The Missing Context Problem in Foundation Models for Healthcare

Jason Fries, PhD jfries@stanford.edu

Center for Biomedical Informatics Research, Stanford University

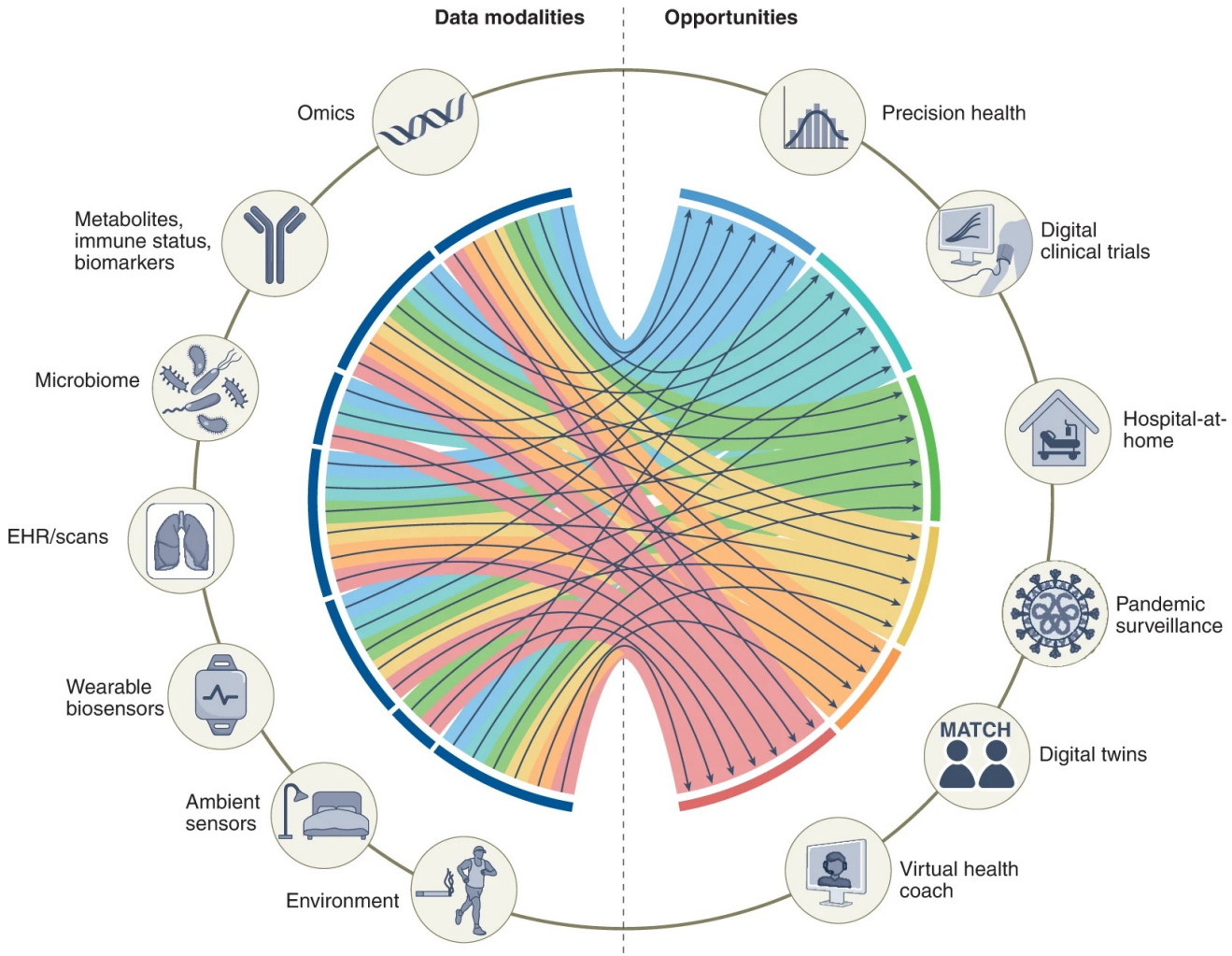
Stanford MEDICINE | Clinical Excellence Research Center

BMIR Stanford Center for Biomedical Informatics Research

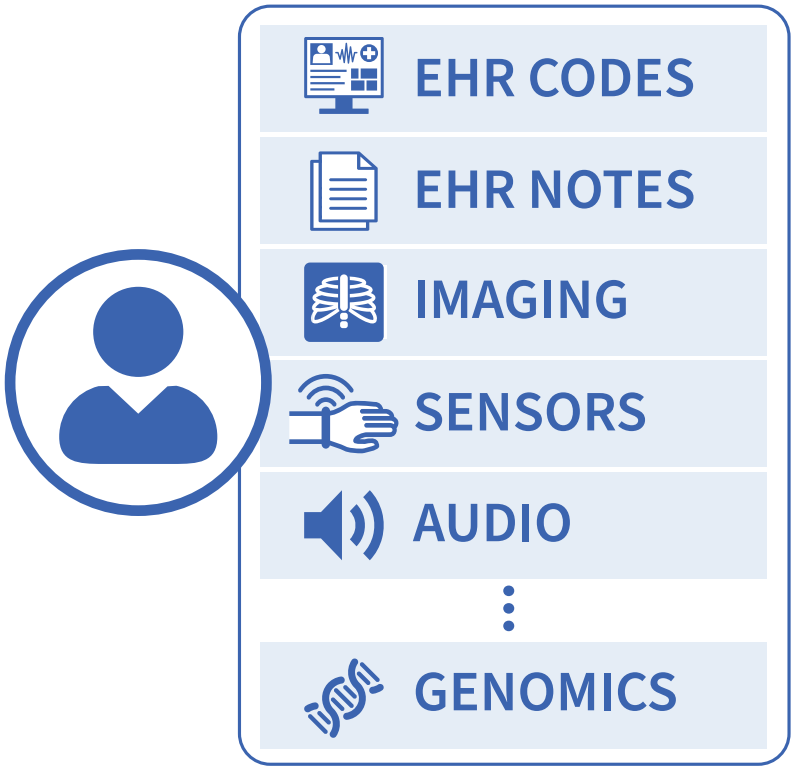
Stanford AIMI Center for Artificial Intelligence in Medicine and Imaging

HAI Stanford University Human-Centered Artificial Intelligence

Human Health and Healthcare is Fundamentally Multimodal

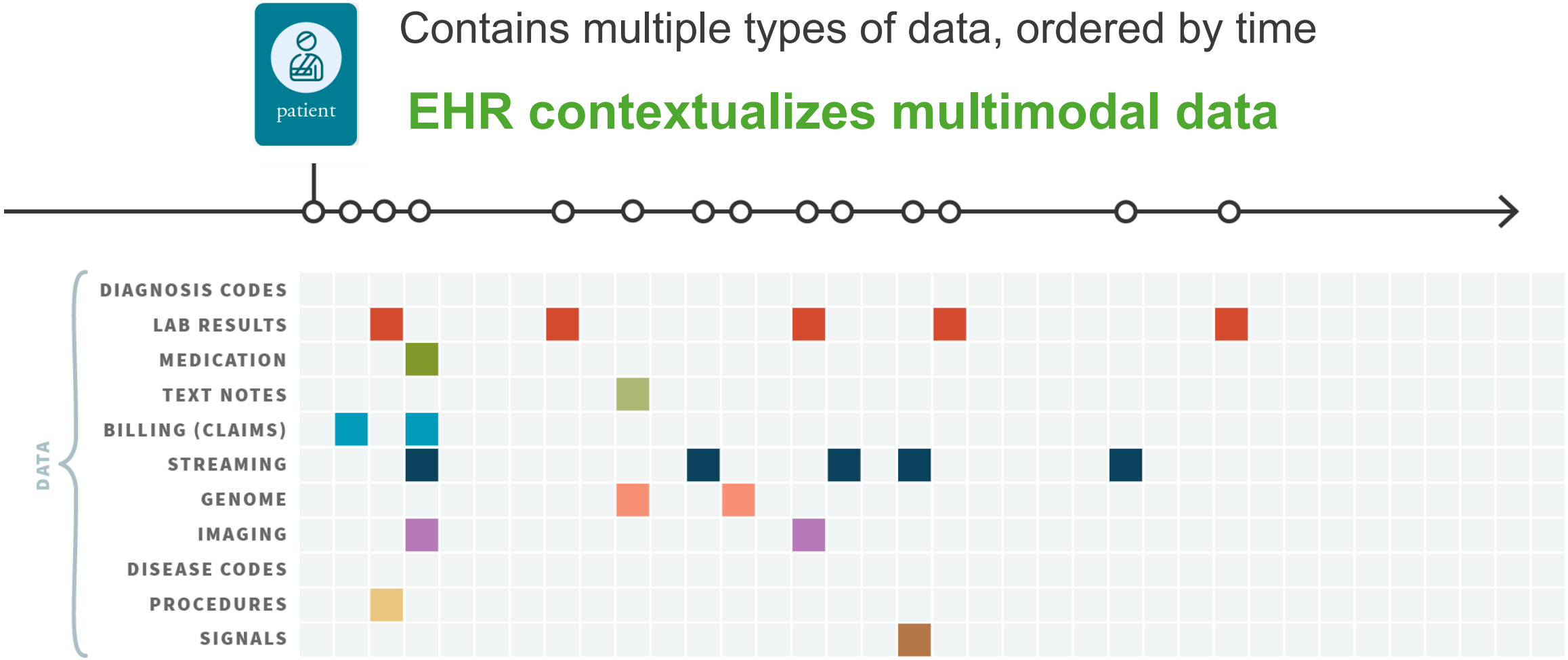


Acosta et al. *Nature Medicine*. 2022



Patient Data

Electronic Health Records (EHRs) are Multimodal



Eras of Pretraining Supervision for Unstructured Data

Manual Labeling

2012

Weakly Supervised Learning

2016

Self-Supervised Learning

2018

LLM-Accelerated Data Curation

2020

Reasoning / Agents Synthetic Data

Today

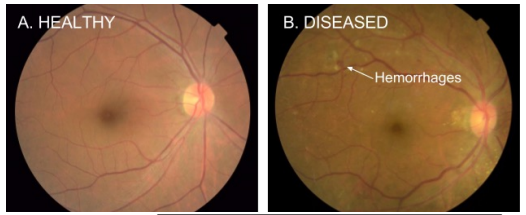
Medical Imaging Datasets

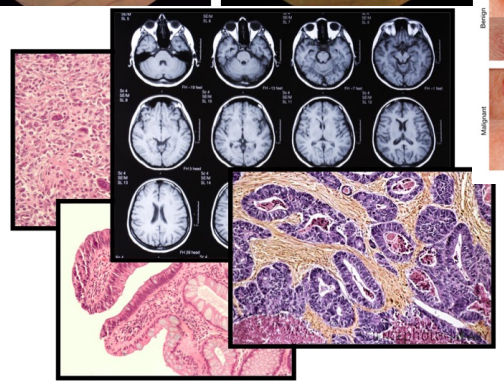
2017 - ChestX-ray14: **112k** x-rays

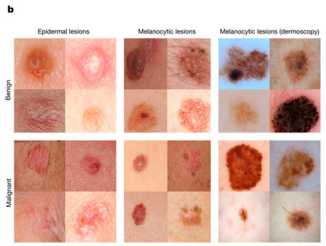
..

2023 - PMC-OA: **1.65M** image / caption pairs

2024 - MedTrinity-25M: **25M** images / annotations







(Gulshan et al. 2016)
(Esteva et al. 2017)

Multimodal Healthcare Foundation Models

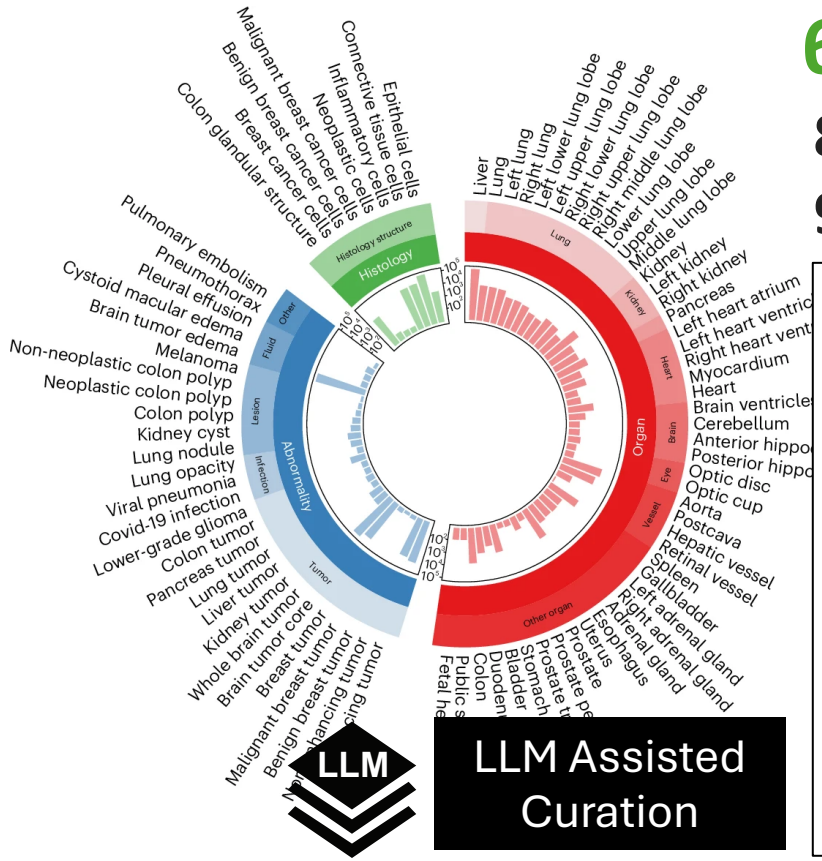
Image

Text Description

Segmentation Mask

...

Aligned Data For Self-Supervised Learning

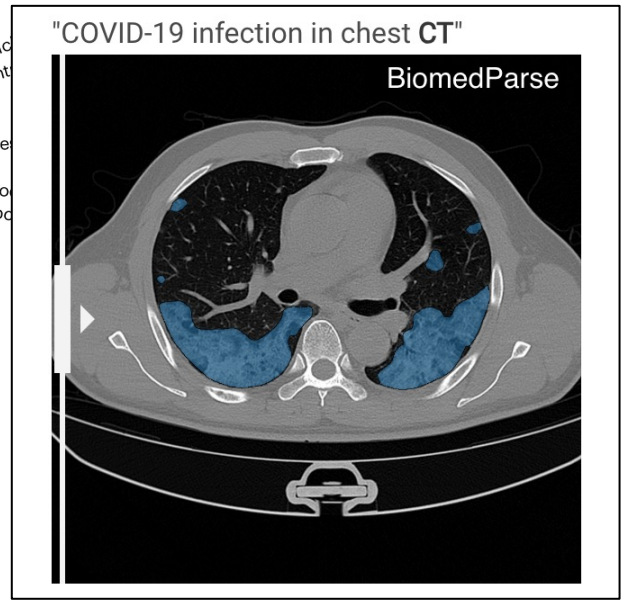


LLM Assisted Curation

6.8M image/mask/description

82 major object types

9 imaging modalities



BiomedParse (Zhao et al. 2024) *Nature Methods*

Healthcare is **Multimodal** and **Longitudinal**

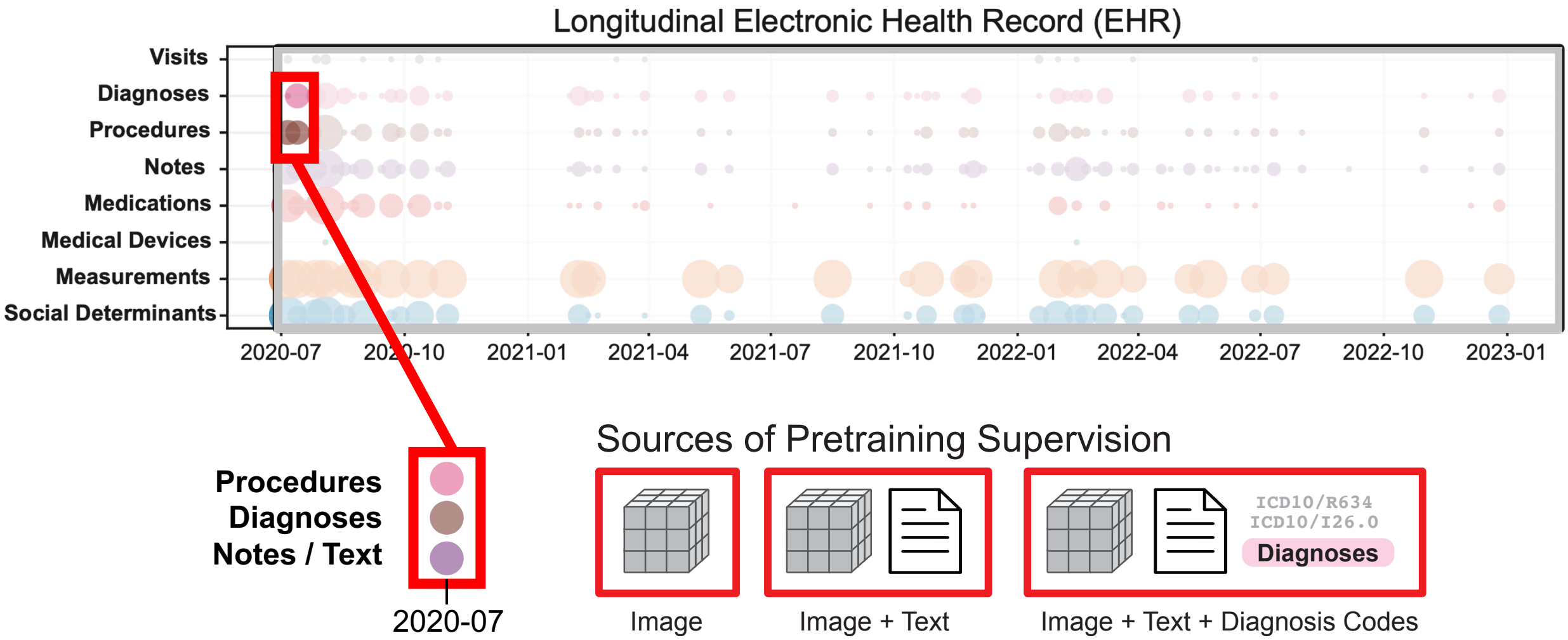
When was this patient diagnosed with cancer?

Is this patient likely to develop gastrointestinal cancer in 20 years, 10 years, or 3 years?

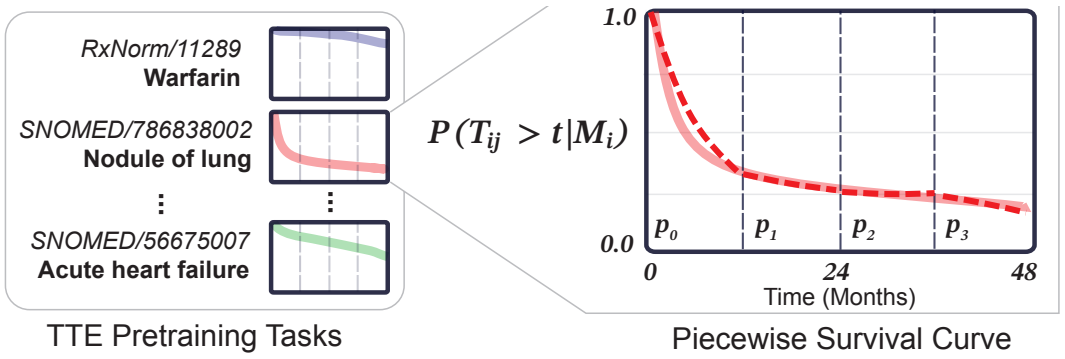
What imaging biomarkers are predictive of developing pulmonary hypertension?

These types of questions require reasoning skills and understanding of longitudinal progression of health

The Missing Context Problem in Healthcare AI



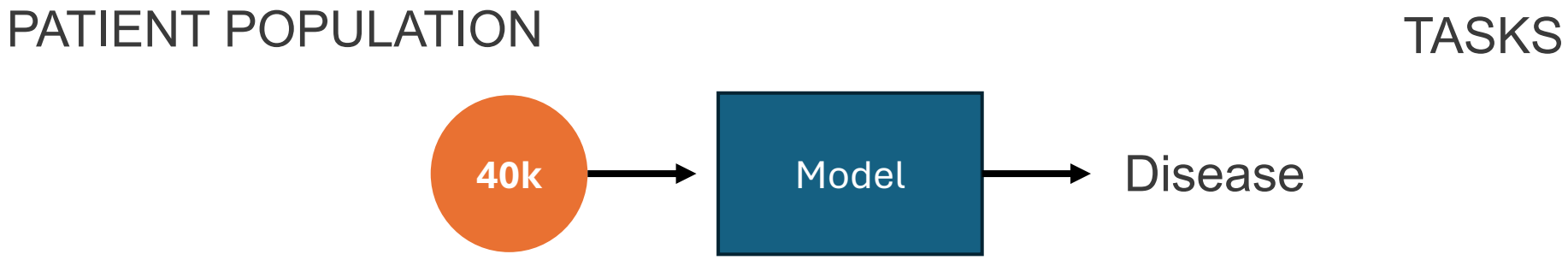
Today's Talk: Learning from Longitudinal EHR Data



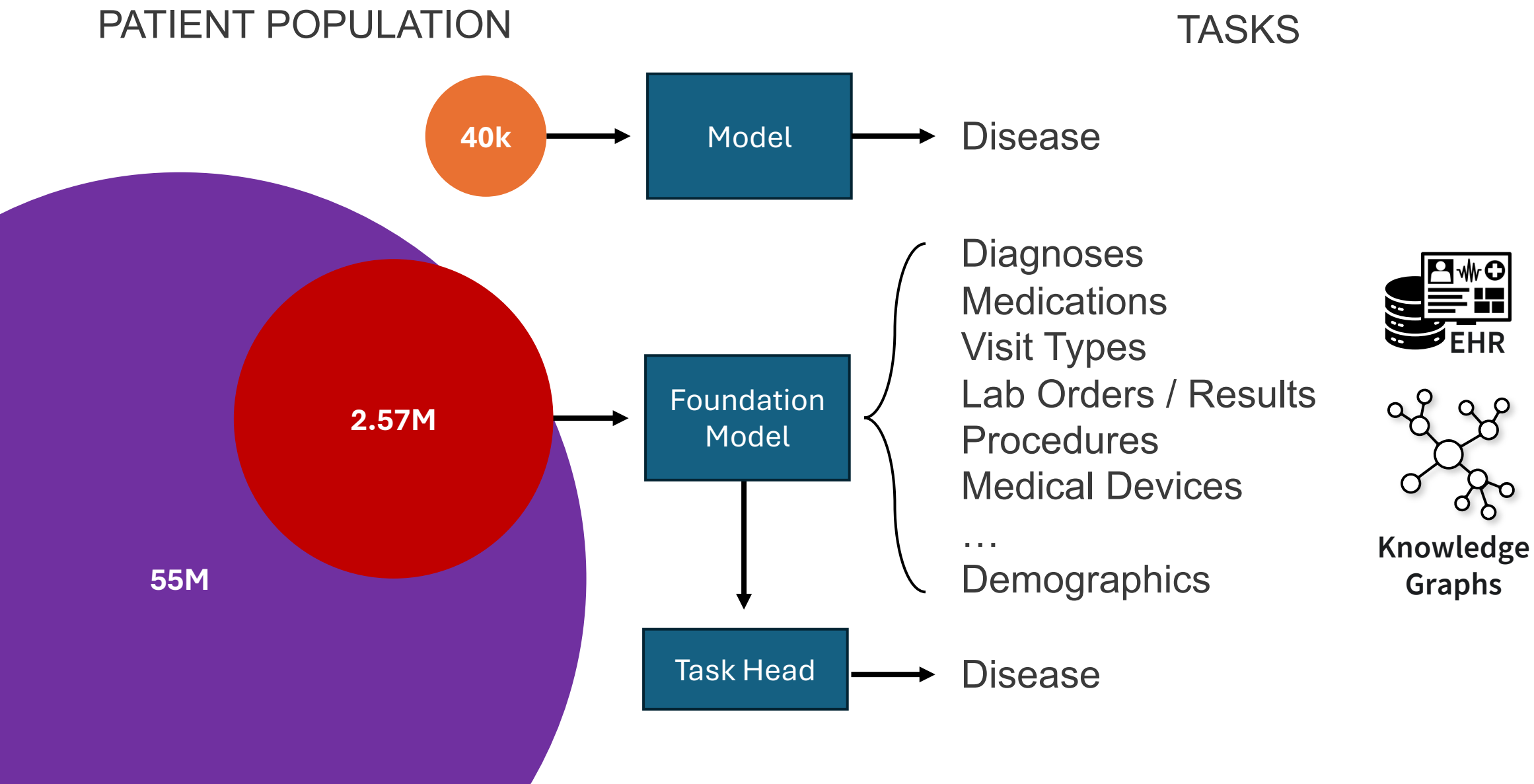
Time-to-Event Pretraining for 3D Medical Imaging

Longitudinal EHR Dataset & Model Releases

Autoregressive EHR Foundation Models



Autoregressive EHR Foundation Models



Benefits of Autoregressive EHR Foundation Models

Improved Few-Shot Learning & Label Efficiency

- **+3.5 to 19%** increase in AUROC [1]
- Match SOTA w/ **95%+ less training data** [4,5]

Robustness to Distribution Shifts

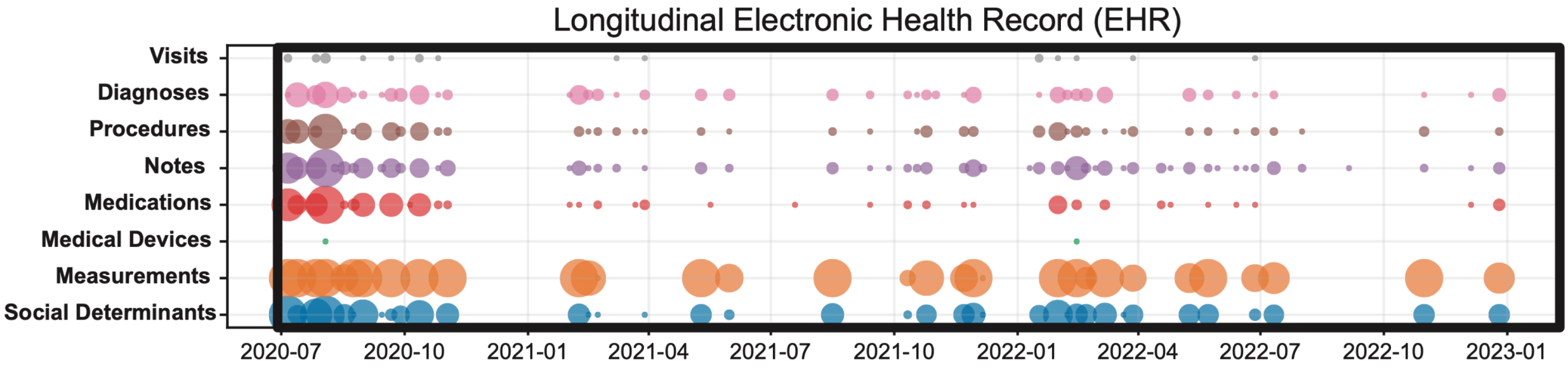
- Improved **temporal robustness** [2]
- Improved **subgroup performance** [3]

Cross-Site Adaptability

- Transfer pretrained models across hospitals
- Require **60-90% less** pretraining data [4]

[1] Steinberg et al. “Language models are an effective representation learning technique for electronic health record data”. *JBI*. 2021.
[2] Guo et al. “EHR foundation models improve robustness in the presence of temporal distribution shift”. *Scientific Reports*. 2023.
[3] Lemmon et al. “Self-supervised machine learning using adult inpatient data produces effective models for pediatric clinical prediction tasks.” *JAMIA*. 2023.
[4] Guo et al. “A Multi-Site Study on the Adaptability of a Shared EHR Foundation Model.” *npj Digital Medicine*. 2024.
[5] Steinberg et al. “MOTOR: A Time-To-Event Foundation Model For Structured Medical Records.” *ICLR 2024*

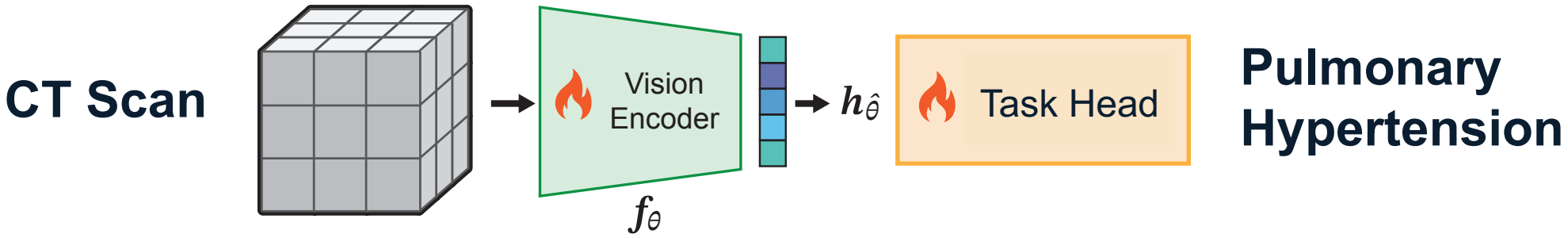
Time-to-Event Supervision (Steinberg, Fries, et al. 2024)



Key Intuition: Predict not only **if something will occur** but **when**

Autogenerate 1000s of time-to-event pretraining tasks

Health Prognosis using 3D Medical Imaging



Sources of Pretraining Supervision

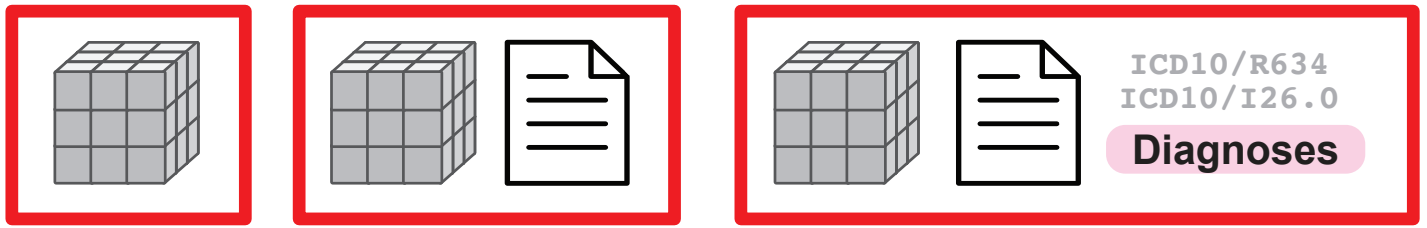
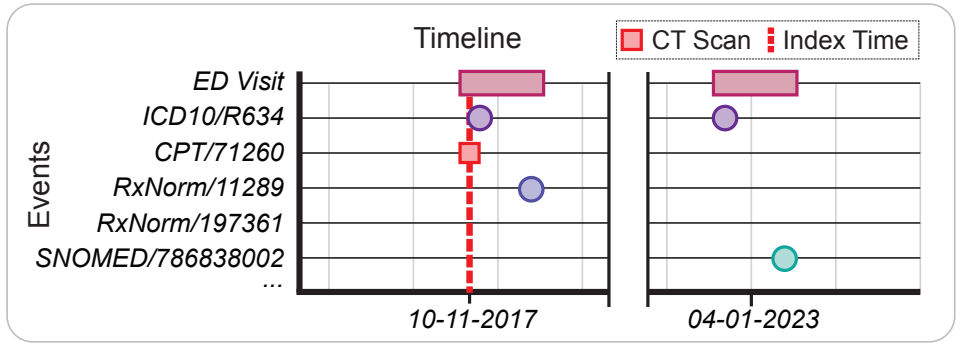
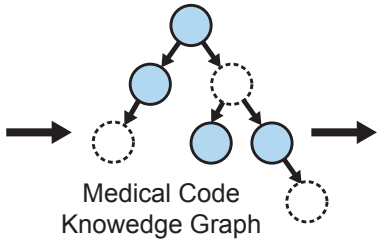


Image | Image + Text | Image + Text + Diagnosis Codes

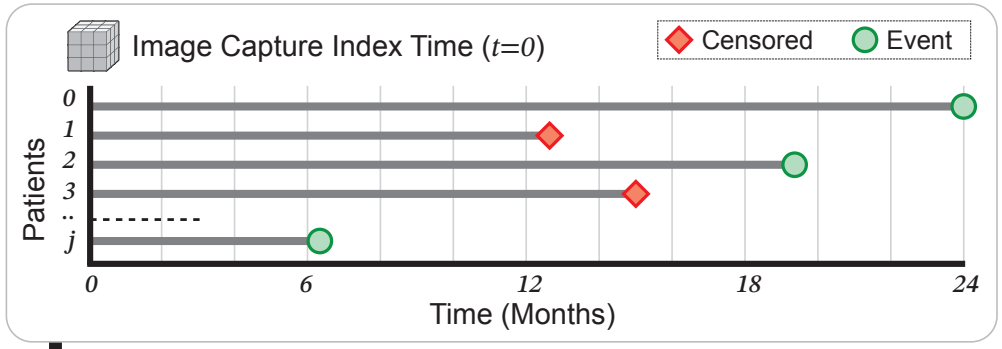
Time-to-Event Pretraining for 3D Imaging



Patient Electronic Health Record (EHR)

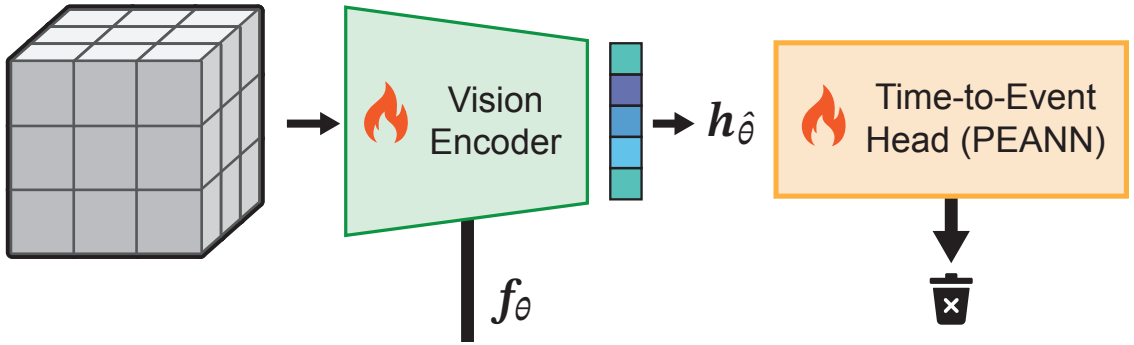


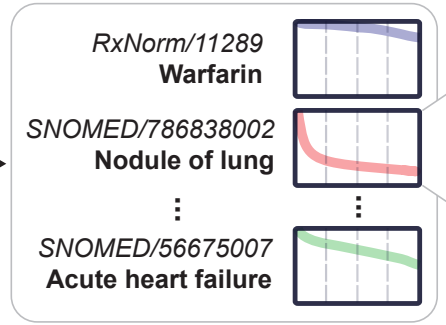
Entropy-Ranked Vertex Cover for Task Selection



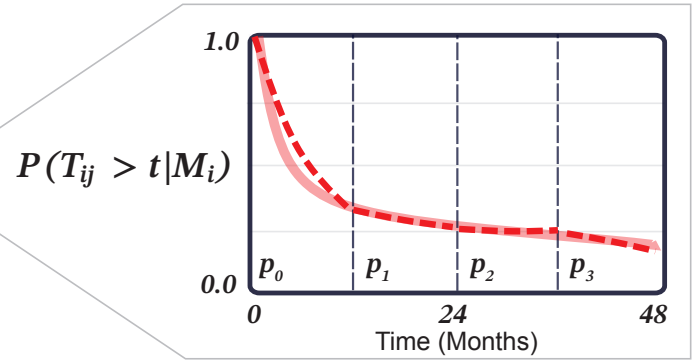
Create Time-to-Event Pretraining Tasks (n=8,192)

TIME-TO-EVENT PRETRAINING





TTE Pretraining Tasks



Piecewise Survival Curve

Training Set:
18,945 CT Scans

- Continued Pretraining of 3 Self-Supervised Vision Encoders
- 8,192 Time-to-Event Pretraining Tasks

Time-to-Event Supervision: Performance Gains

How does TTE supervision compare to popular self-supervised methods (i.e., MAE)?

23.7% increase in AUROC
29.4% improvement in Harrell's C-index
54% improvement in calibration (Integrated Brier Score)

How does local context supervision compare vs. longitudinal?

15.7% increase in AUROC from longitudinal vs. local supervision

Current self-supervised methods struggle to learn prognostic pixel biomarkers

2024 Multimodal, Longitudinal EHR Dataset Releases

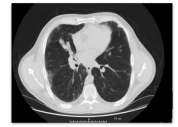
EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models

6,739 Patients

Tabular

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

19,402 Patients



CT Scans | Tabular | Radiology Notes

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

267 Patients

Tabular | All Clinical Notes

26k Unique Patients
270M Coded Clinical Events
70k Clinical Notes, **128** Note Types
23k CT Scans (**2.4M** 2D images)

https://redivis.com/ShahLab/

REDIVIS

EHR Foundation Model Releases

Hugging Face https://huggingface.co/StanfordShahLab

MOTOR: Many Outcome Time Oriented Representations **2024**

CLMBR: Clinical language modeling-based representations **2023**

Context Clues: **Evaluating Long Context Models for Clinical Prediction Tasks on EHRs**
(Wornow, Bedi, et al. 2024)

Context Lengths: 512, 1024, 2048, 4096, 8192, 16384
Architectures: GPT, Llama, Hyena, Mamba
Pretraining Dataset: **2.57M** EHRs

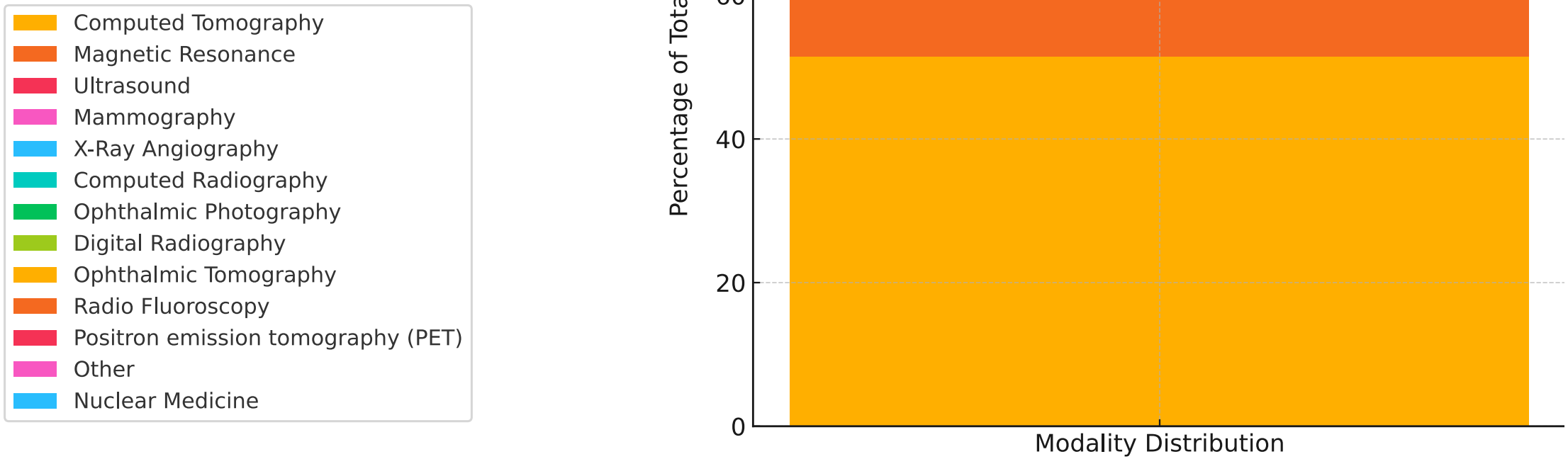
Future: One Healthcare System's Data

30 TB EHR

1B Conversations

6 PB Imaging

(+1PB/year pathology)



Future: Building Multi-Site EHR Foundation Models

- Leverage Existing Communities to Train **Site-Specific EHR Foundation Models**
- Build as **Mixture of Experts**

https://www.ohdsi.org/data-standardization/

Team Science

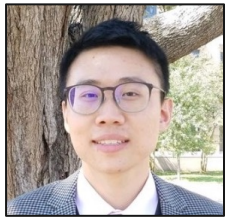
Shah Lab and Collaborating Researchers

Jason Fries

Ethan Steinberg

Michael Wornow

Suhana Bedi

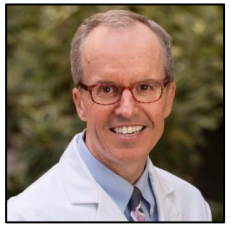

Frazier Huo

Alejandro Lozano

Jeya Maria Jose Valanarasu

Louis Blankemeier

Akshay Chaudhari


Curtis Langlotz

Nigam Shah

Governance, Privacy, and Licensing

Mariko Kelly

Julie Marie Romero

Jonathan Gortat

Scott Edmiston

Reed Sprague

Technology & Digital Solutions

Natasha Flowers

Joseph Mesterhazy

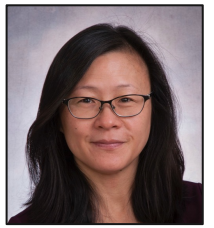
Priya Desai

Somalee Datta

Todd Ferris

External Collaborating Researchers

Lawrence Guo


Lillian Sung