Foundation Models for Electronic Health Records

BIODS 271: Foundation Models for Healthcare

May 28, 2025

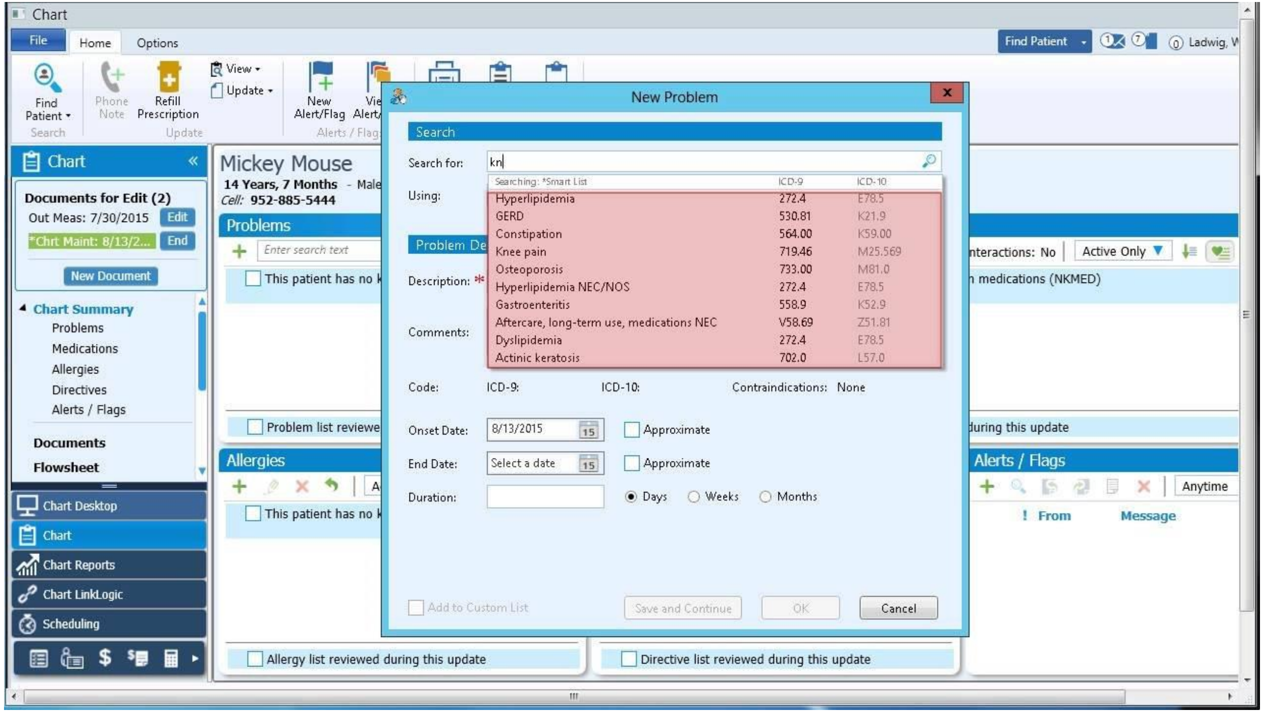
Jason Fries, PhD Research Scientist, Shah Lab
Center for Biomedical Informatics Research

Outline

- **Overview: EHR Data & Tasks**
  - Electronic Health Records (EHRs)
  - AI for Healthcare Tasks
- **Modeling: FMs for Structured EHRs**
  - Formulating Self-Supervision
  - Pretraining Objectives
- **Evaluation**
- **Future: Research Opportunities**

Overview:
EHR Data & Tasks

Electronic Health Records (EHR)

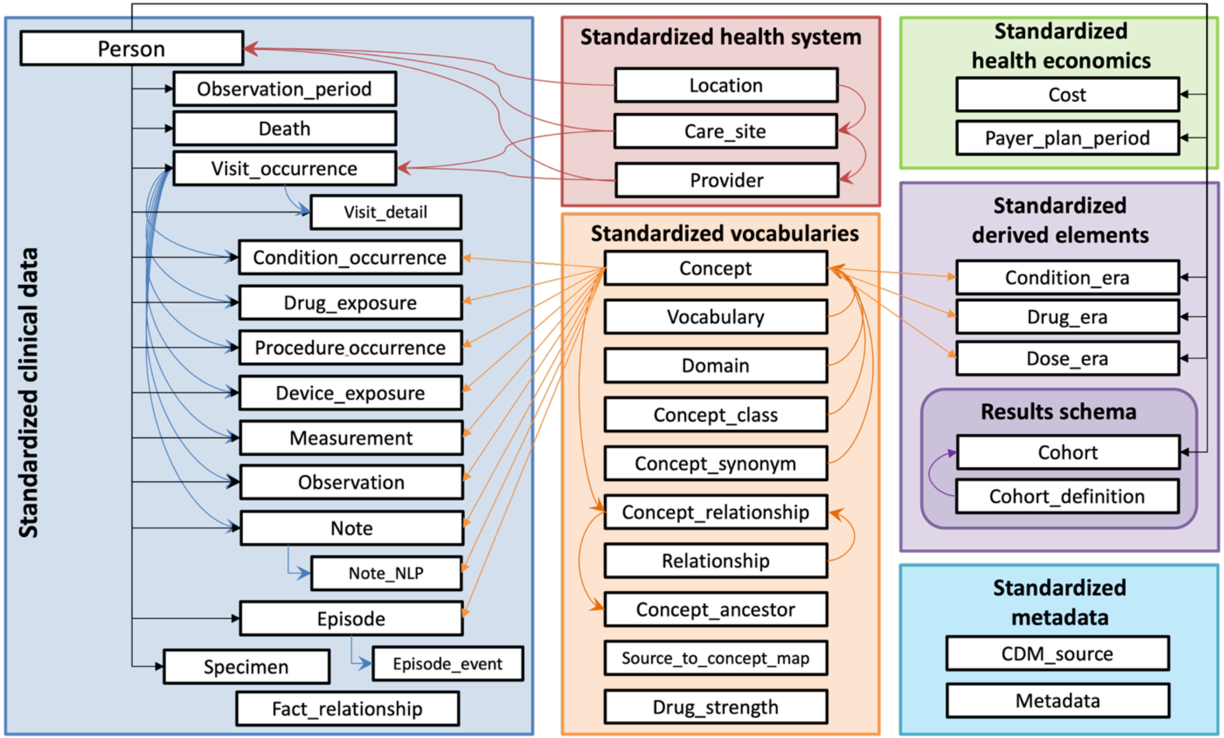


Healthcare View

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

Epic Cerner

Electronic Health Records (EHR)



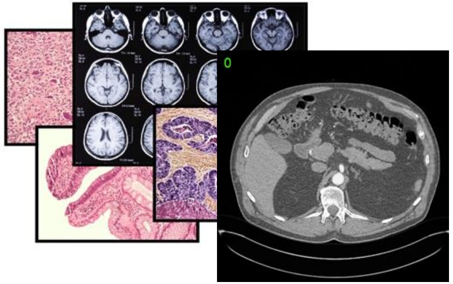
Data Scientist View

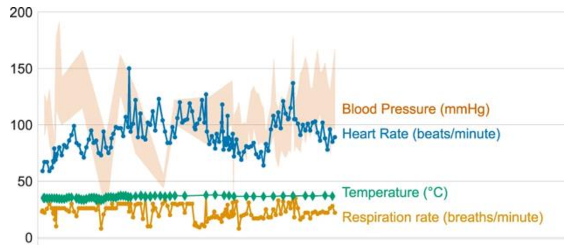
- Relational databases
- Some data model (Epic, OMOP, i2b2)
- Apply functions to all patients

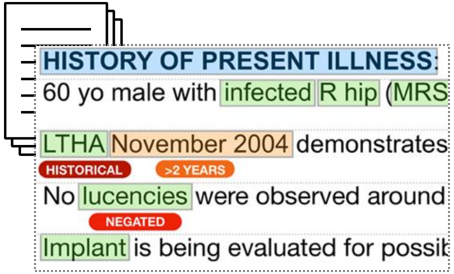
Healthcare Data is Inherently Multimodal

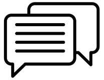
Labs | Vitals | Medication List
Notes | Past Medical History
Problem List | Social History
...
Care Plan | Treatment Plan

Tabular Data

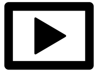








Audio /Conversations



Video

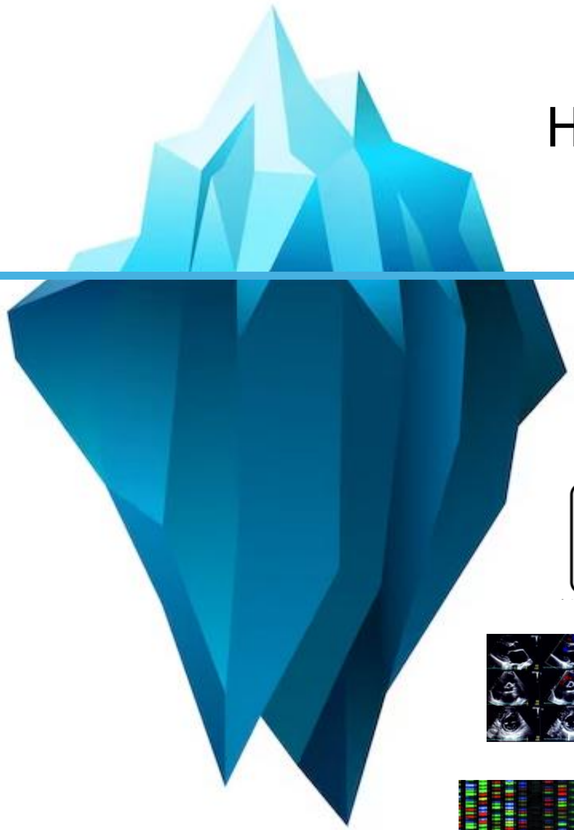
Genomics

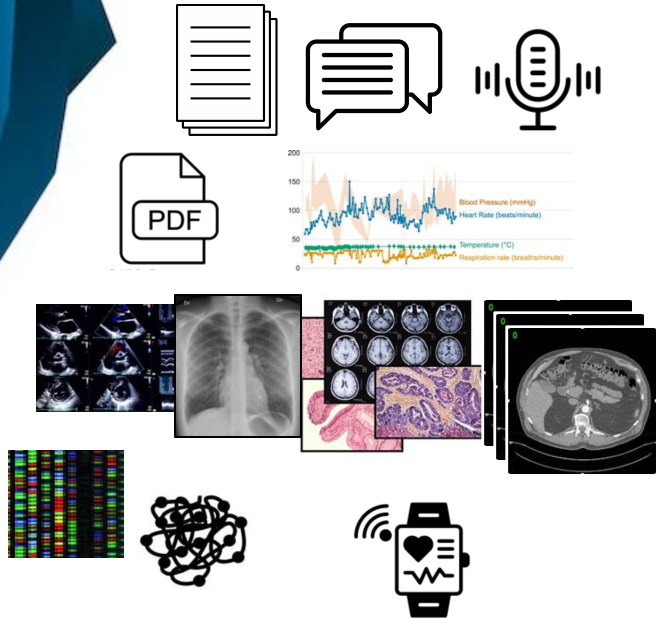
STRUCTURED DATA

UNSTRUCTURED DATA

Hospital data is growing at a rate of **36% per year**

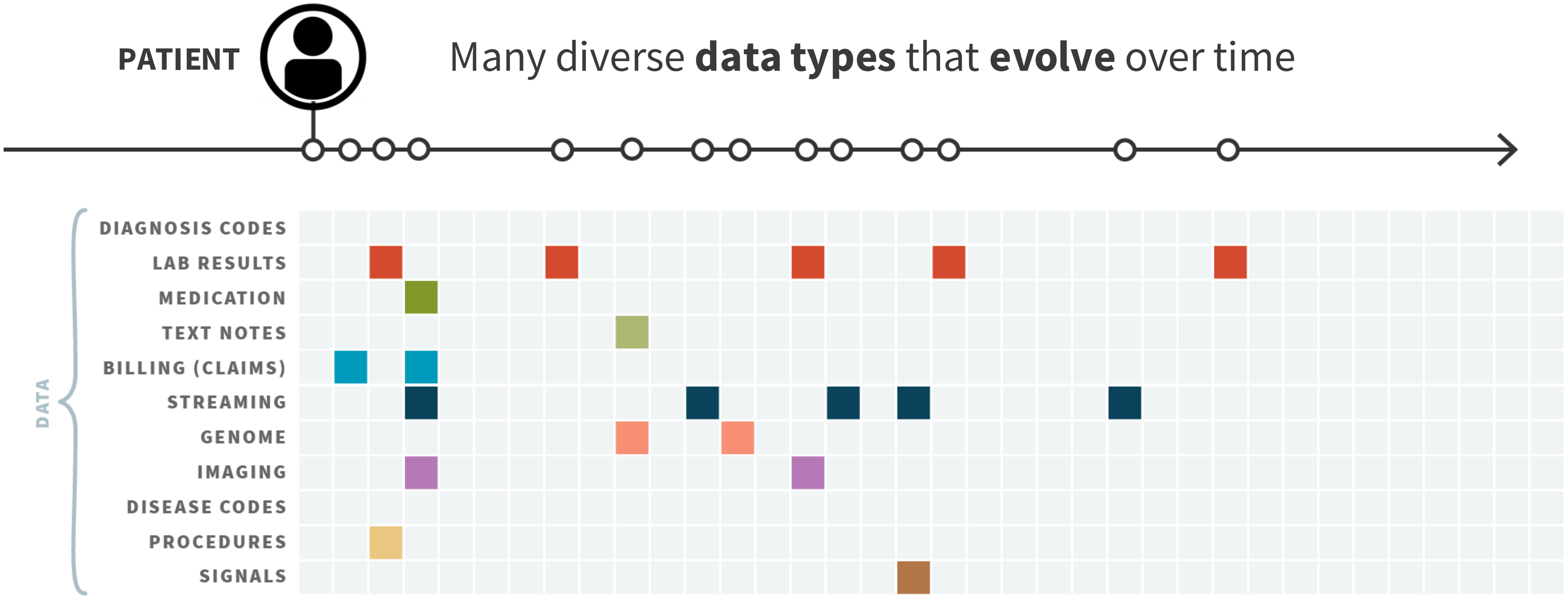
World Economic Forum, Dec. 2019





Hard to use for medical decision making

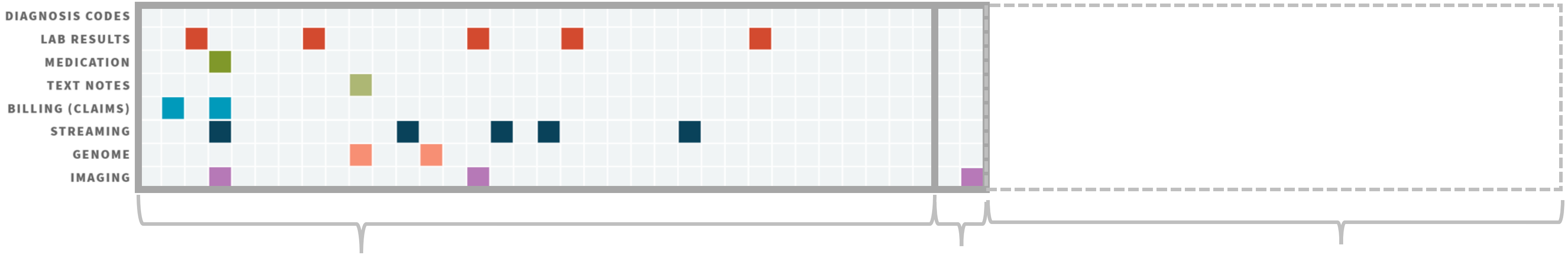
Electronic Health Records (EHRs) are Multimodal Timelines



Longitudinal EHRs provide a **holistic view of multimodal data**

AI for Healthcare Requires Temporal Reasoning

Diagnosis (Classification) | **Prognosis**



Example ML Applications

What Occurred in the Past?

- Chart summarization
- Clinical trial recruitment

What is Occurring Now?

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

Predict Future Risks & Intervention Benefits

- What is the likelihood that this patient will develop lung cancer?

Stakeholders

Clinicians

Whether to Treat | **How to Treat**

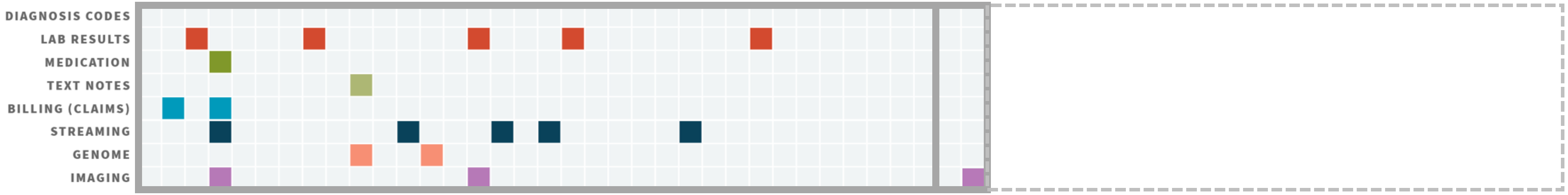
subject to

Policy | Capacity to Act

Intervention Properties

Foundation Models Are Essential for AI in Healthcare

Diagnosis (Classification) **Prognosis**



What Occurred in the Past?

What is Occurring Now?

Predict Future Risks & Intervention Benefits

Example ML Applications

?

?

?

Many stakeholder groups with **distinct needs**

Stakeholders

- Clinicians
- Hospital Administrators
- Insurance Providers
- Pharma
- Regulatory Agencies
- Patient s
- ...
- Researchers

How Can AI Improve Healthcare?

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

Check for updates

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

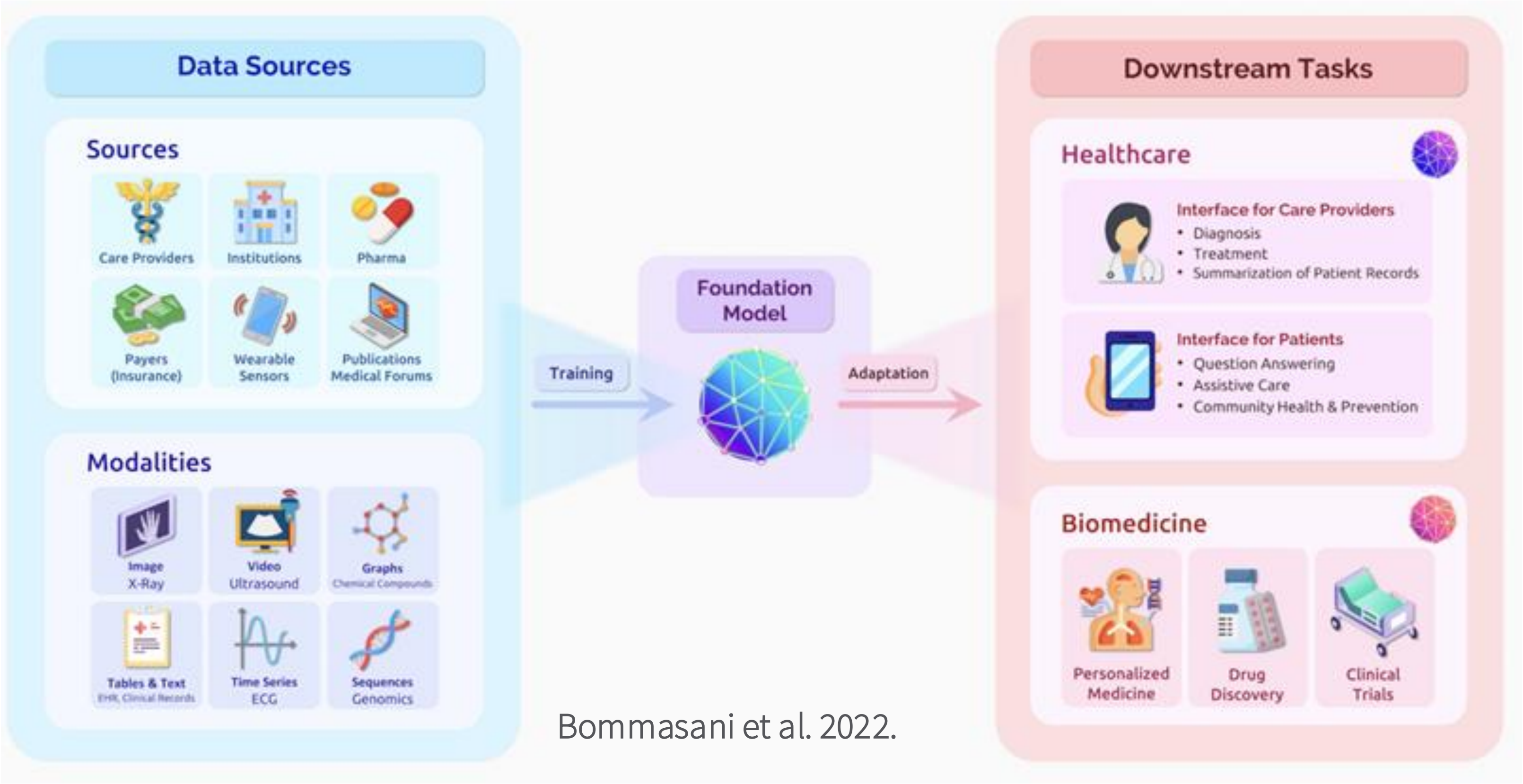
A Sketch of Healthcare Tasks

- **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)

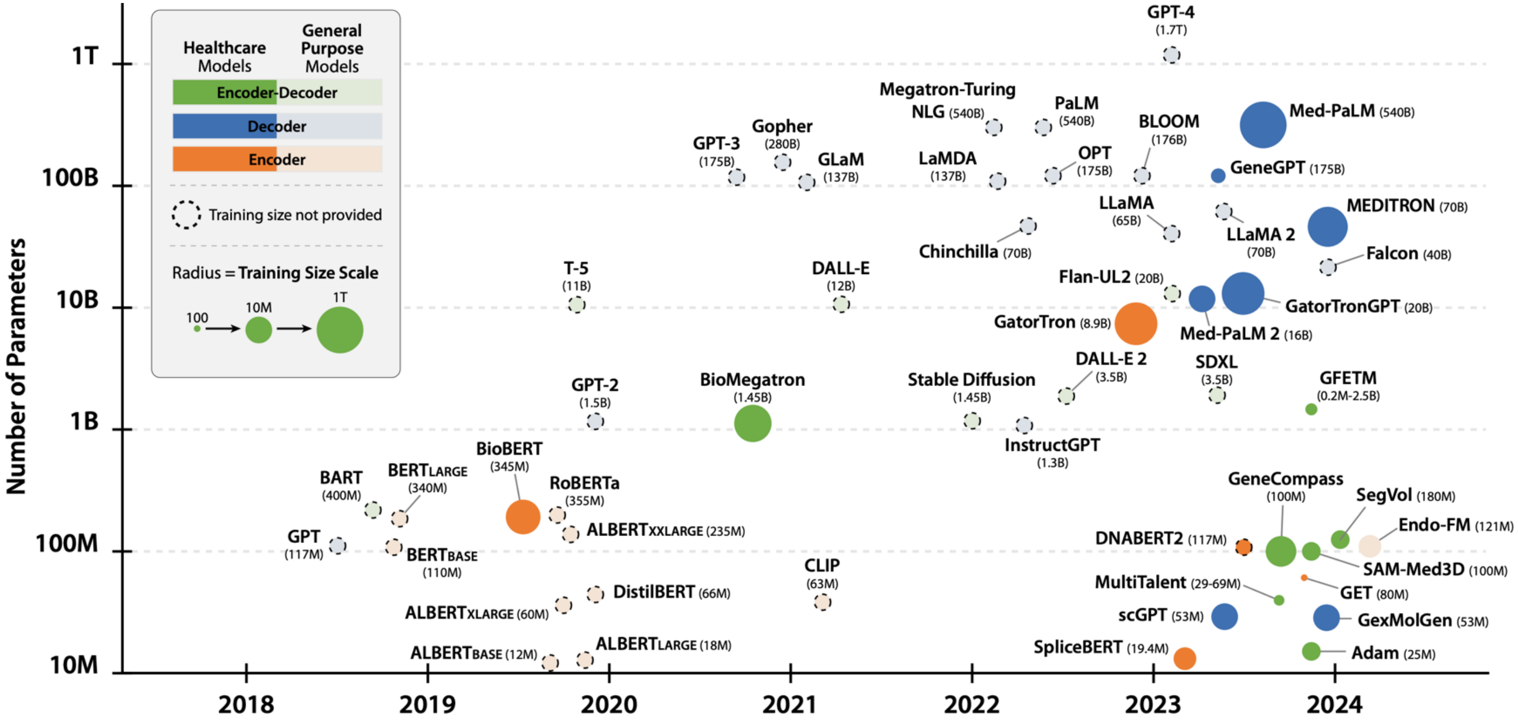
- **Research**
  - Causal inference (e.g. drug trials and observational studies)
  - Identify off-label drug benefits

Foundation Models and AI's "Industrial Age"



Bommasani et al. 2022.

Opportunity for AI to reimagine how we **interact and understand medical data**



Khan et al., "A Comprehensive Survey of Foundation Models in Medicine," 2025.

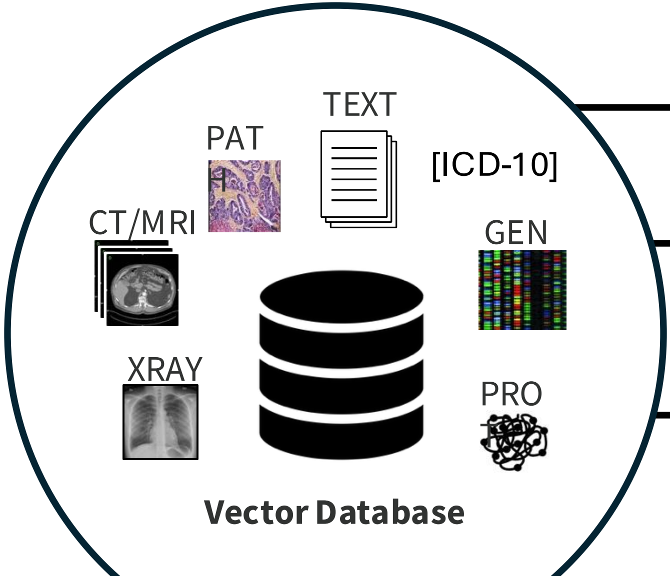
14

The Future

We must build systems for patient timeline data that are **fast**, **multimodal**, and **interactive**

"I can't just go to the medical records department to have them pull 500 charts on a certain type of patient."

Byrne Lee, MD, Clinical Professor,
Surgical Oncology, Stanford Health Care



Design **chat interfaces**

Find **similar patients** to inform decision making

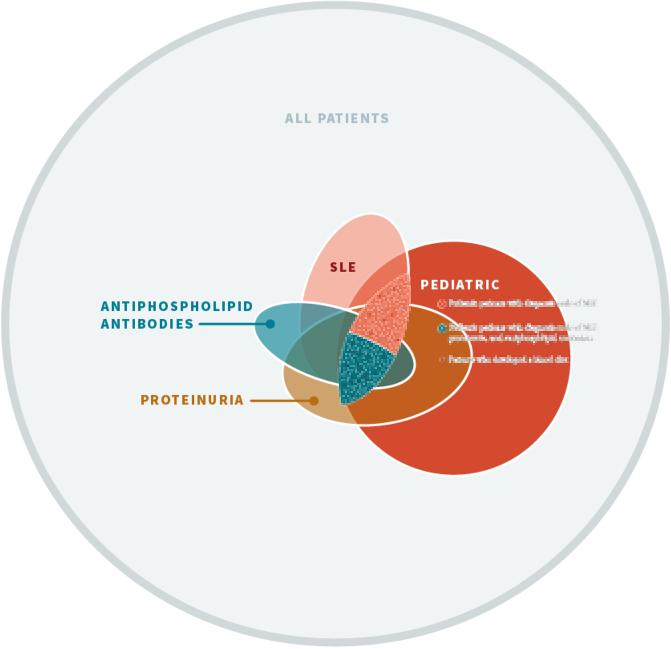
Automate **feedback loops** to **improve** model embeddings

Modeling:
Pretraining Objectives

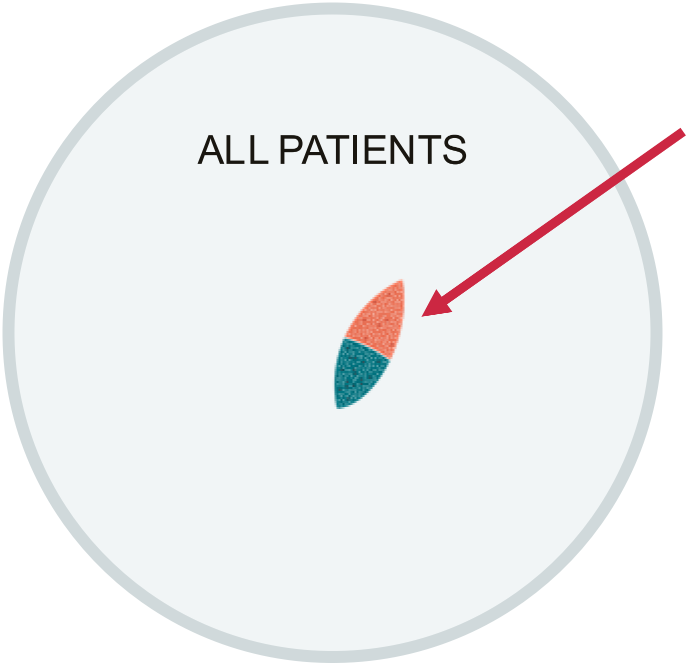
Classic Approach to Building and Patient Model

MEET LAURA

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



Classic Approaches Often Fail Due to Limited Data

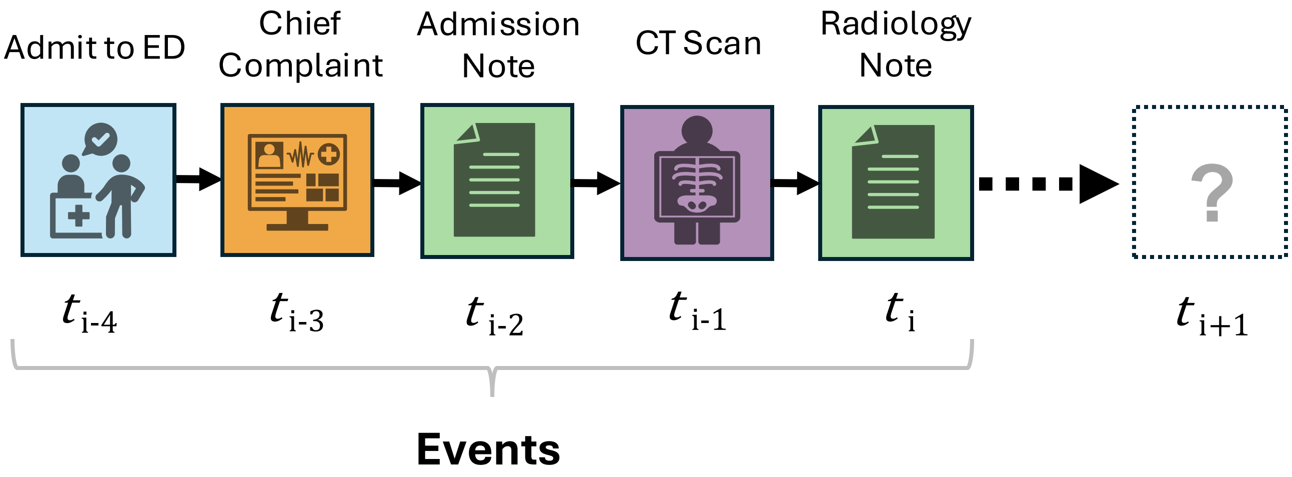


Classic medical ML approach is to train on a small **target cohort**

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

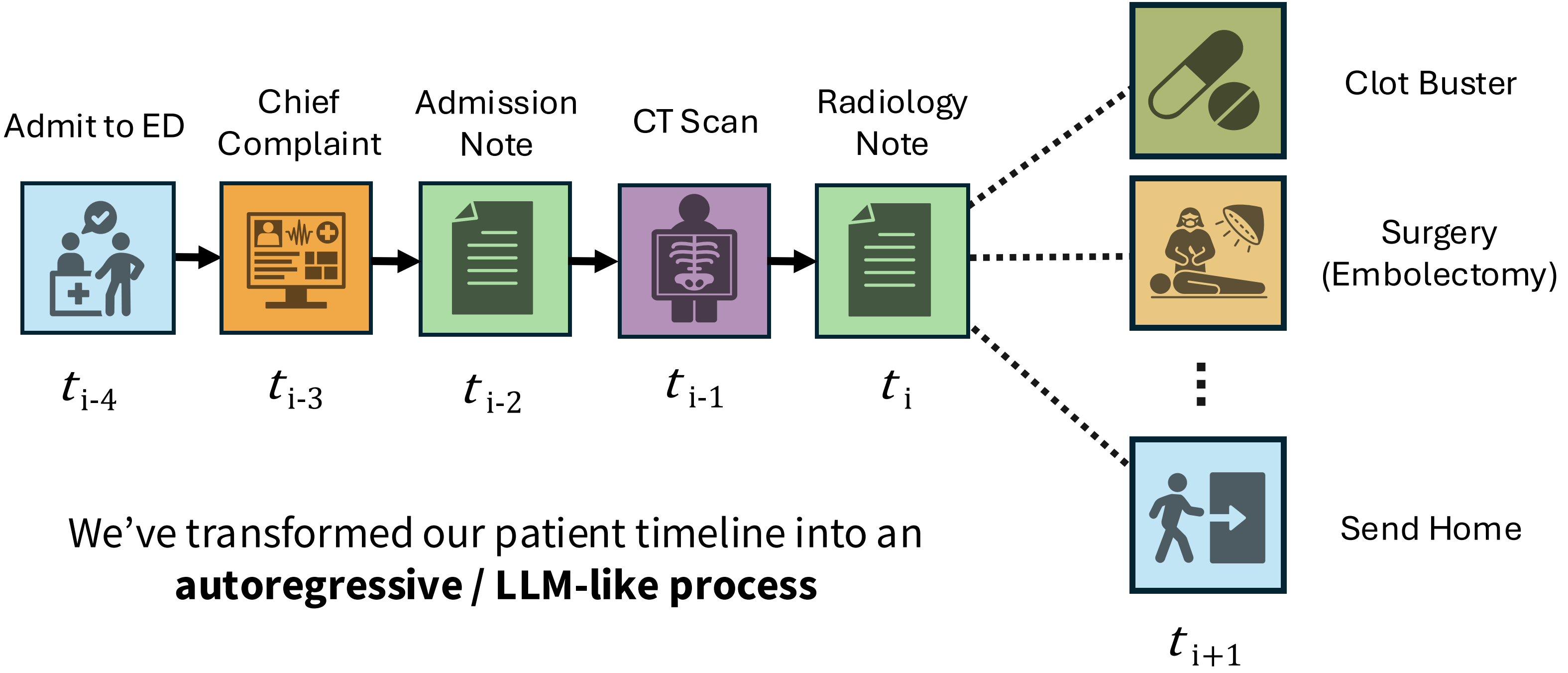
Modeling Patient Timelines for AI

PATIENT CASE: Patient **presents to ED** with sudden onset **shortness of breath**, **pleuritic chest pain**, and **tachycardia**. Concern for **pulmonary embolism**.



Modeling Patient Timelines for AI

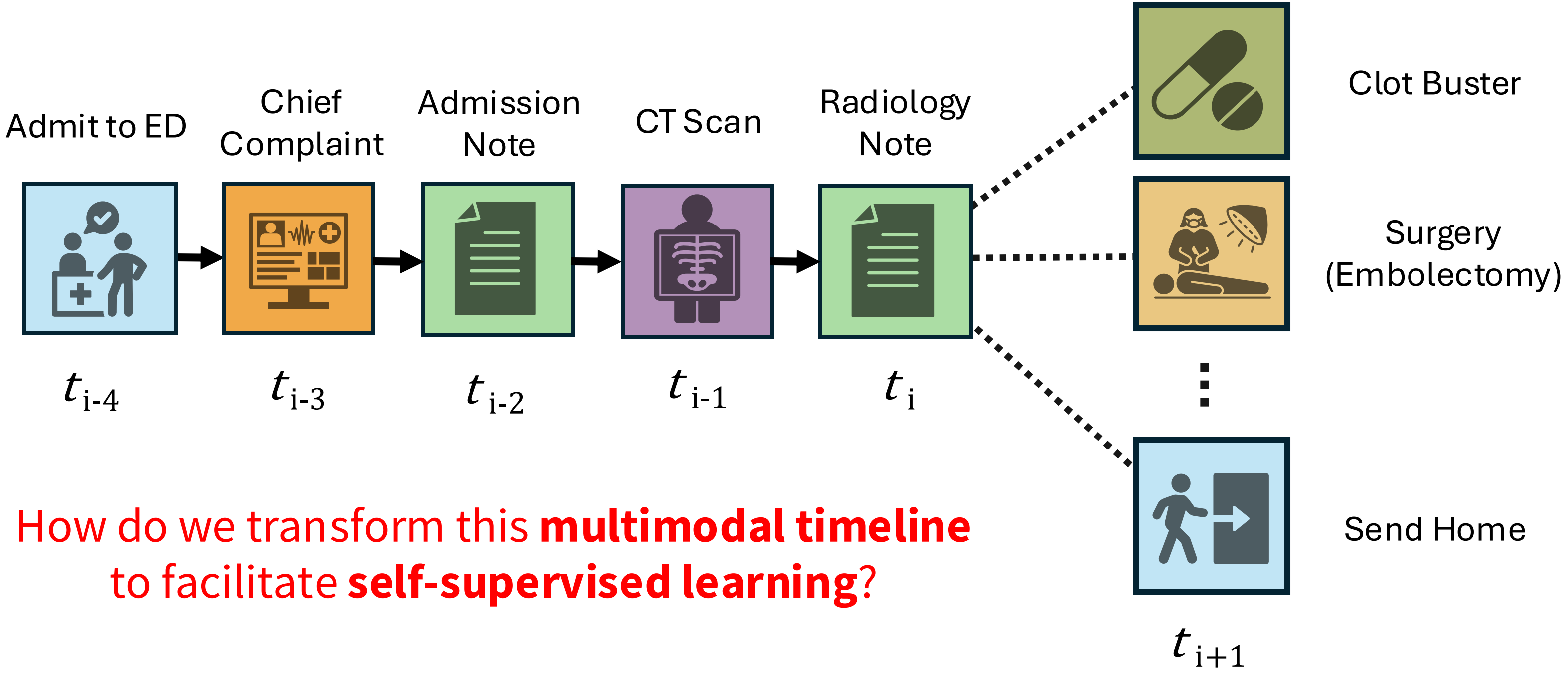
PATIENT CASE: Patient **presents to ED** with sudden onset **shortness of breath**, **pleuritic chest pain**, and **tachycardia**. Concern for **pulmonary embolism**.



We've transformed our patient timeline into an **autoregressive / LLM-like process**

Modeling Patient Timelines for AI

Hypothesis: A model that accurately **predicts future health states**, based on patient history, **encompasses many proposed use cases of medical AI**

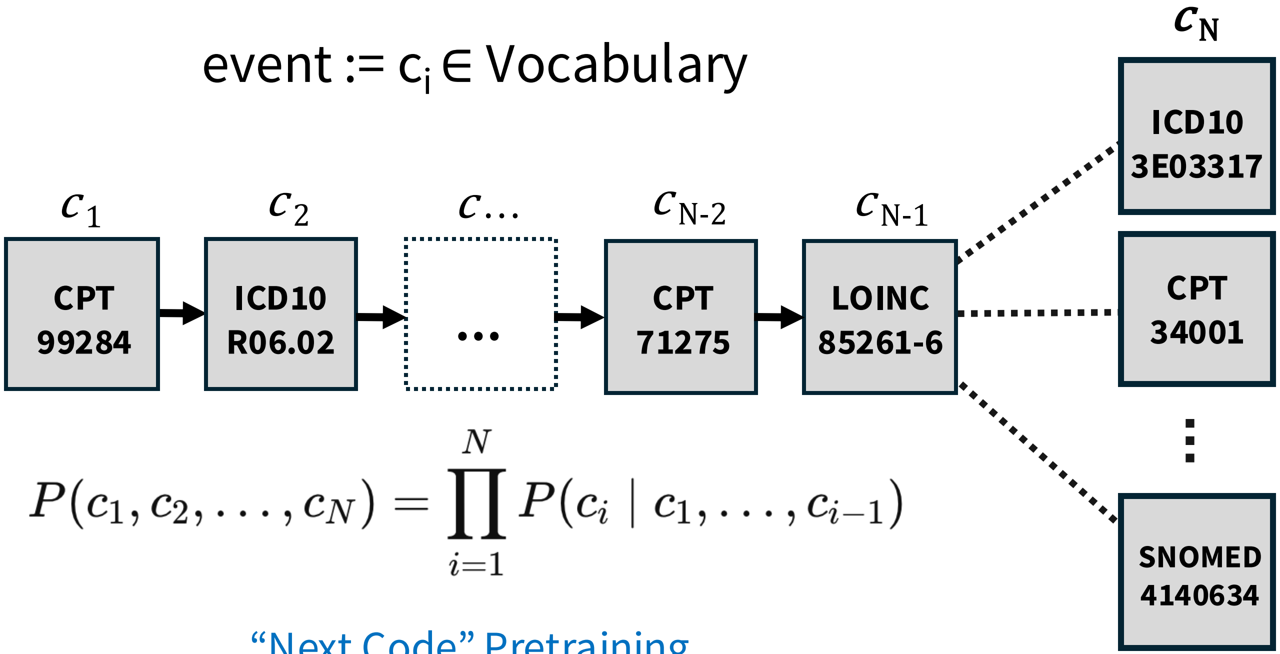


How do we transform this **multimodal timeline** to facilitate **self-supervised learning**?

Modeling Structured EHR Timelines

Map events to ontologies to define a “language” based on medical codes

event := $c_i \in$ Vocabulary

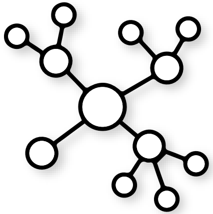


$$P(c_1, c_2, \dots, c_N) = \prod_{i=1}^{N} P(c_i \mid c_1, \dots, c_{i-1})$$

“Next Code” Pretraining

Ontology Mapping

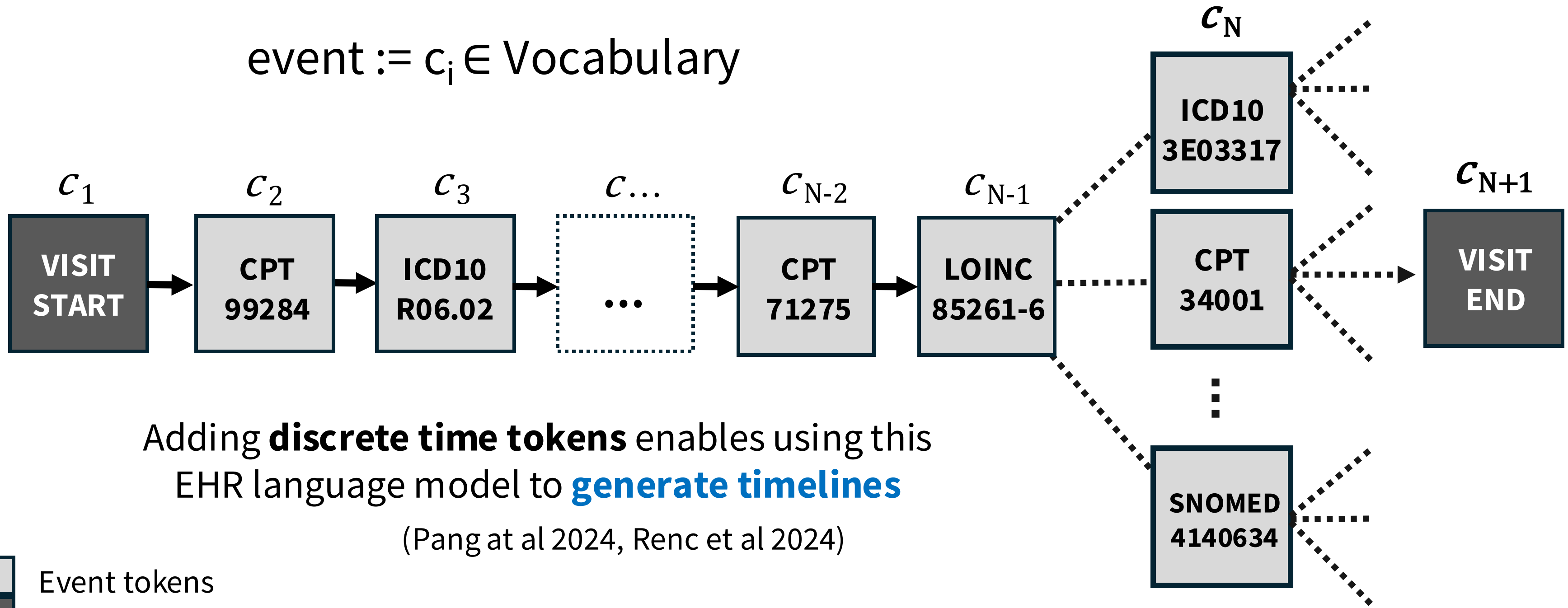
Admit to ED → **CPT 99284**
Shortness of Breath → **ICD10 R06.02**
Chest Pain → **ICD10 R07.1**
Tachycardia → **ICD10 R00.0**
Admission Note → **LOINC 47039-3**
CT Scan → **CPT 71275**
Radiology Note → **LOINC 85261-6**
Thrombolytic → **ICD10 3E03317**
Embolectomy → **CPT 34001**
Discharge to Home → **SNOMED 4140634**



Modeling Structured EHR Timelines

Map events to ontologies to define a "language" based on medical codes

event := $c_i \in$ Vocabulary

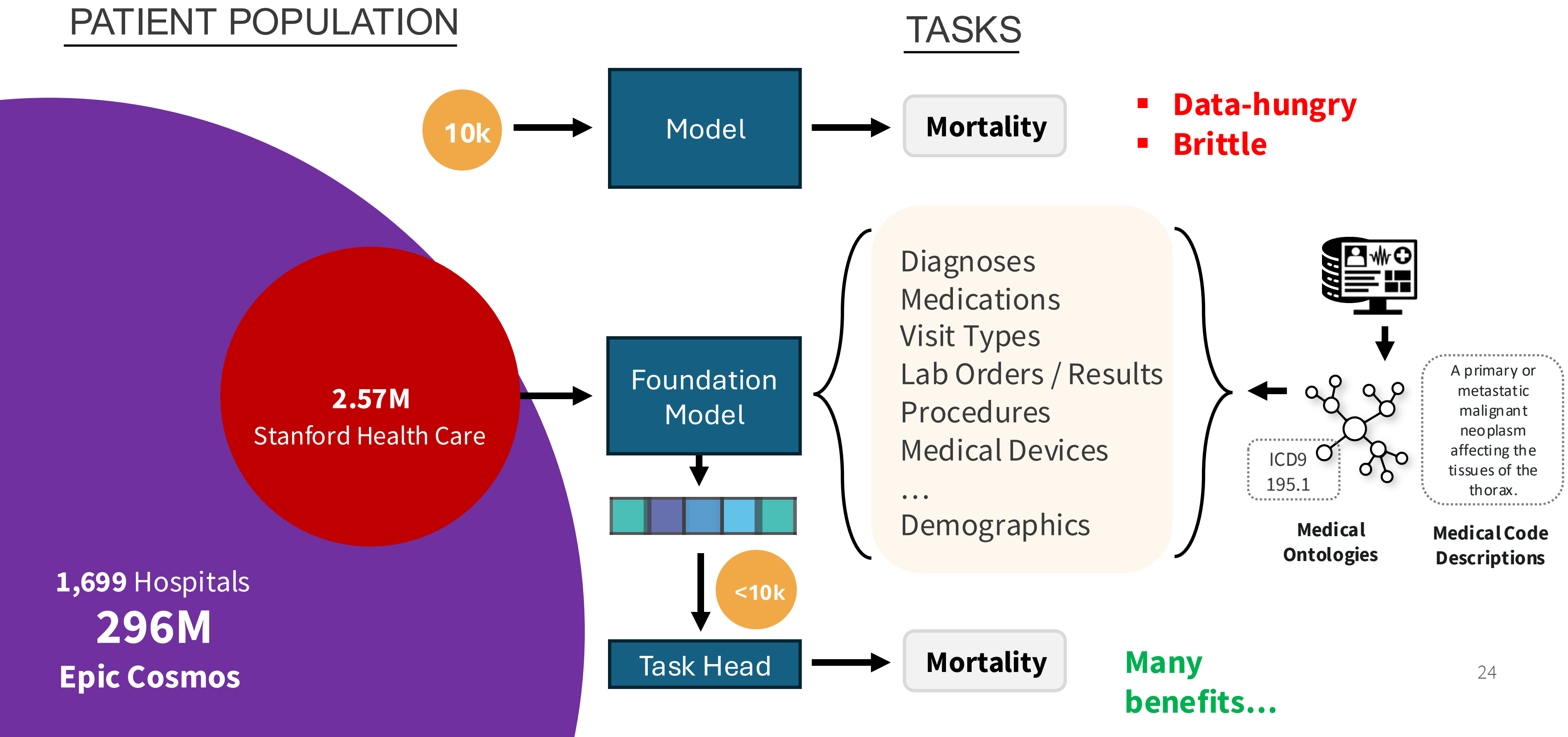


Adding **discrete time tokens** enables using this EHR language model to **generate timelines**

(Pang at al 2024, Renc et al 2024)

Event tokens
Time tokens

Self-Supervised Training of an EHR Foundation Model



Self-Supervised Pretraining Objectives for Structured Event Data

BERT-Style (Masked Language Modeling)

- BEHRT (Li et al. 2020)
- MedBERT (Rasmy et al. 2021)
- CEHR-BERT (Pang et al 2021)
- ClaimPT (Zeng et al. 2022)
- *et alia*

GPT-Style (Autoregressive)

- CLMBR (Steinberg et al. 2020)
- TransformEHR (Yang et al. 2023)
- CEHR-GPT (Pang et al 2024)
- ETHOS (Renc et al. 2024)

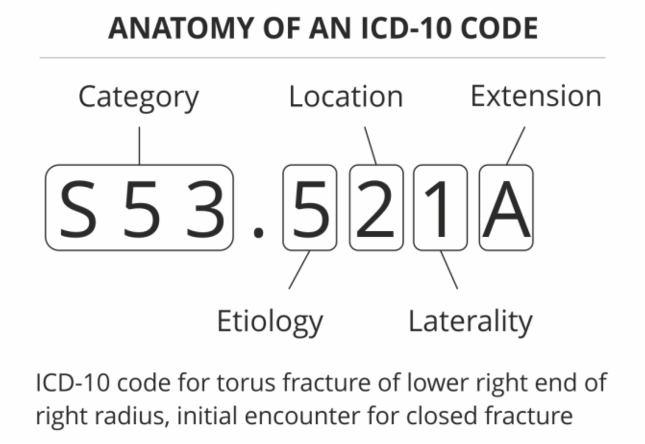
Time-to-Event

- MOTOR (Steinberg et al. 2024)

Won't talk about masked language modeling

Will focus on structured (medical code) models

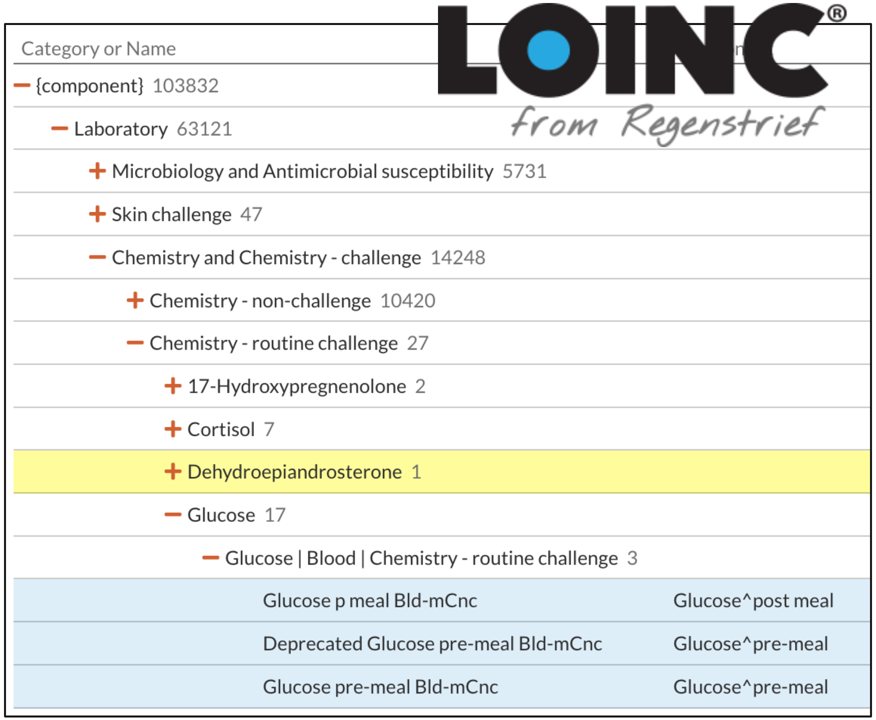
Structured Data: Medical Vocabularies



https://blogs.halodoc.io/

- Controlled Vocabularies
- **Knowledge Graphs**

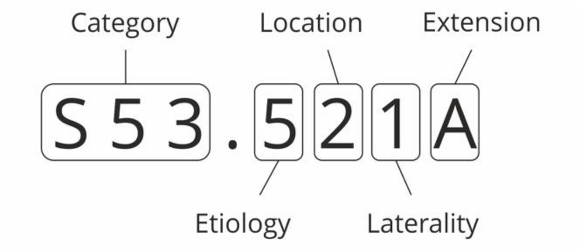
$\text{code}_i \in \text{Vocabulary}$



More Like NLP Now, but Key Differences!

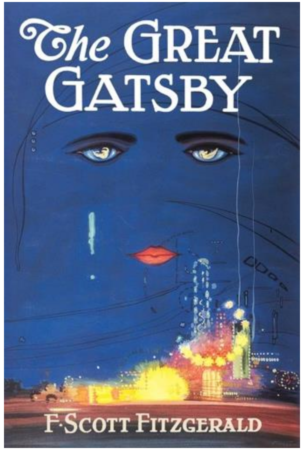
Tokenization / Vocabulary

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



Sequence Properties

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



50% Patients
>= 68k tokens

GPT-Style (Autoregressive)

- CLMBR (Steinberg et al. 2020)
- TransformEHR (Yang et al. 2023)
- CEHR-GPT (Pang et al 2024)
- ETHOS (Renc et al. 2024)

Self-Supervised Pretraining in Natural Language

S = Where are we going

Previous words (Context)

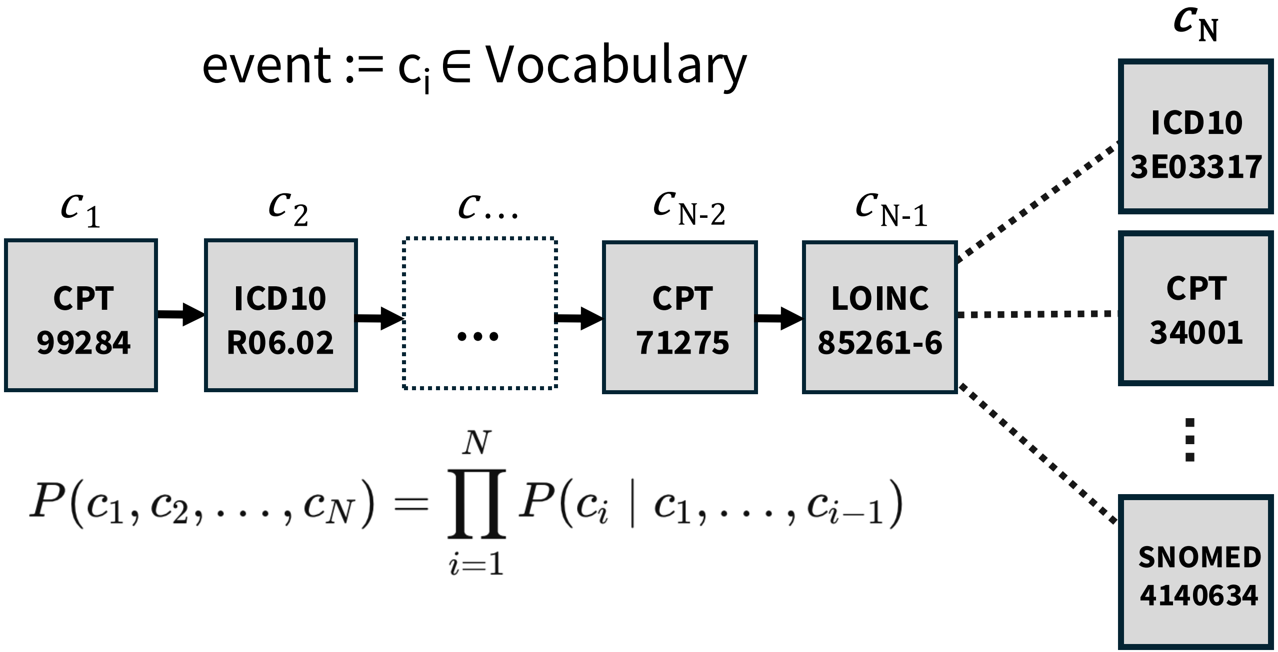
Word being predicted

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

$$
\begin{aligned}
P_{(w_1, w_2, \ldots, w_n)} &= p(w_1)p(w_2|w_1)p(w_3|w_1, w_2)...p(w_n|w_1, w_2, .., w_{n-1}) \\
&= \prod_{i=1}^{n} p(w_i|w_1, ..., w_{i-1})
\end{aligned}
$$

Next Code Pretraining

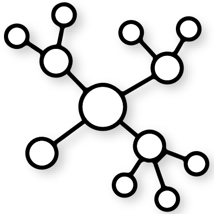
event := $c_i \in$ Vocabulary



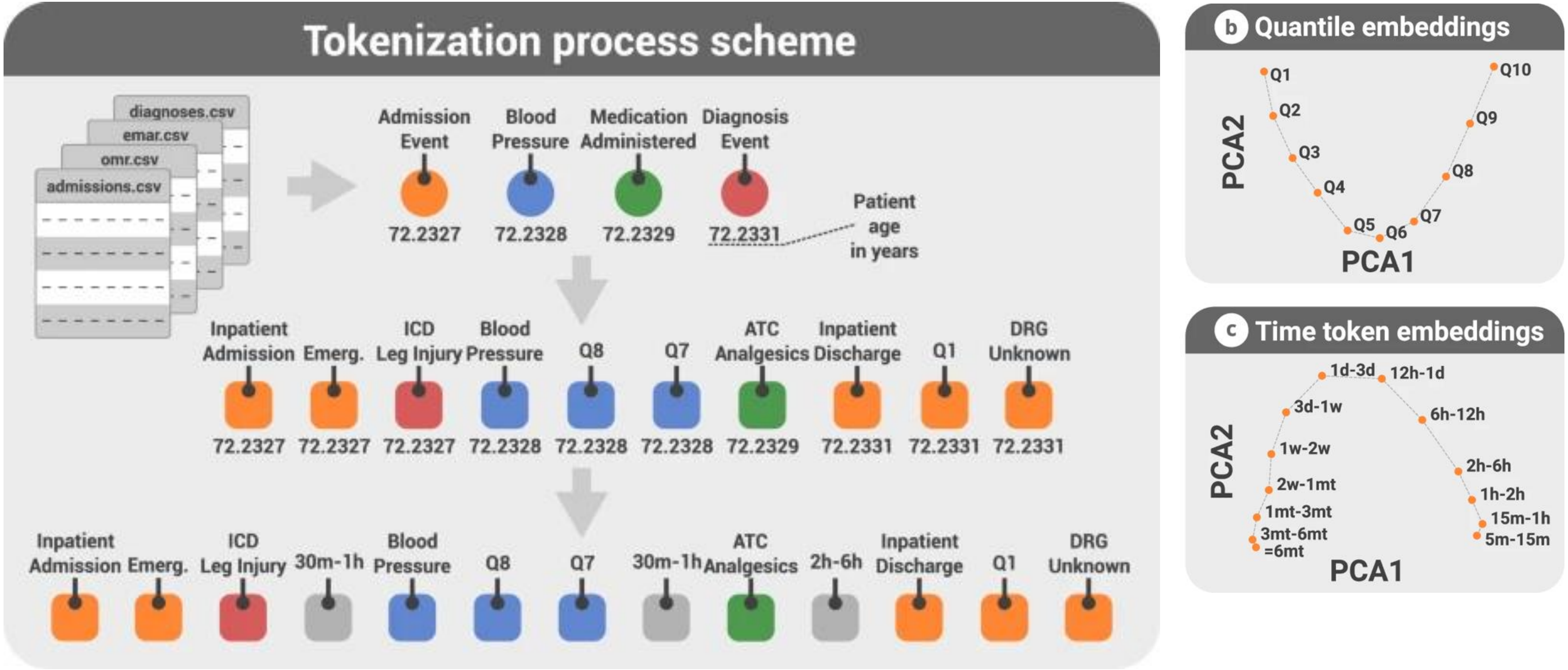
$$P(c_1, c_2, \dots, c_N) = \prod_{i=1}^{N} P(c_i \mid c_1, \dots, c_{i-1})$$

Ontology Mapping

Admit to ED → **CPT 99284**
Shortness of Breath → **ICD10 R06.02**
Chest Pain → **ICD10 R07.1**
Tachycardia → **ICD10 R00.0**
Admission Note → **LOINC 47039-3**
CT Scan → **CPT 71275**
Radiology Note → **LOINC 85261-6**
Thrombolytic → **ICD10 3E03317**
Embolectomy → **CPT 34001**
Discharge to Home → **SNOMED 4140634**

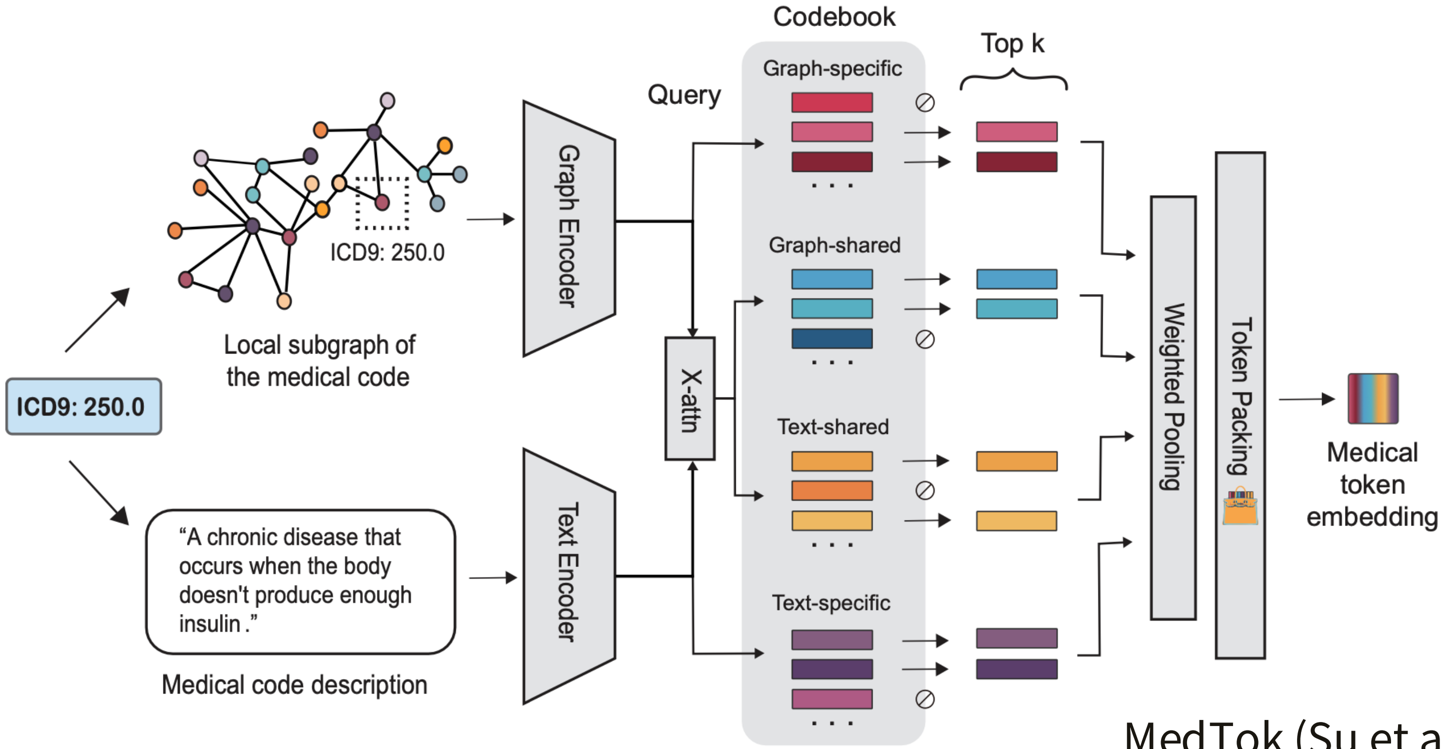


Tokenization



ETHOS (Renc et al. 2024)

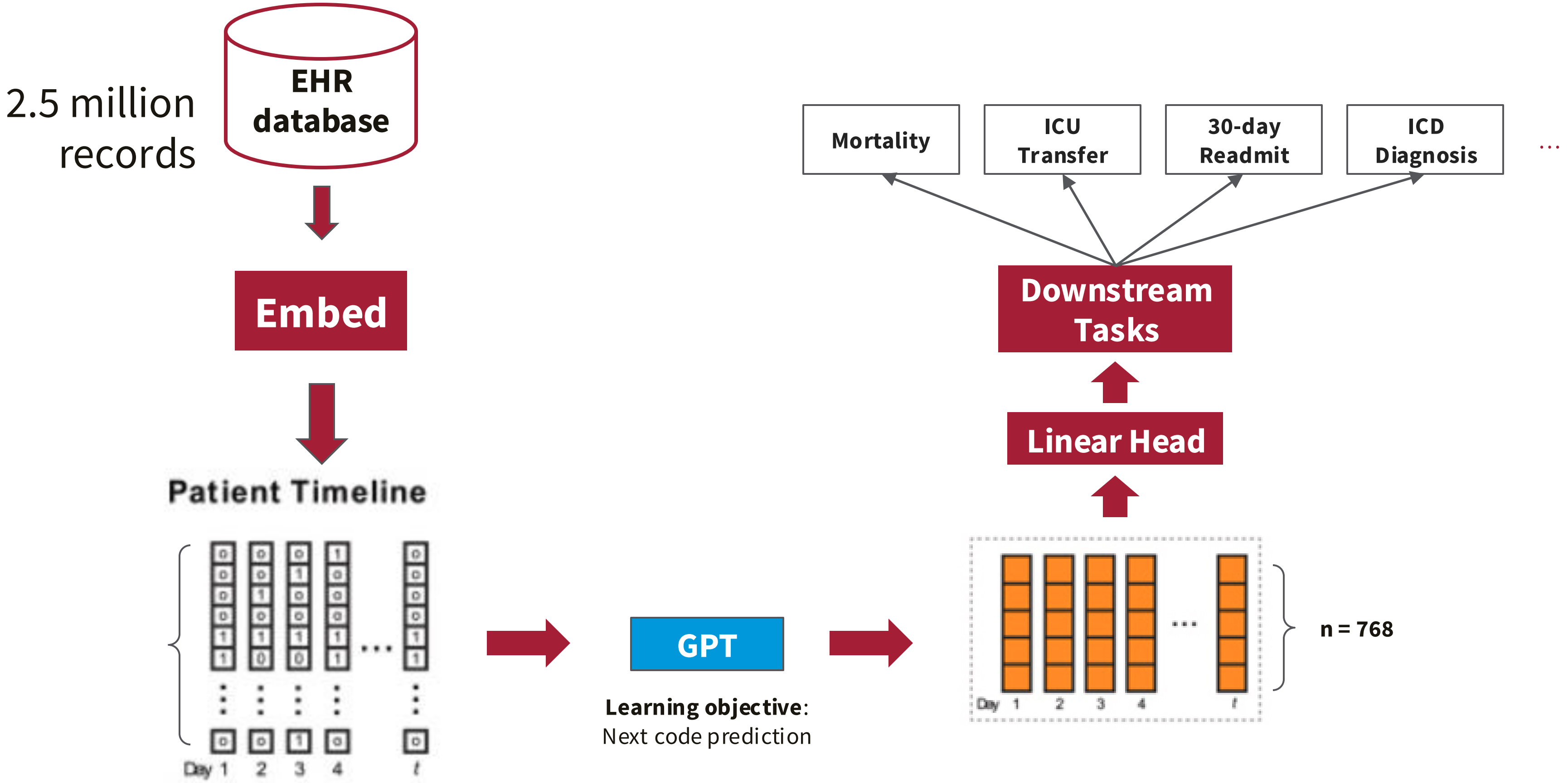
Generalized Tokenizer



MedTok (Su et al. 2025)

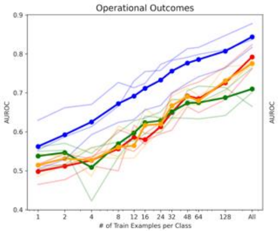
Drop-in Replacement

GPT-based Approach



Validating Benefits of EHR Foundation Models

Data Efficiency



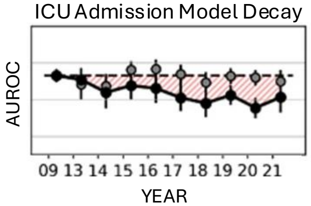
SOTA **few-shot learning**

SOTA **overall performance**

(Wornow et al. 2023)

(Steinberg et al. 2020)

Robustness



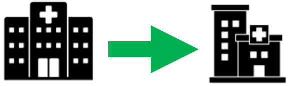
Improved robustness to **temporal distribution shifts**

(Guo et al. 2023)

Improved performance across key **subgroups** (pediatrics)

(Lemmon et al. 2023)

Cross-Site Adaptability



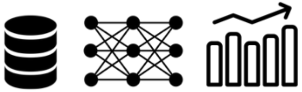
Hospital A → Hospital B

Transfer **pretrained models** across hospitals

Require **up to 90% less** pretraining data

(Guo et al. 2024)

Reproducible EHR Benchmarking

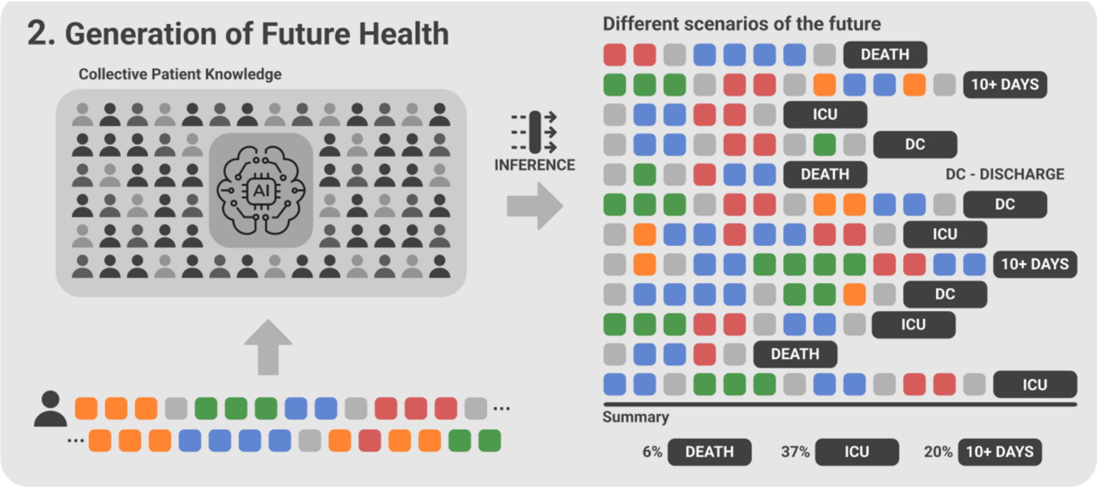


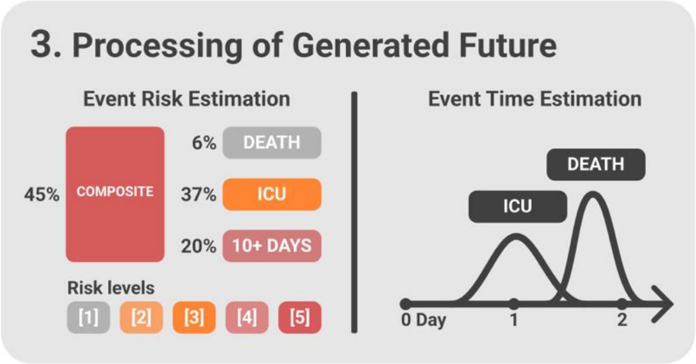
First **externally verifiable** evaluation of **EHR foundation models** on longitudinal data

(Wornow et al. 2025)
(Arnrich et al. 2024)
(Steinberg et al. 2024)
(Wornow et al. 2023)
(Huang et al. 2023)

Publication Venue
- Medical / Informatics
- Computer Science

Zero-Shot Patient Classification





ETHOS samples from model generations to estimate future event risk

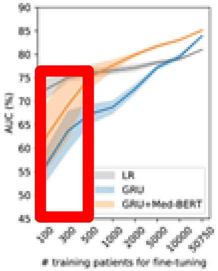
Time-to-Event Modeling

Data (Label) Efficiency of EHR Foundation Models

Label Efficiency: How many **labeled examples** are needed to train a high-performing model?

BERT-Style (Masked Language Modeling)

- BEHRT (Li et al. 2020)
- MedBERT (Rasmy et al. 2021)
- CEHR-BERT (Pang et al 2021)
- ClaimPT (Zeng et al. 2022)
- *et alia*

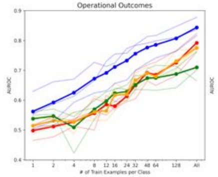


MedBERT

- Trained on **28M patients**
- Performance with **< 500 examples worse than logistic regression**

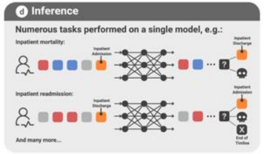
GPT-Style (Autoregressive)

- CLMBR (Steinberg et al. 2020)
- TransformEHR (Yang et al. 2023)
- CEHR-GPT (Pang et al 2024)
- ETHOS (Renc et al. 2024)



CLMBR

- Trained on **2.57M patients** (3.5B tokens)
- SOTA **few-shot** learning using **embeddings**



ETHOS

- Trained on **200k patients** (MIMIC-VI)
- **Zero-shot** abilities using **generation**

Autoregressive Modeling at Smaller Scales

Autoregressive LLMs can capture long-distance dependencies given **sufficient data and parameters**

Natural Language

≥ 7B parameters
≥ 500B-1T tokens

EHR

143M parameters
3.5B tokens

285x
less data

Can we train a **small, data-constrained** EHR foundation model to learn embeddings that capture more information about the future?

Key Concepts in Time-to-Event Modeling

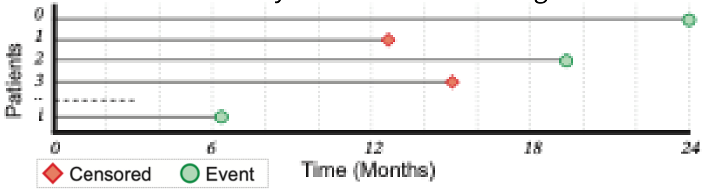
Model the **time until an event occurs** (e.g., death) while accounting for **censoring**

Censoring

Event times are **not fully observed by end of a study period**

$\boxed{(X_i, T_i)}$ BIASE D (X_i, T_i, δ_i) $\delta_i = \begin{cases} 1 & \text{event observed} \\ 0 & \text{censored} \end{cases}$

Mortality Event Time Censoring

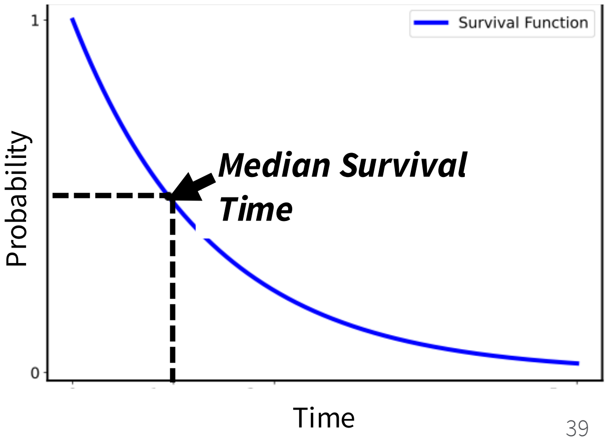


Survival Function

The probability that an event has not occurred as of time t

$$S(t) = \Pr(T > t)$$

Survival Curve



Hazard Rate Function

Instantaneous risk of an event at time t, given survival up to t

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}$$

Event's "speed" at each moment

$$S(t) = \exp\left(-\int_0^t h(u)\,du\right)$$

Survival depends on cumulative hazard over time

Learn a patient representation $R_i = f_\theta(X_i)$ for estimating **personalized hazard rates**

Intuition Behind the Pretraining Objective

Hypothesis: Multi-task learning (MTL) will capture generalizable TTE features

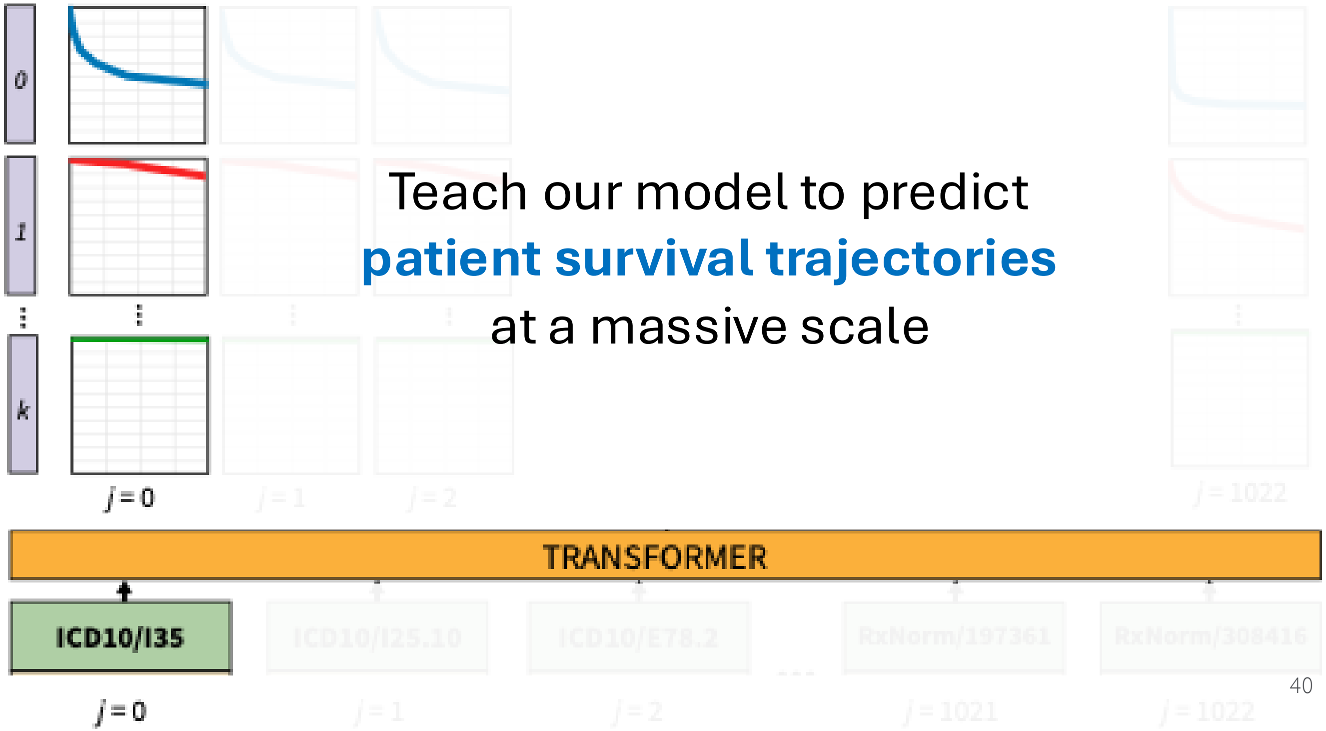
Select k TTE pretraining tasks

$k \leq$ 16,392

Event j

Shared Decoder

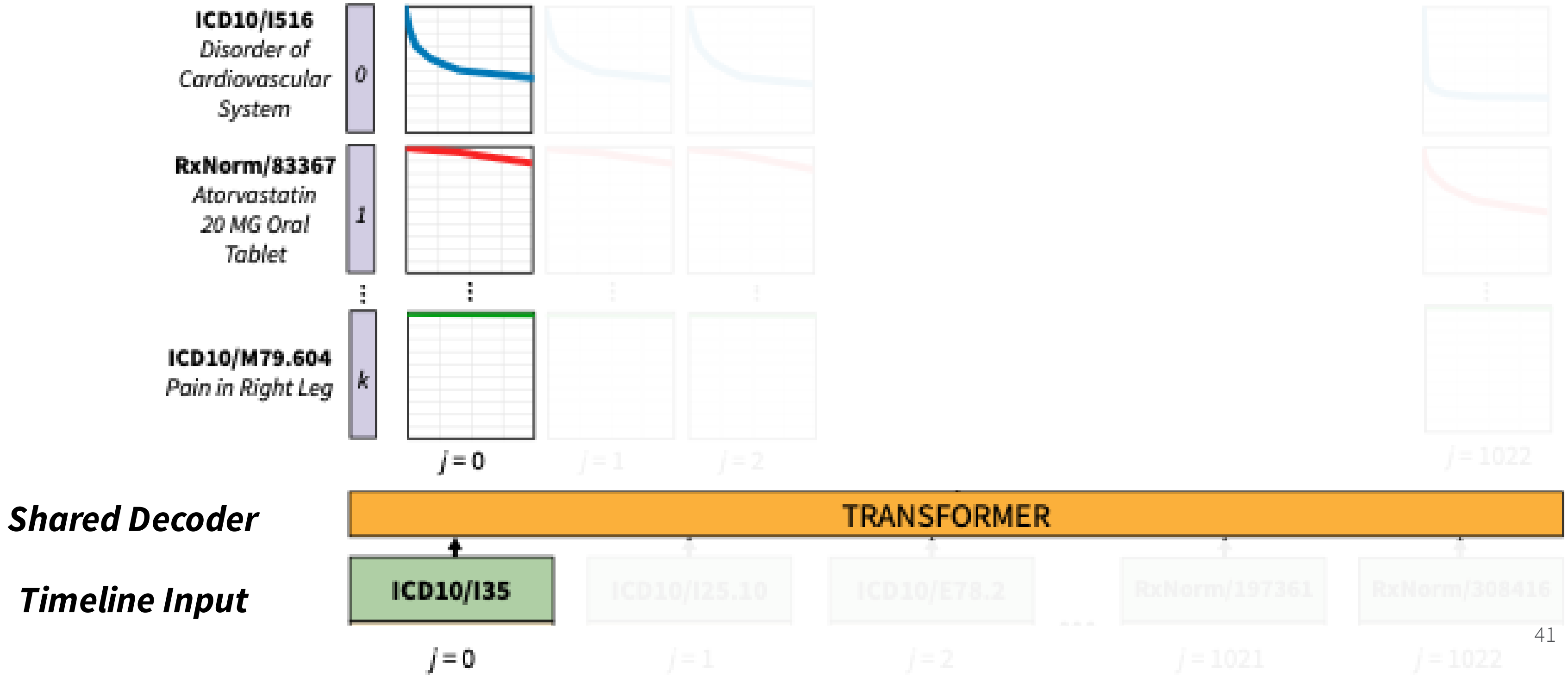
Timeline Input



Teach our model to predict **patient survival trajectories** at a massive scale

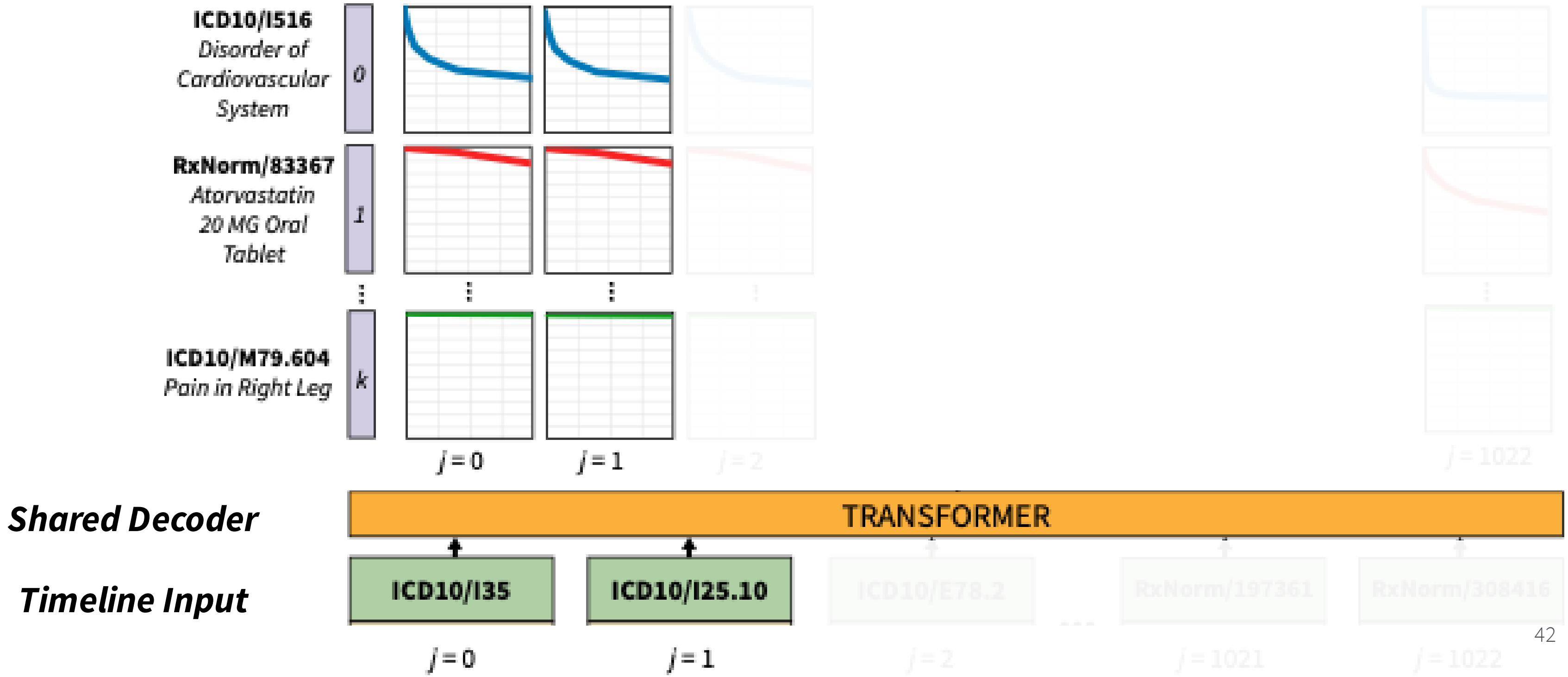
Intuition Behind the Pretraining Objective

Hypothesis: Multi-task learning (MTL) will capture generalizable TTE features



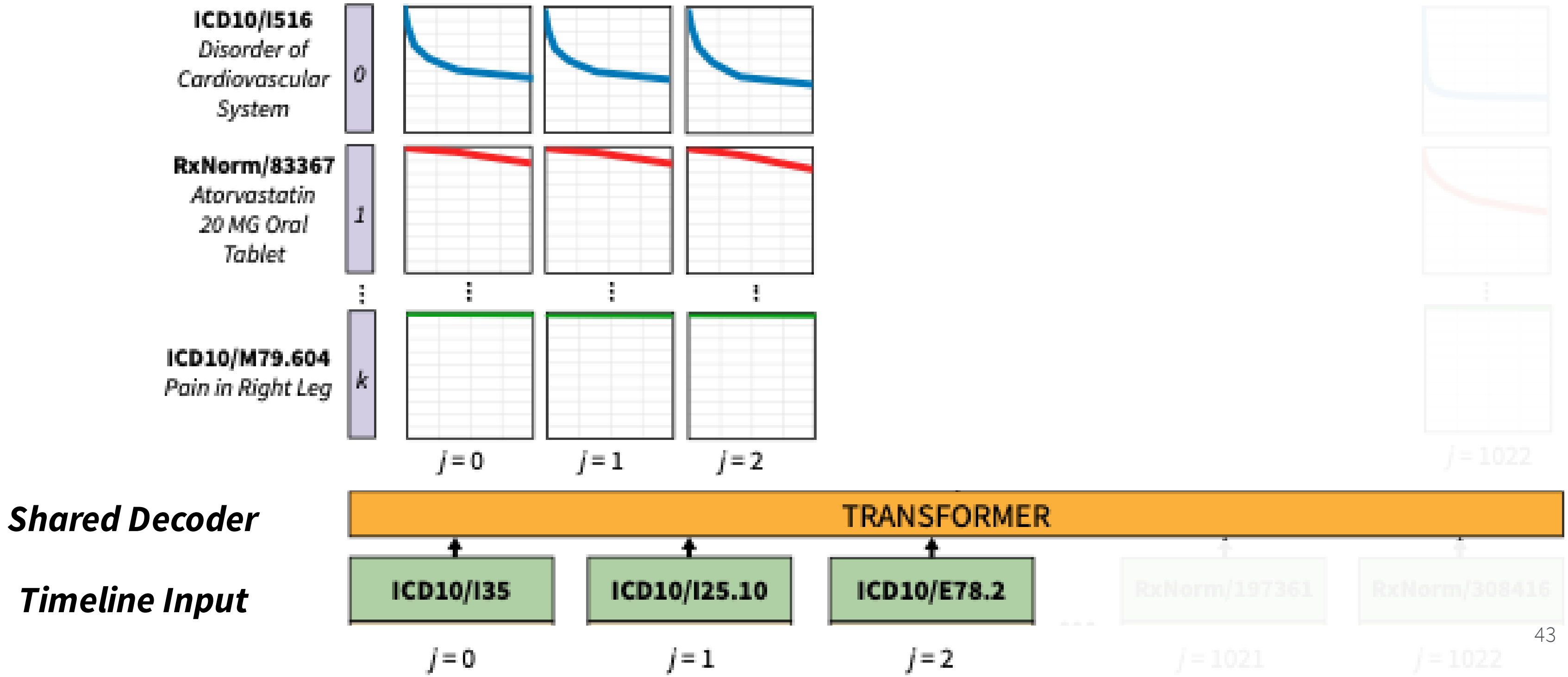
Intuition Behind the Pretraining Objective

Hypothesis: Multi-task learning (MTL) will capture generalizable TTE features



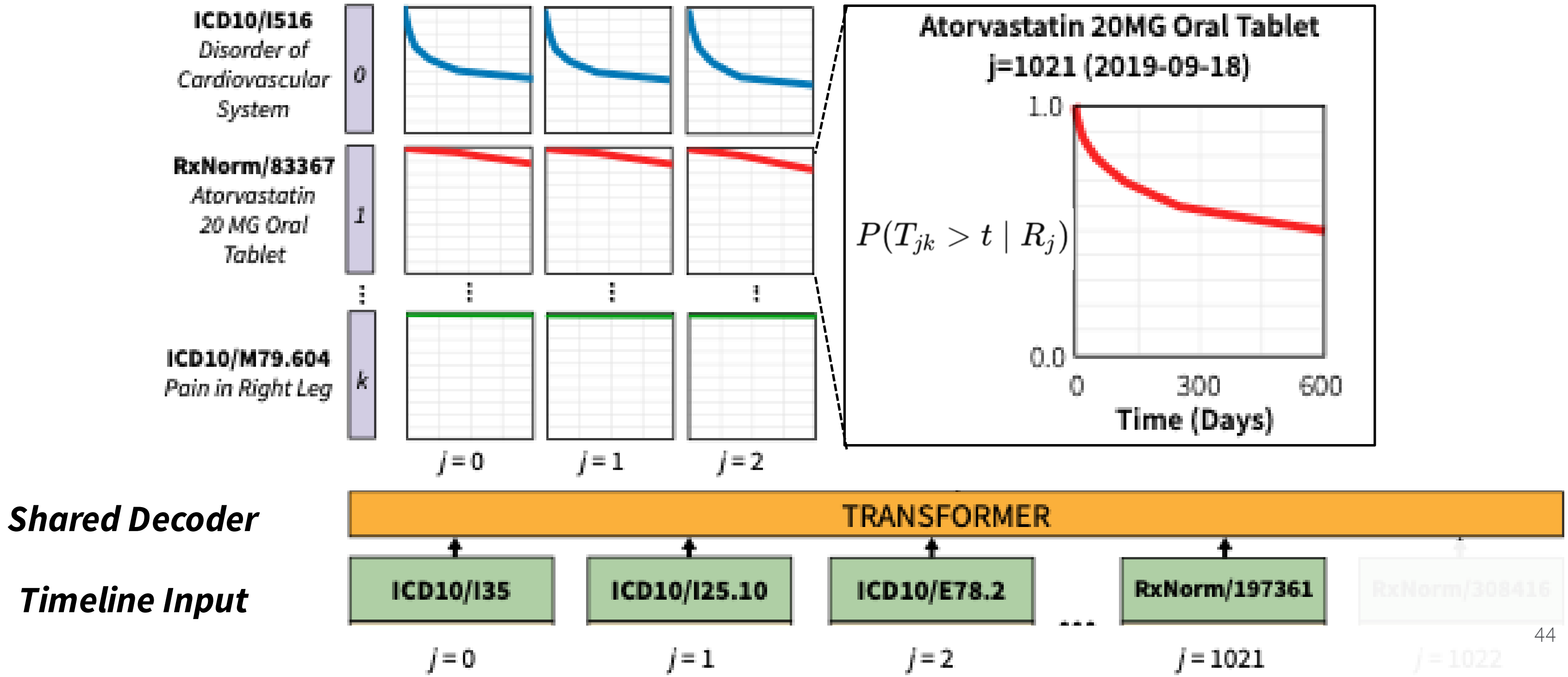
Intuition Behind the Pretraining Objective

Hypothesis: Multi-task learning (MTL) will capture generalizable TTE features



Intuition Behind the Pretraining Objective

Hypothesis: Multi-task learning (MTL) will capture generalizable TTE features

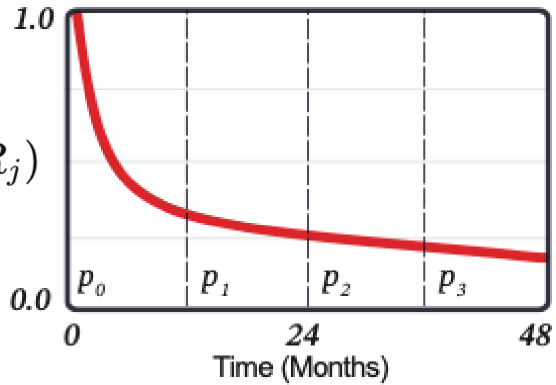


Pretraining Objective

Deep Piecewise Exponential Model

- Partition time into **pieces** for more expressive risk modeling
- For **piece** p, interval start and end time: $[S_p, E_p)$
- **Hazard rate** is constant within this interval

$P(T_{jk} > t \mid R_j)$



For a patient with **event j, task k, and piece p**

Piecewise Hazard Function

t is within piece p

$$h_{jk}(t) = \sum_{p=1}^{P} I(S_p \leq t < E_p)\, \lambda_{jkp}$$

hazard rate for piece p

Survival Function

$$S_{jk}(t) = \prod_{p=1}^{P} \exp\left(-\lambda_{jkp}\left(\min(t, E_p) - S_p\right) I(t \geq S_p)\right)$$

time-independent task embedding

Hazard Rate

$$\lambda_{jkp} = \exp(W_p R_j \cdot \hat{\beta}_k)$$

patient representation as of j

TRANSFORMER f_θ

piece-specific linear projection

Pretraining Objective

Loss Function

Minimize the negative log-likelihood of the observed event times across all tasks and time pieces

$$\min_{\Theta} \ \mathcal{L}(\Theta) = -\underbrace{\sum_{j,k}}_{\text{all events and tasks}} \sum_{p=1}^{P} \Big[\underbrace{\delta_{jkp}\left(\log \lambda_{jkp} - \lambda_{jkp} U_{jkp}\right)}_{\text{event happens in piece } p} + \underbrace{\left(1-\delta_{jkp}\right)\left(-\lambda_{jkp} U_{jkp}\right)}_{\text{no event in piece } p}\Big]$$

U represents the amount of time an event is at risk within a given time interval

Datasets & Tasks

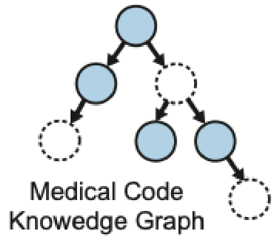
Datasets

STANFORD STARR-OMOP (EHR)

2.7M Patients

3.5B Events

Pretraining Tasks



Entropy-Ranked Vertex Cover for Task Selection

Intuition: We pick k tasks that **maximize diversity** by selecting nodes whose values are **least predictable** given their parents

$k \leq$ 16,392

Evaluation Tasks

Celiac Disease | **Stroke**

Pancreatic Cancer | **NAFLD**

Heart Attack | **Lupus**

ICD-10

Rule-based labeling

We remove these tasks from the pretraining set

CheXpert

13 Chest X-ray Findings

NLP-based

Measures generalization to labels not derived from codes

Results: MOTOR vs. Baselines

MOTOR-Scratch (no pretraining) largely **underperforms** compared to baselines

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

Results: MOTOR vs. Baselines

But with **pretraining...**

MOTOR-Probe & **MOTOR-Finetune** outperform **SOTA on all tasks**

Avg improvement: **+4.6%**

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**

Results: Autoregressive vs. TTE Pretraining

Overall Performance

Objective	Celiac	HA	Lupus	NAFLD	Cancer	Stroke
RSF	0.729	0.836	0.787	0.802	0.824	0.840
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**

Autoregressive beats SOTA (RSF)

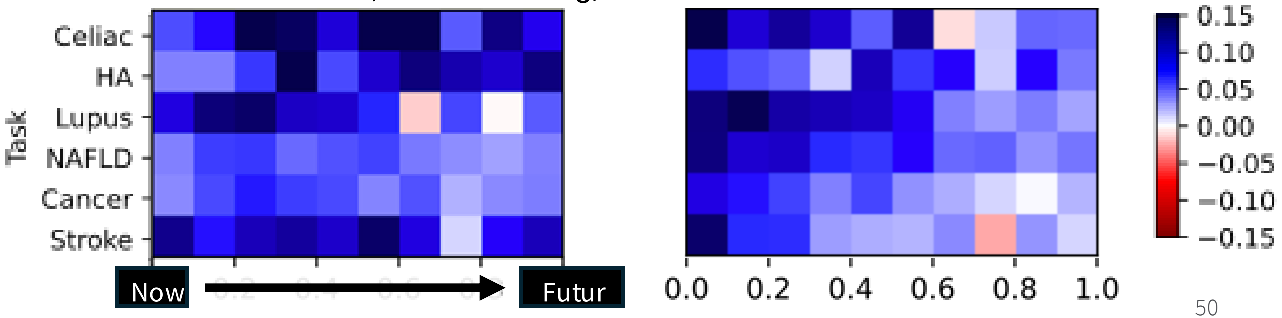
…but **TTE beats autoregressive** by **~2%**

Performance Comparison over Long Time Horizons

Pretraining is the key driver of performance

Performance Deltas of MOTOR with TTE Pretraining Versus:

MOTOR-Scratch (No Pretraining) *Random Survival Forests*



Time Horizon (Percentile of Event Times)

Results: Autoregressive vs. TTE Pretraining

Overall Performance

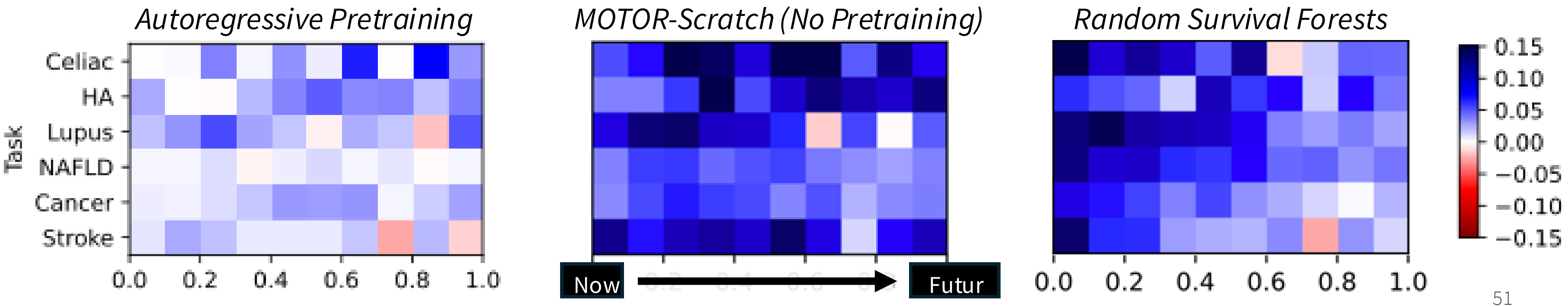
Objective	Celiac	HA	Lupus	NAFLD	Cancer	Stroke
RSF	0.729	0.836	0.787	0.802	0.824	0.840
Next Code	0.774	0.862	0.842	0.860	0.860	0.857
Time-to-Event	**0.802**	**0.887**	**0.863**	**0.864**	**0.865**	**0.875**

Autoregressive beats SOTA (RSF)

…but **TTE beats autoregressive** by **~2%**

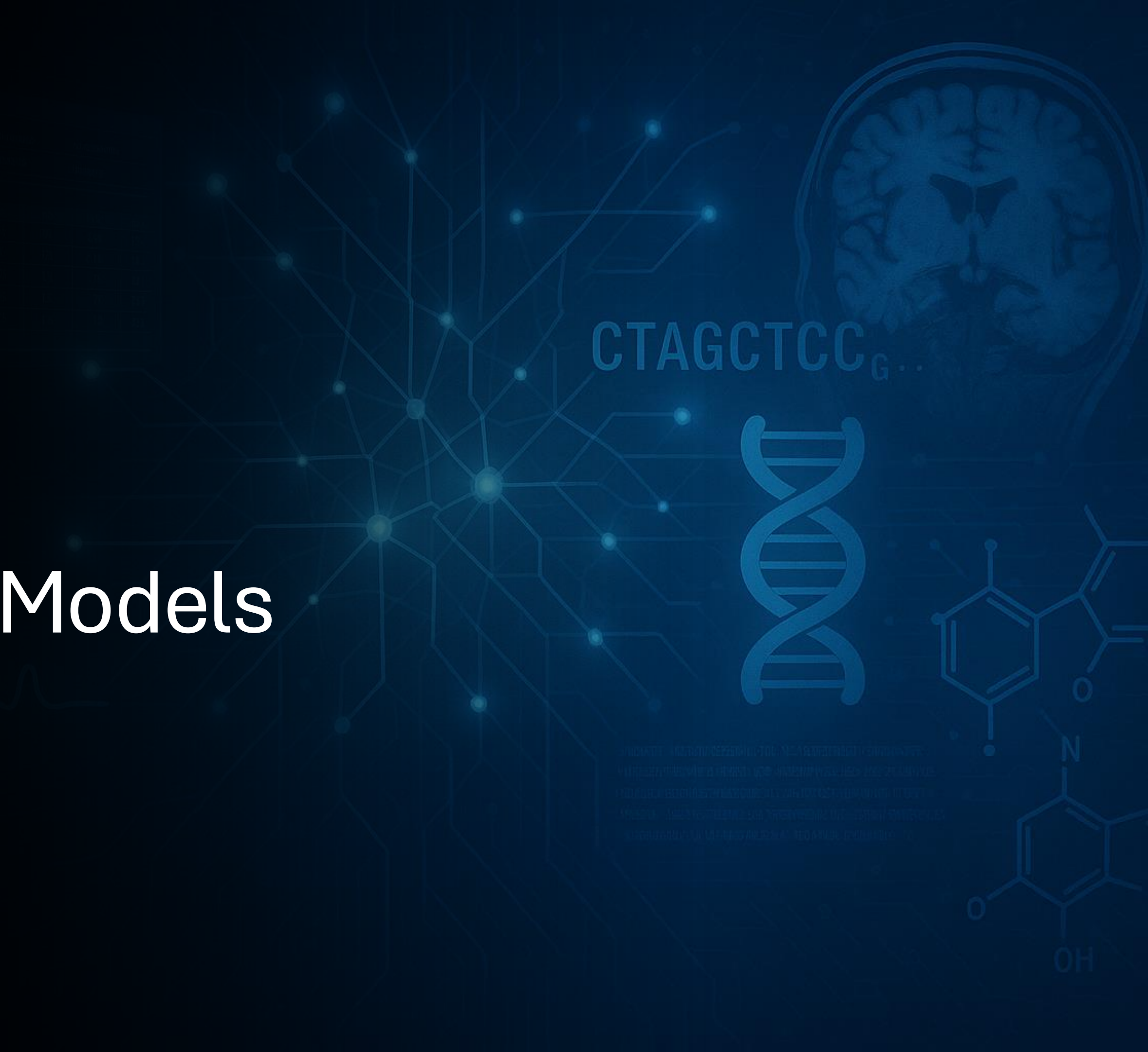
Performance Comparison over Long Time Horizons

Performance Deltas of MOTOR with TTE Pretraining Versus:



Time Horizon (Percentile of Event Times)

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

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

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

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

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

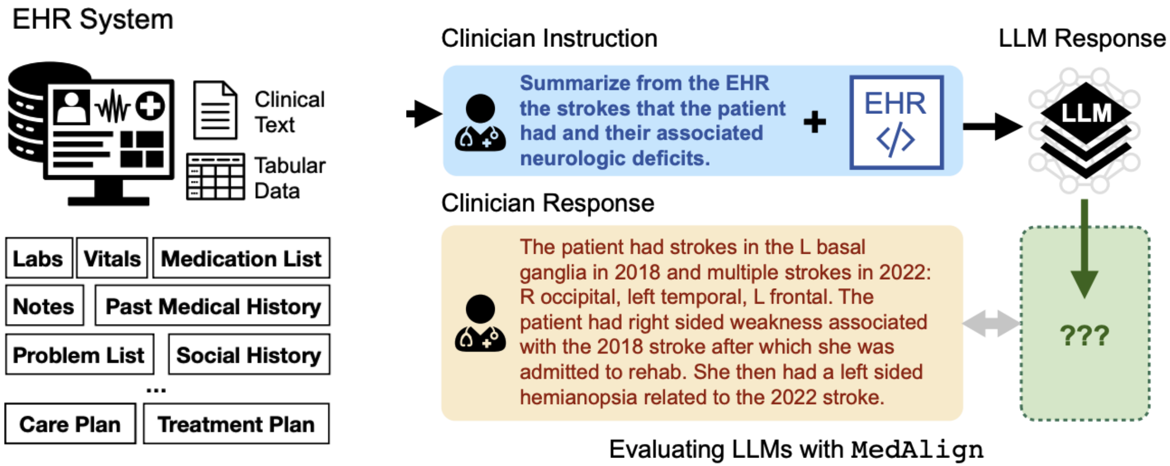
USMLE
United States Medical Licensing Exam


```
<record>
    <visit type="Emergency Room Visit" start="10/08/2018 20:00">
        <day start="10/08/2018 20:00">
            <person>
                Birth:7/19/1966
                Rac
                Gen
                Eth
                Age
                Age
            </perso
            <condit
                <co
            </condi
            <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>
```


Longitudinal Patient Timelines

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.

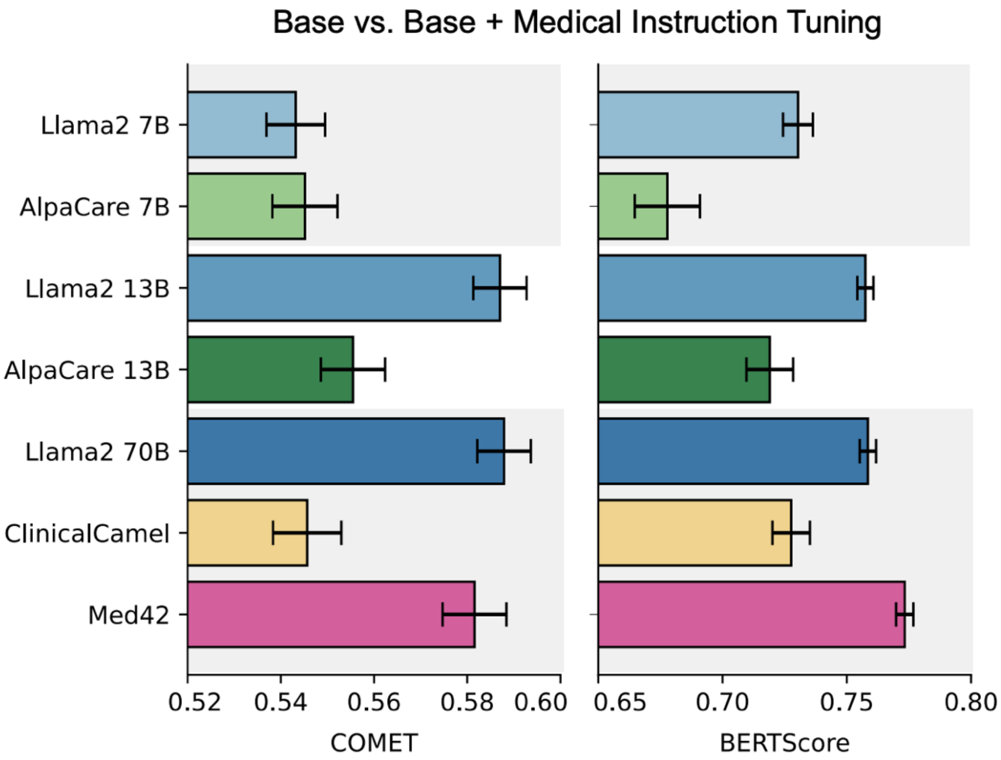
Instruction Tuning: Aligning with Clinical Needs

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!

Longitudinal, Multimodal EHR Dataset Releases

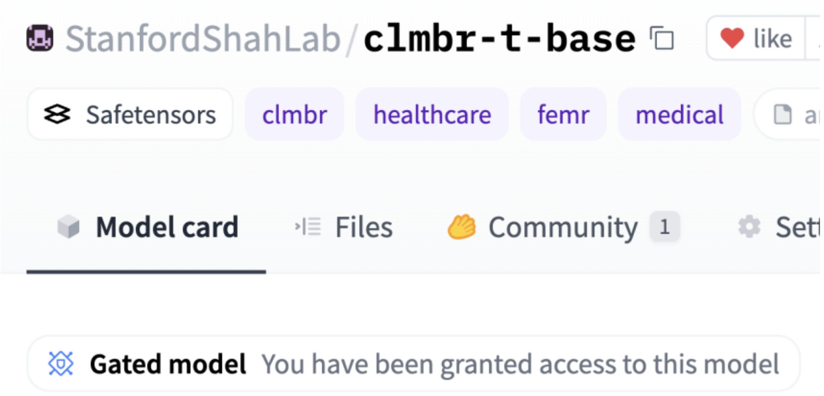
Dataset	Task	Technical Challenge	Example	Tabular	Images	Notes
EHRSHOT	Risk Stratification	Few-Shot Learning	*What is the likelihood that this patient gets a diagnosis of pancreatic cancer within the next year?*	✓	✗	✗
INSPECT	Time-to-Event Modeling	Multimodal Learning	*When is chronic pulmonary hypertension most likely to develop*	✓	✓	✓
MedAlign	Instruction Following	Long-Context Learning & Temporal Reasoning	*From this EHR, summarize the patient's history of strokes and the resulting neurologic deficits.*	✓	✗	✓

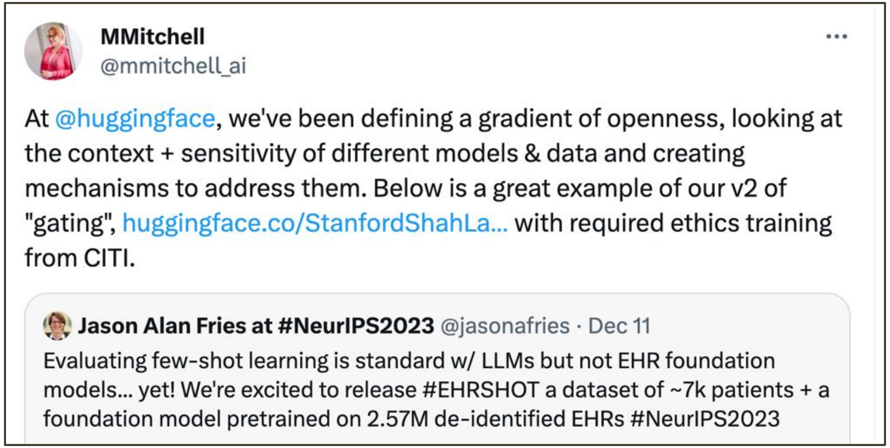
26k Patients **295M** **442k** Visits

REDIVIS

https://redivis.com/ShahLab

Enabling Open Science





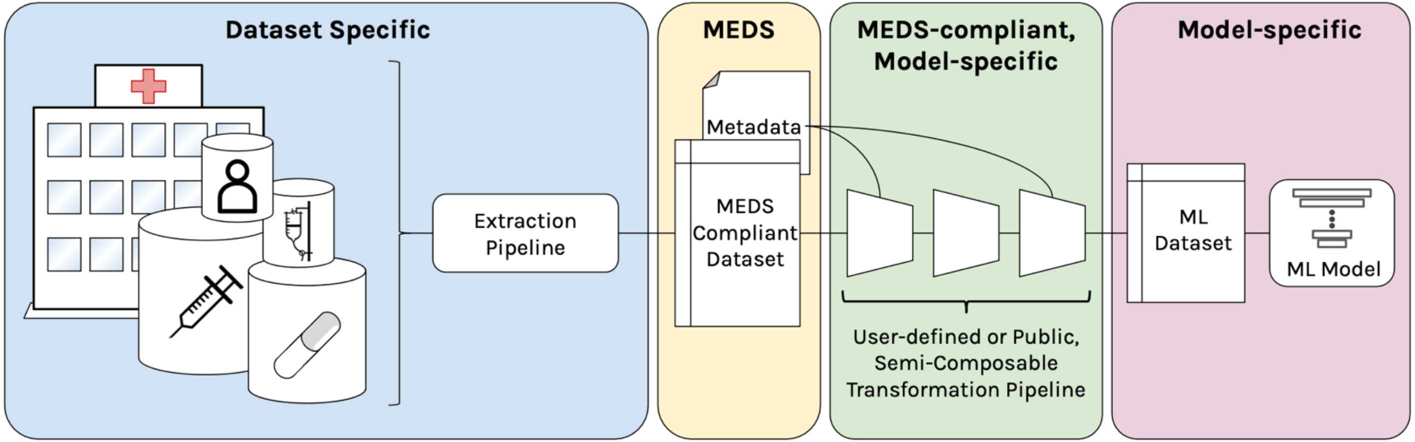
First EHR model hub release!

- 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

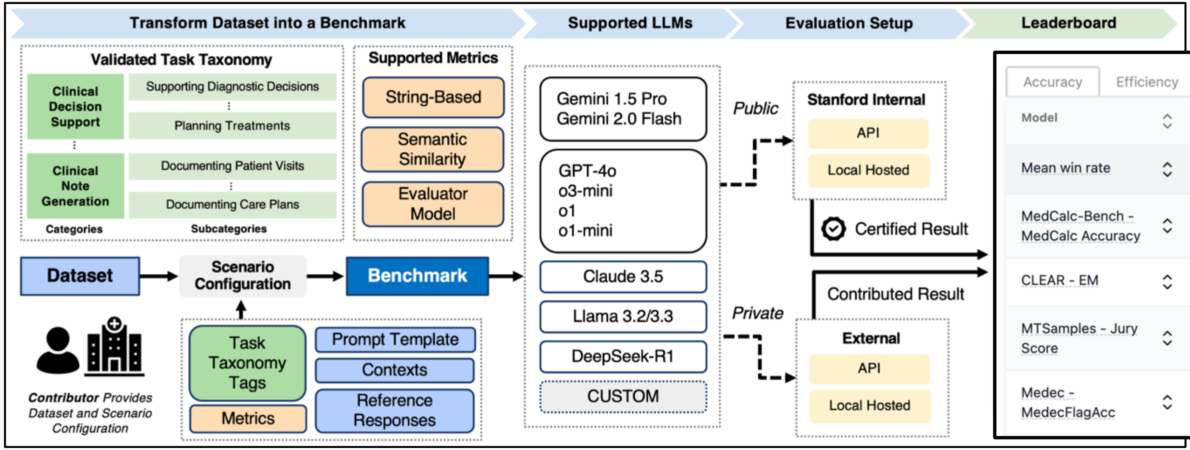
https://github.com/Medical-Event-Data-Standard/meds

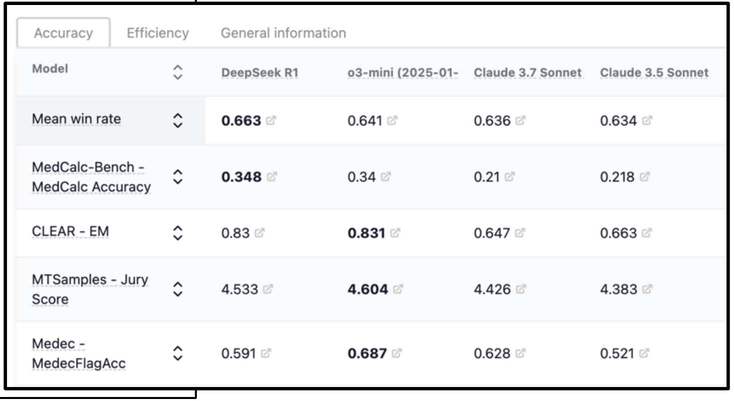
Opportunities: Datasets & Benchmarks

Stanford MedHELM

Community evaluation framework for benchmarking healthcare LLMs

https://medhelm.stanford.edu/



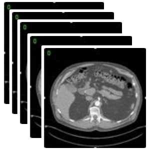


Model	DeepSeek R1	o3-mini (2025-01-	Claude 3.7 Sonnet	Claude 3.5 Sonnet
Mean win rate	**0.663**	0.641	0.636	0.634
MedCalc-Bench - MedCalc Accuracy	**0.348**	0.34	0.21	0.218
CLEAR - EM	0.83	**0.831**	0.647	0.663
MTSamples - Jury Score	4.533	**4.604**	4.426	4.383
Medec - MedecFlagAcc	0.591	**0.687**	0.628	0.521

Future:
Research Opportunities

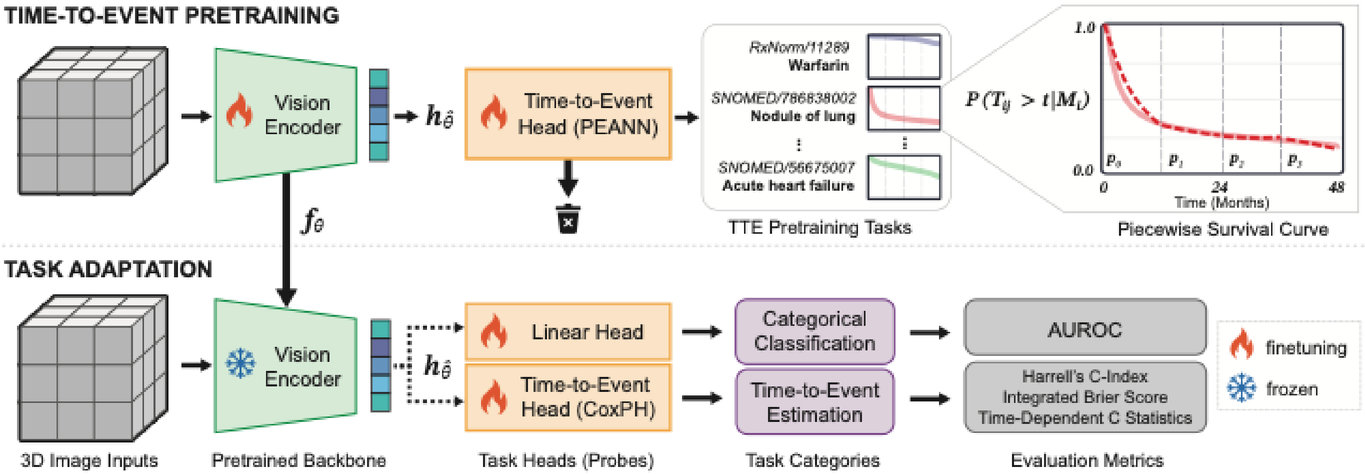
Multimodal Time-to-Event Pretraining

Pulmonary Embolisms



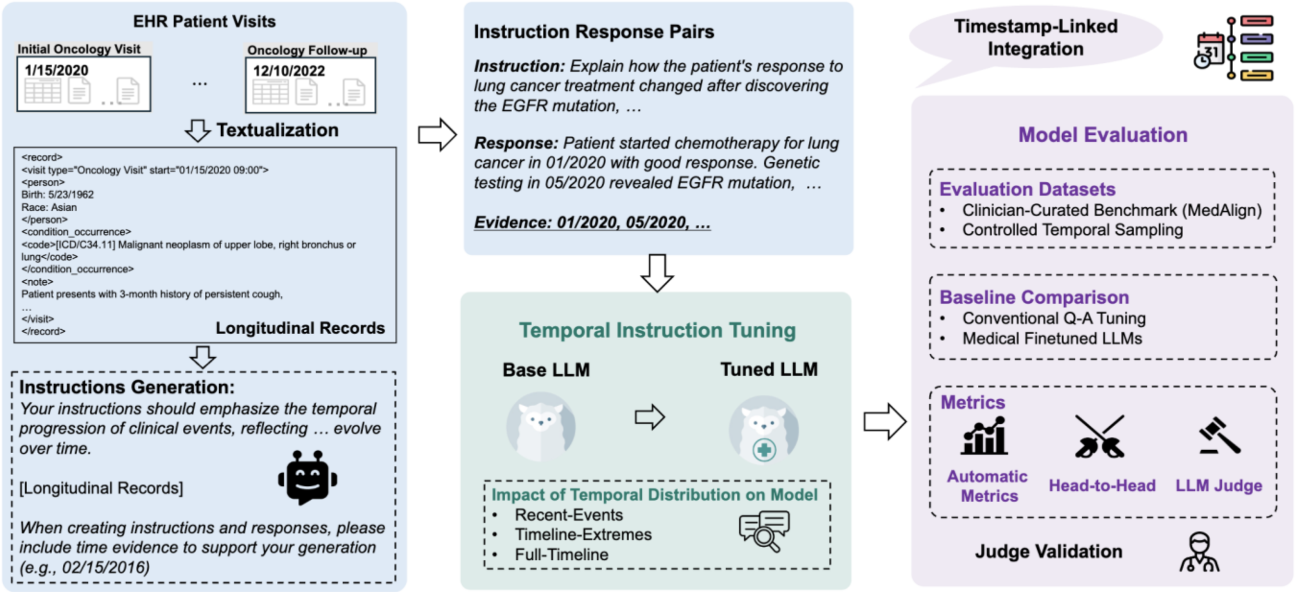
18,945 CT Scans
(4.2 Million 2D images)

- Same pretraining setup as MOTOR
- **Single time point** (not dynamic)
- Pretraining a 3D image encoder



Time-to-Event Pretraining for 3D Medical Imaging
Huo et al. ICLR 2025.

Synthetic Data Generation

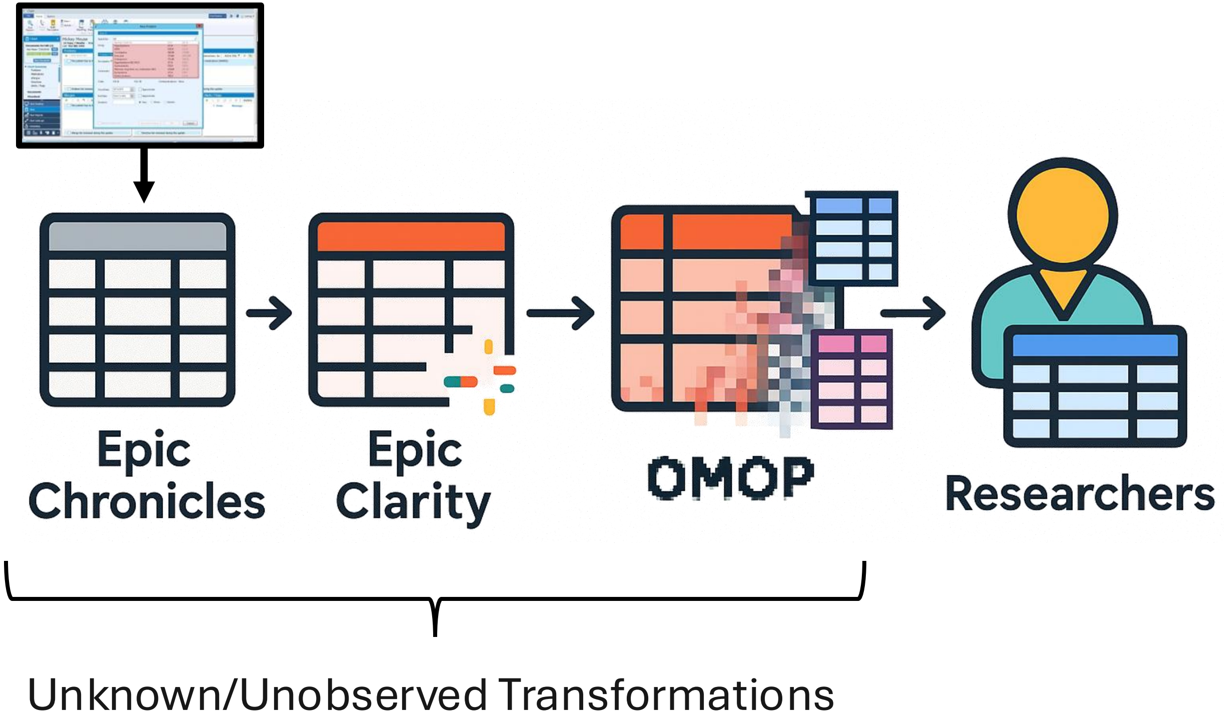


Use Real EHRs to Generate **Synthetic Post-Training Data**

TIMER: Temporal Instruction Modeling and Evaluation for Longitudinal Clinical Records

Cui et al. 2025. Preprint

Data-Centric AI: Data Quality

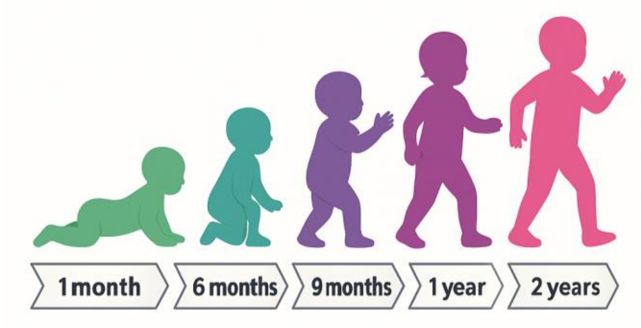


Researcher View

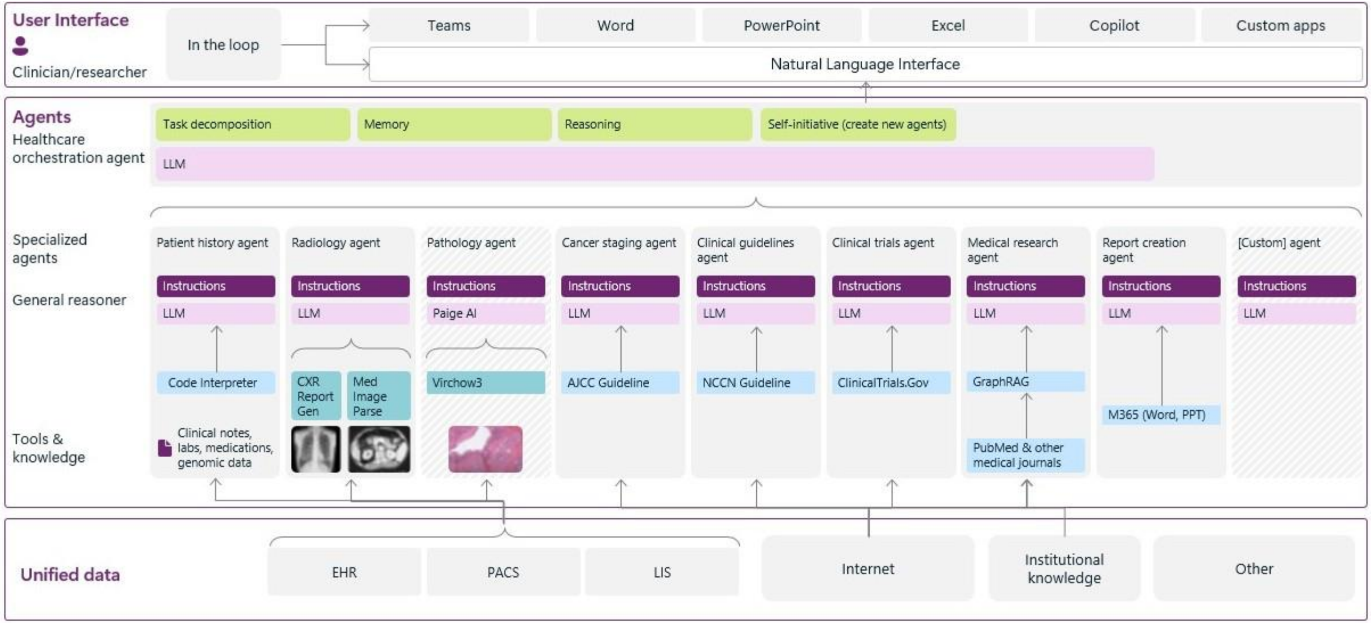
EHR data is typically transformed in **hidden ways**

Data-Centric AI: Training Mixtures

- **Exclusion biases in training data**
- General **data scarcity** (e.g., rare diseases)
- **Limited EHR datasets and benchmarks** for pediatric populations
- Unique data processing challenges
  - Example: Child and mother combined in a single patient record
- **Limited patient history** vs. adults
- Rapid developmental changes



Human-AI Teaming & Agentic Systems



Collaboration with Microsoft + Agent Orchestrator Platform

Thank You!

jason-fries@stanford.edu

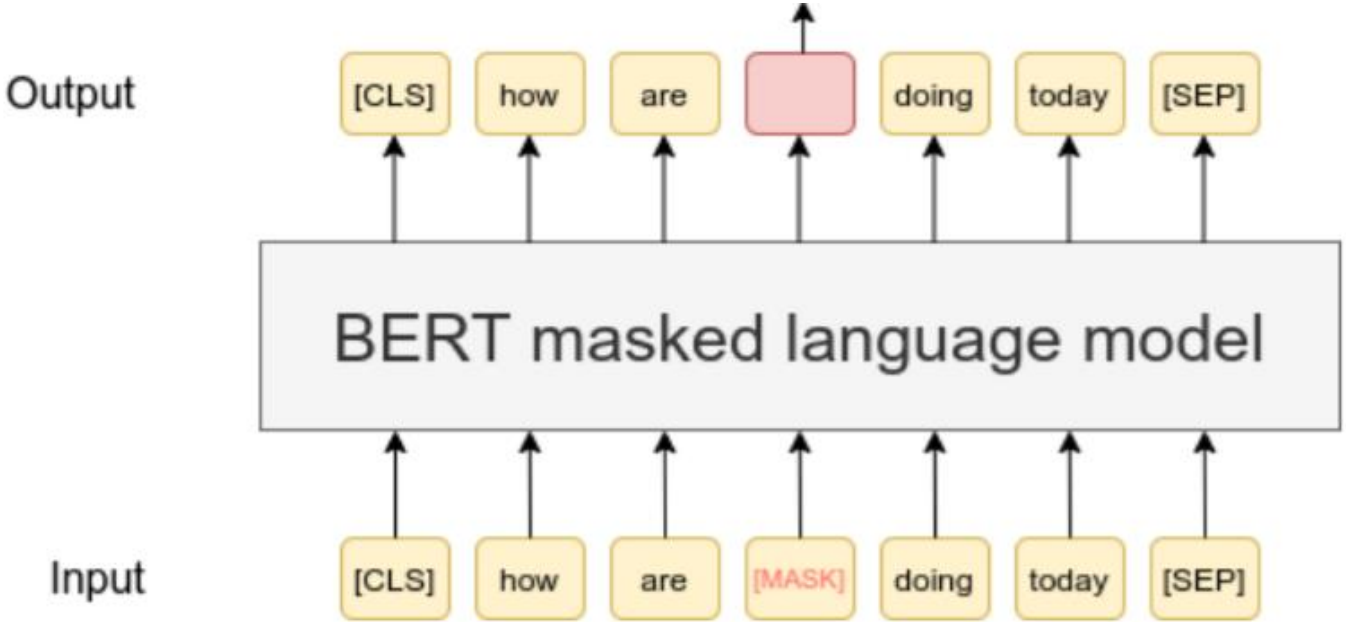
Appendix

BERT-Style (Masked Language Modeling)

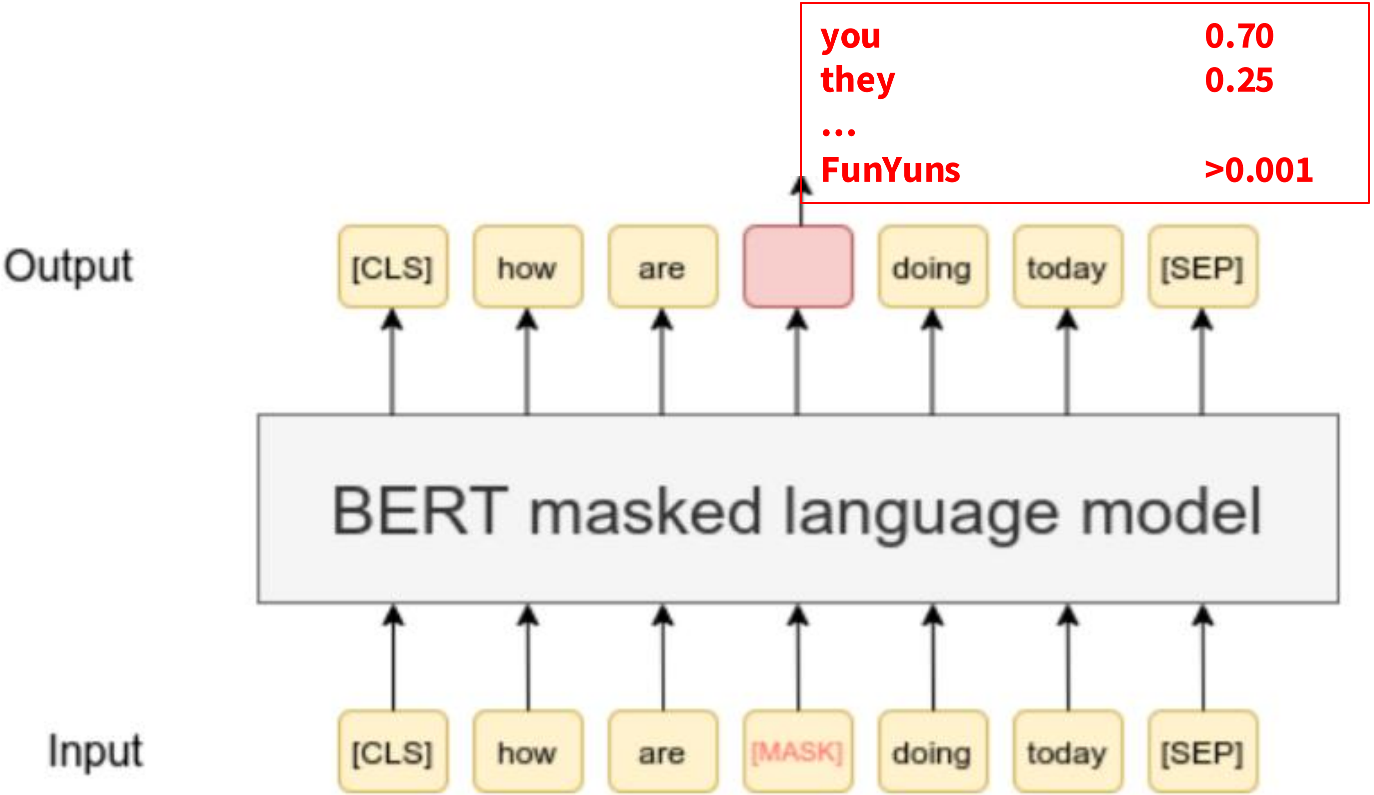
BEHRT (Li et al. 2020)
MedBERT (Rasmy et al. 2021)
ClaimPT (Zeng et al. 2022)

Corruption-based (Masking) Pretraining Objective

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



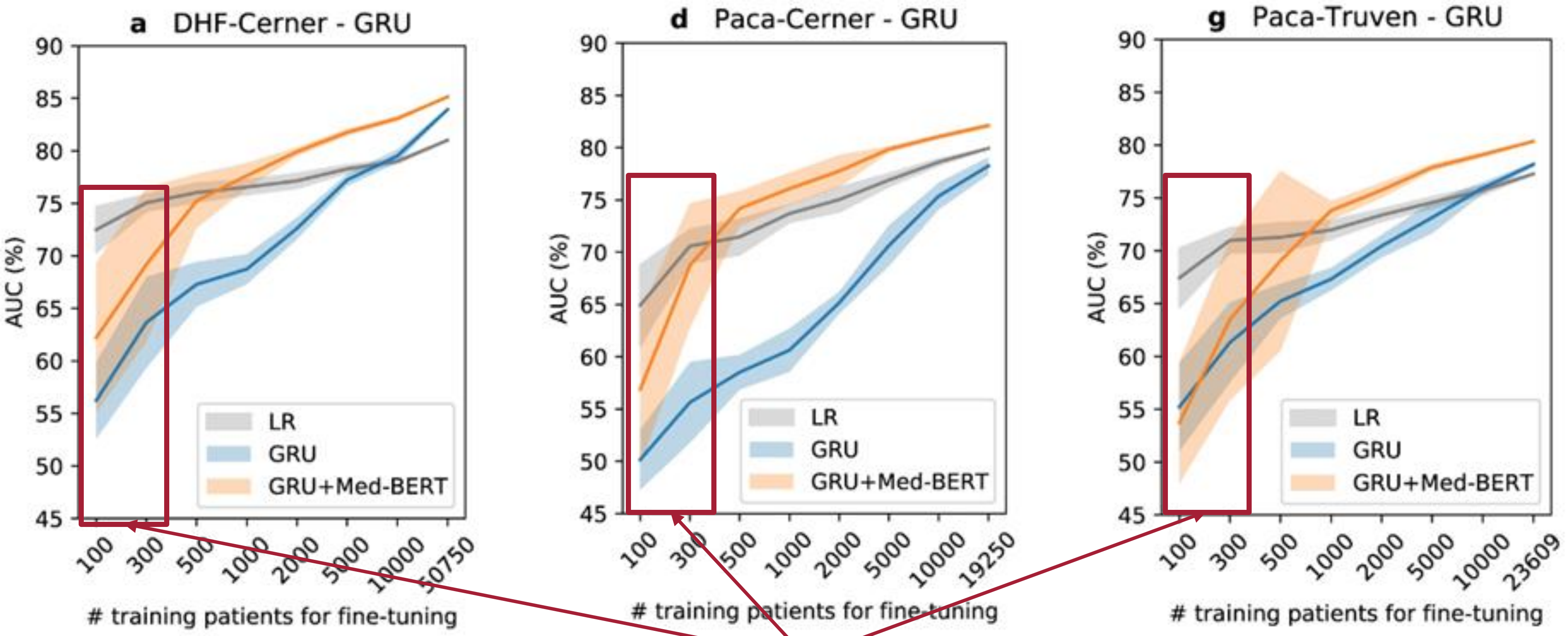
Corruption-based (Masking) Pretraining Objective



BERT-based Architecture (BEHRT)

Li et al. 2020

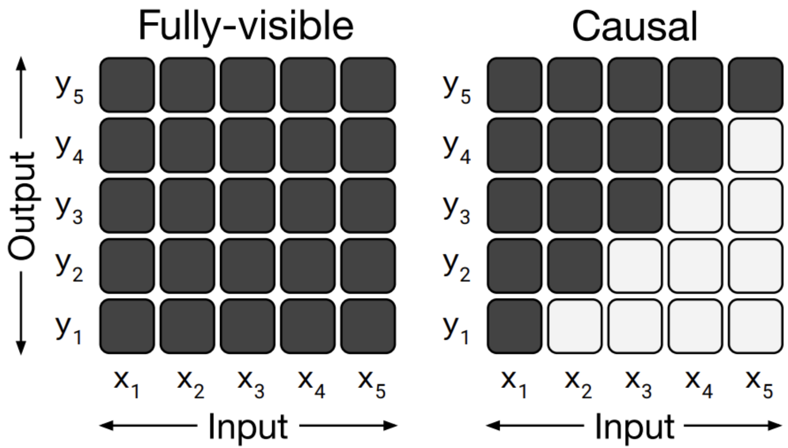
Better performance than baselines (MedBERT)



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**

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.