public Model Weights
MIMIC-Extract			5		
Purushotham 2018			3		
Harutyunyan 2019			4		
Gupta 2022					
COP-E-CAT					
Xie 2022					
eICU					
EHR PT					
FIDDLE					
HiRID-ICU					
Solares 2020					

…but they offer a narrow viewpoint of patients…

Benchmark	ICU/ED Visits	All Other Visit Types	# of Tasks	Few Shot Eval	Public Model Weights
MIMIC-Extract	✅		5		
Purushotham 2018	✅		3		
Harutyunyan 2019	✅		4		
Gupta 2022	✅	🟡	4		
COP-E-CAT	✅	🟡			
Xie 2022	✅	🟡			
eICU	✅				
EHR PT	✅				
FIDDLE	✅				
HiRID-ICU	✅		6		
Solares 2020	✅	✅	2		

Almost all are **sourced from a single dataset** called "**MIMIC-III**" which contains **~40k patients** from **one hospital**

Wornow et al. 2023

…a limited set of tasks…

Benchmark	ICU/ED Visits	All Other Visit Types	# of Tasks	Few Shot Eval	Public Model Weights
MIMIC-Extract	✅		5		
Purushotham 2018	✅		3		
Harutyunyan 2019	✅		4		
Gupta 2022	✅	🟡	4		
COP-E-CAT	✅	🟡	4		
Xie 2022	✅	🟡	3		
eICU	✅		4		
EHR PT	✅		11		
FIDDLE	✅		3		
HiRID-ICU	✅		6		
Solares 2020	✅	✅	2		

Wornow et al. 2023

…are not designed for few-shot evaluation…

Benchmark	ICU/ED Visits	All Other Visit Types	# of Tasks	Few Shot Eval	Public Model Weights
MIMIC-Extract	✅		5		
Purushotham 2018	✅		3		
Harutyunyan 2019	✅		4		
Gupta 2022	✅	🟡	4		
COP-E-CAT	✅	🟡	4		
Xie 2022	✅	🟡	3		
eICU	✅		4		
EHR PT	✅		11	✅	
FIDDLE	✅		3		
HiRID-ICU	✅		6		
Solares 2020	✅	✅	2		

Wornow et al. 2023

…and do not publish model weights

Benchmark	ICU/ED Visits	All Other Visit Types	# of Tasks	Few Shot Eval	Public Model Weights
MIMIC-Extract	✅		5		
Purushotham 2018	✅		3		
Harutyunyan 2019	✅		4		
Gupta 2022	✅	🟡	4		
COP-E-CAT	✅	🟡	4		
Xie 2022	✅	🟡	3		
eICU	✅		4		
EHR PT	✅		11	✅	
FIDDLE	✅		3		
HiRID-ICU	✅		6		
Solares 2020	✅	✅	2		

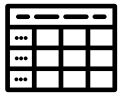
Transparency and reproducibility are key to advance science and build trust!

Wornow et al. 2023

Releasing New Medical Datasets

EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models **2023**

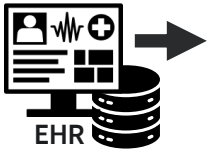
6,739 Patients

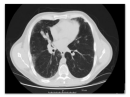


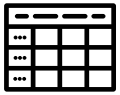
SPOTLIGHT

INSPECT: A Multimodal Dataset for Patient Outcome Prediction of Pulmonary Embolisms **2023**

19,402 Patients



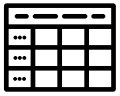


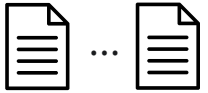


NeurIPS Datasets & Benchmarks 2023

MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records **2023**

267 Patients





ML4H Symposium 2024

BEST THEMATIC PAPER

AAAI 2024

ORAL

EHRSHOT

Benchmark	ICU/ED Visits	All Other Visit Types	# of Tasks	Few Shot Eval	Public Model Weights
MIMIC-Extract	✅		5		
Purushotham 2018	✅		3		
Harutyunyan 2019	✅		4		
Gupta 2022	✅	🟡	4		
COP-E-CAT	✅	🟡	4		
Xie 2022	✅	🟡	3		
eICU	✅		4		
EHR PT	✅		11	✅	
FIDDLE	✅		3		
HiRID-ICU	✅		6		
Solares 2020	✅	✅	2		
EHRSHOT	✅	✅	**15**	✅	✅

Transparency

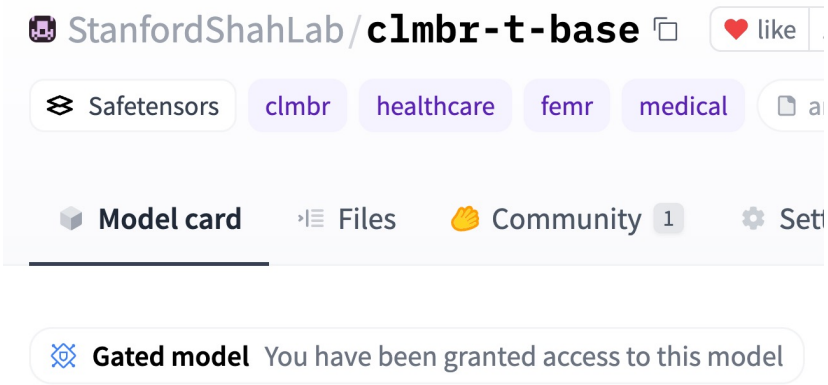
Representative data

Broad range of tasks

Few-shot eval

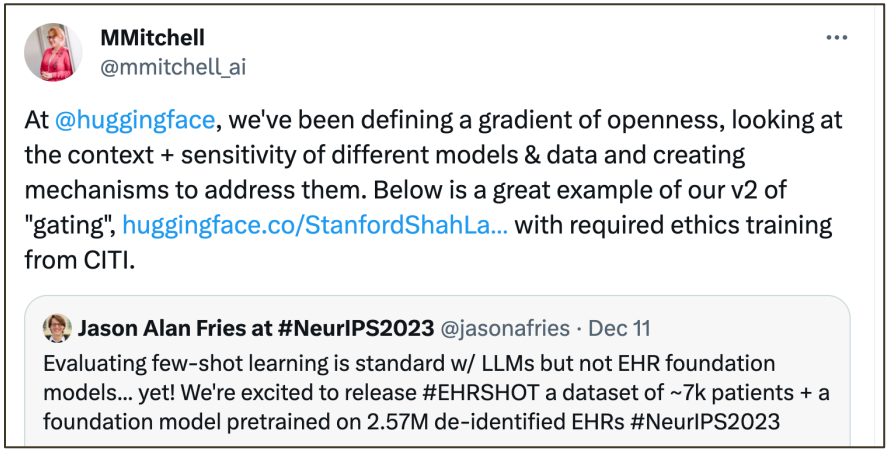
Wornow et al. 2023

Enabling Open Science



First EHR model hub release!

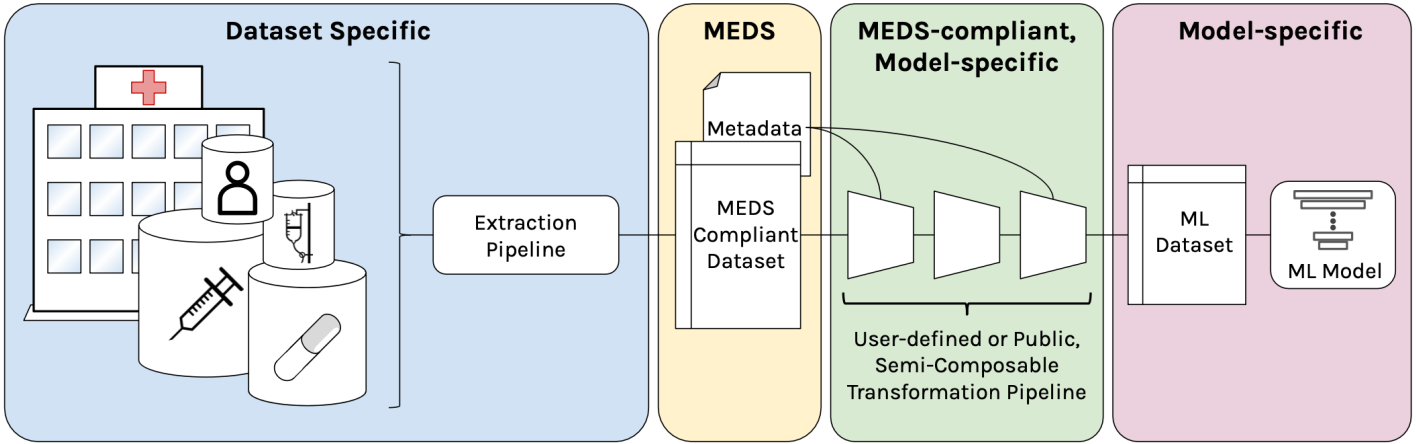
- Gated model on Hugging Face
- Requires **CITI ethics training**
- **Non-commercial use only**



Margaret Mitchell

Chief AI Ethics Scientist, Hugging Face

Medical Event Data Standard (MEDS)

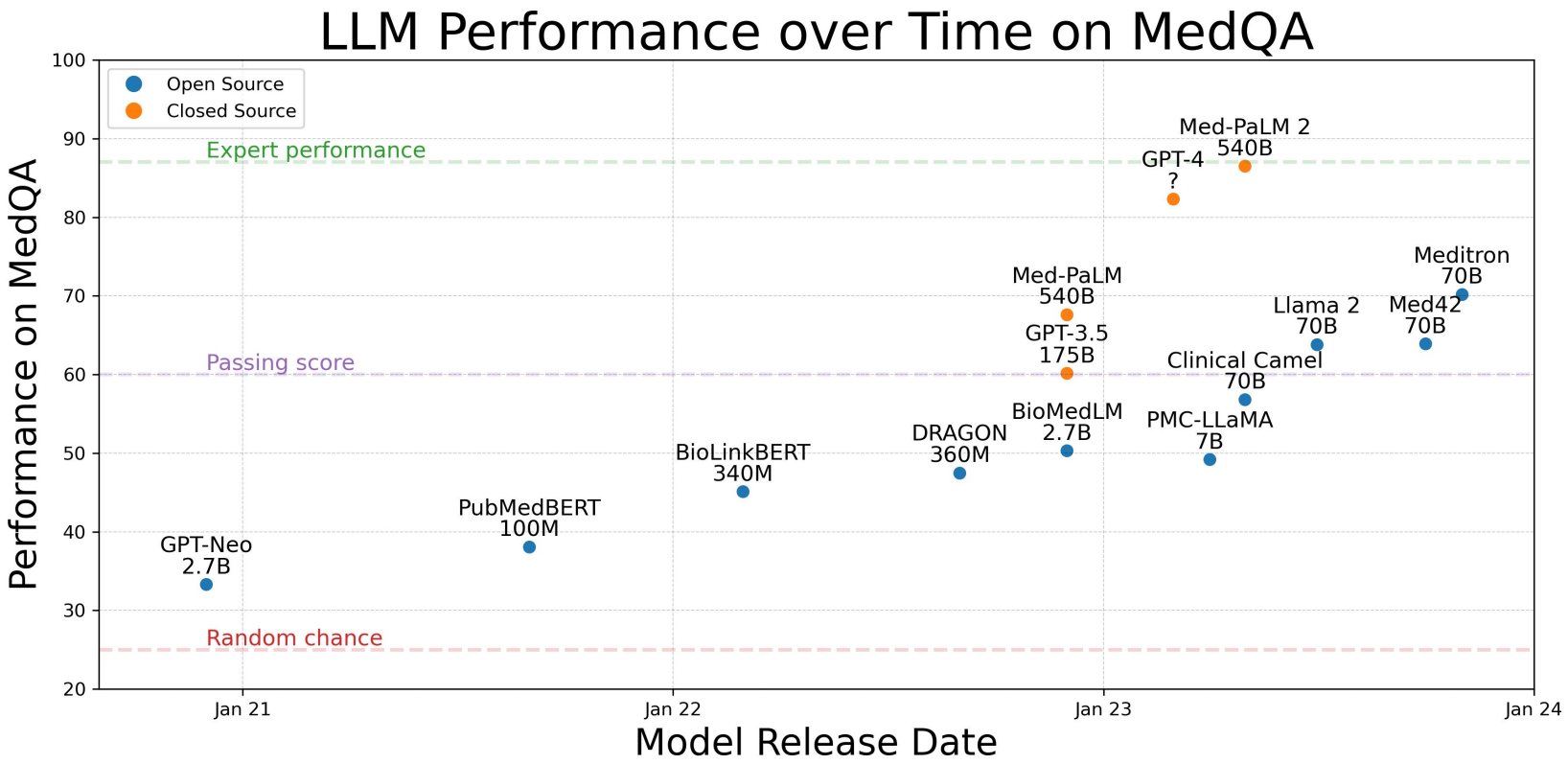


Open Data Schema for Health AI Practitioners

Bert Arnrich, Edward Choi, Jason A. Fries, Matthew B. A. McDermott, Jungwoo Oh, Tom J Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, Robin van de Water

https://github.com/Medical-Event-Data-Standard/meds

Of LLMs and Medical Knowledge…



Multiple Choice vs. Longitudinal Patient Timelines

MedQA

Question: A 35-year-old man is brought to the emergency department by a friend 30 minutes after the sudden onset of right-sided weakness and difficulty speaking. [...] **Which of the following is the most appropriate next step in diagnosis?**

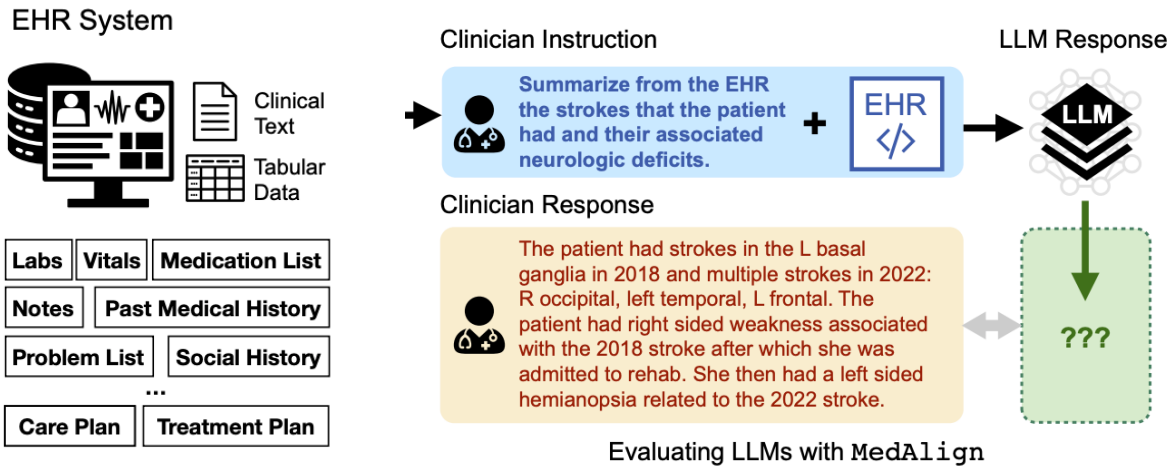
(A) Echocardiography with bubble study
(B) Adenosine stress test
(C) Cardiac catheterization
(D) Cardiac MRI with gadolinium
(E) CT angiography

USMLE
United States Medical Licensing Exam

```
<record>
    <visit type="Emergency Room Visit" start="10/08/2018 20:00">
        <day start="10/08/2018 20:00">
            rson>
             Birth:7/19/1966
             Rac
             Gen
             Eth
             Age
             Age
            erso
            ndit
             <co
          ondi
          <visit_
             <co
          </visit
          <measur
             <co
          </measu
          <proced
             <co
          </proce
    <observation start="10/08/2018 08:10 PM">
        <code>[LOINC/LP21258-6] Oxygen saturation 96 %</code>
    </observation>
    <note type="emergency department note" start="10/08/2018 08:10 PM">
        Emergency Department Provider Note Name: Jessica Jones, MD MRN: [1234555]
        ED Arrival: 10/08/2018 Room #: 17B History and Physical Triage: 52 year old woman
        with unknown past medical history presenting with right sided weakness since about
        2 hours ago. Last known normal 5:45pm. She said she was feeling well and then suddenly
        noticed that her right arm and leg went limp. She denies taking any blood thinners,
        and has had no recent surgeries. NIHSS currently graded at an 8: 4 no movement in R
        arm and 4 no movement in R leg CT head is negative for any bleed or any early ischemic
        changes. INR is 1.0, Plt 133. Discussed with patient the severity of symptoms and the
        concern that they are caused by a stroke, and that IV tPA is the best medication to
        reduce the risk of long term deficits. Patient is agreeable and IV tPA was given at
        8:20pm. Initially SBP 210/100, labetalol 5mg IV x1 given and came down to 180/90.
        IV tPA given after this point. Patient will need to be admitted to the ICU, with close
        neurological monitoring. Plan for head CT 24 hours post IV tPA administration, stroke
        workup including LDL, HA1C, echo, tele monitoring. Local neurology consult in AM.
    </note>
    <measurement start="10/08/2018 08:15 PM">
        <code>[LOINC/70182-1] NIHSS 8 </code>
```

Encode Patient Timelines as XML, JSON, FHIR, etc.

Instruction Tuning: Aligning with Clinical Needs



MedAlign: A Clinician-Generated Benchmark Dataset for Instruction Following with Electronic Medical Records [1]

- **15** clinicians / **7** specialties
- 983 instructions, 303 responses
- Assess **real information needs**

[1] Fleming et al. “A Clinician-Generated Benchmark Dataset for Instruction Following with Electronic Medical Records”. *AAAI*. 2024.

Instruction Tuning: Aligning with Clinical Needs

Table 2: MEDALIGN instruction categories and example instructions.

Category	Example Instruction	Gold	All
Retrieve & Summarize	Summarize the most recent annual physical with the PCP	223	667
Care Planning	Summarize the asthma care plan for this patient including relevant diagnostic testing, exacerbation history, and treatments	22	136
Calculation & Scoring	Identify the risk of stroke in the next 7 days for this TIA patient	13	70
Diagnosis Support	Based on the information I've included under HPI, what is a reasonable differential diagnosis?	4	33
Translation	I have a patient that speaks only French. Please translate these FDG-PET exam preparation instructions for her	0	2
Other	What patients on my service should be prioritized for discharge today?	41	75
Total		303	983

Clinicians spend 49% of their day interacting with EHRs! **>66% of instructions** were **"retrieve & summarize"** data from the EHR.

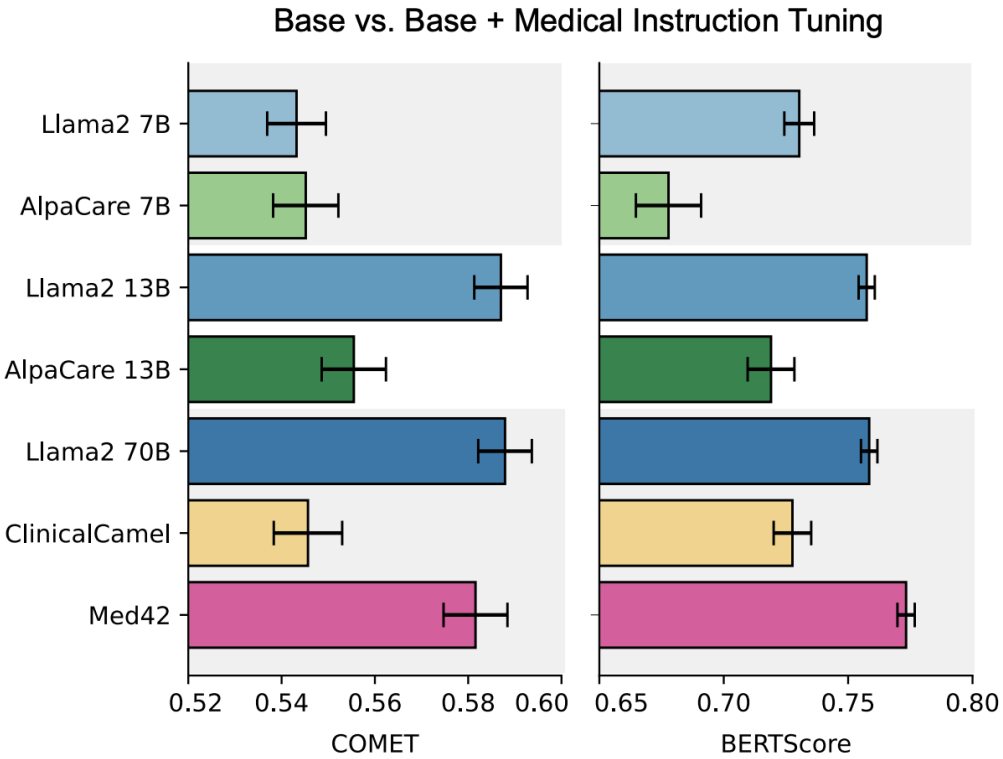
Instruction Tuning: Aligning with Clinical Needs

Model	Context	Correct ↑	WR ↑	Rank ↓
GPT-4 (MR)	32768†	**65.0%**	0.658	2.80
GPT-4	32768	60.1%	**0.676**	**2.75**
GPT-4	2048*	51.8%	0.598	3.11
Vicuña-13B	2048	35.0%	0.401	3.92
Vicuña-7B	2048	33.3%	0.398	3.93
MPT-7B-Instruct	2048	31.7%	0.269	4.49

GPT-4 **35% Error Rate**

[1] Fleming et al. “A Clinician-Generated Benchmark Dataset for Instruction Following with Electronic Medical Records”. *Under Review*. 2023.

Instruction Tuning in Medical LLMs



Current short instruction tuning tasks for medicine (e.g., MedQA) **actually hurt performance on MedAlign**

A Single Benchmark Does NOT Tell the Whole Story!

Opportunities: The Road Ahead

Open Weights are Critical to Fair & Secure Models

Why Anthropic and OpenAI are obsessed with securing LLM model weights

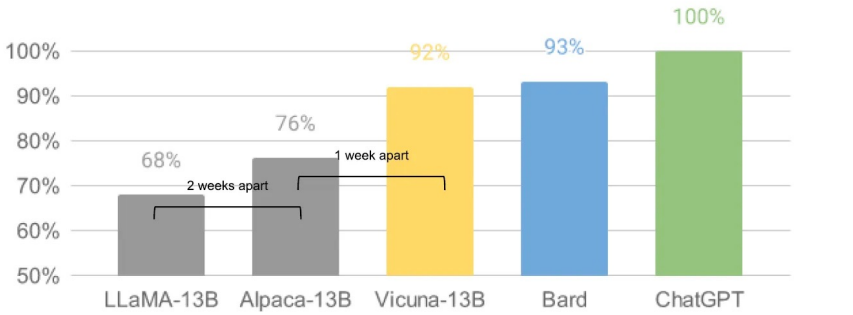
Transparency (training data, model weights) is critical for fair and secure models

OpenAI

"Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."

Calls for the Academic Community

Smaller Models, Cheaper to Train



*GPT-4 grades LLM outputs. Source: https://vicuna.lmsys.org/

Lead Building Open, Reproducible Medical Base Models

Reimagine Model Evaluation

Knowledge Retrieval

Real Clinical Workflows

Human-AI Collaboration

AI will augment existing roles

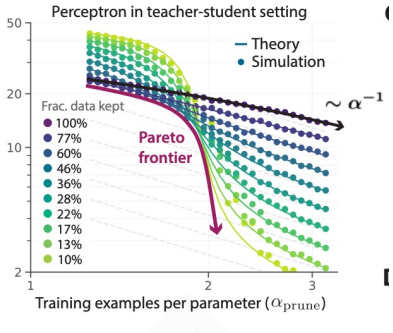
We need to **measure human + AI performance**

Data-Centric AI for EHRs

We don't have enough EHR data for today's generative AI

- **340M** people in the U.S.
- Only ~**550B tokens** of structured EHR data
  - Llama-2: **2 trillion tokens**
  - Mistral-7B: **8 trillion token**

Beat Power Law Scaling



Data Pruning?

Sorscher et al. 2022

Synthetic Data Generation

Cover gaps in training data?

CEHR-GPT (Pang et al. 2024)

Synthetic data: breaking the data logjam in machine learning for healthcare

Mihaela van der Schaar · Zhaozhi Qian · Boris van Breugel · Nick Maxfield

February 3, 2023 · 27 min read

Thank You!

jason-fries@stanford.edu