Innovating at the Intersection of Clinical and Data-Centric AI

Jason Alan Fries, PhD

Department of Biomedical Data Science

Division of Computational Medicine
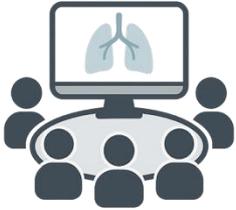
Department of Medicine

Stanford University

1

A patient with advanced lung cancer has been treated for 24 months and now shows progression in a few isolated areas, with mixed response.

When presenting to a **tumor board**, clinicians need to piece together:

- **Difficult to reason longitudinally**
- **Difficult to find similar cases**

- New genomic mutations driving treatment resistance
- Pathology reports
- Radiology scans
- Clinical history

*"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

Electronic Health Records (EHRs) are Multimodal Timelines

PATIENT — Many diverse **data types** that **evolve** over time
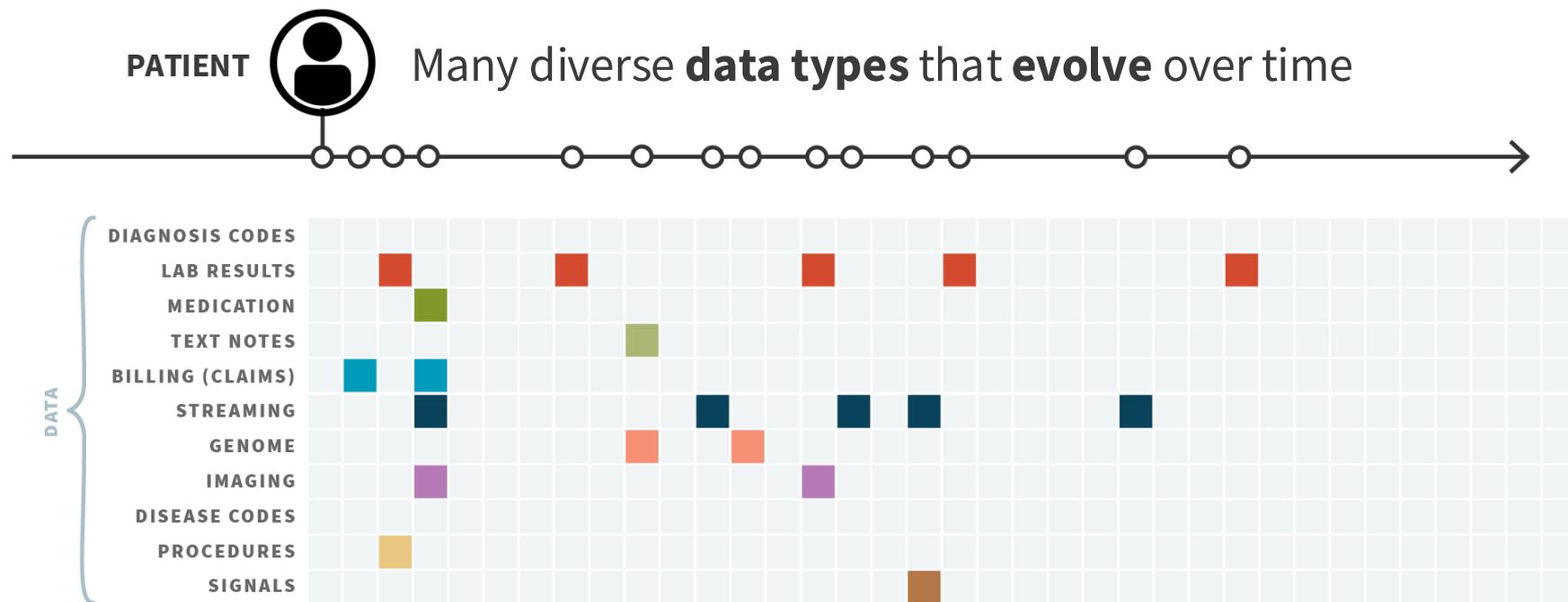
Longitudinal EHRs provide a **holistic view of multimodal data**

Healthcare AI has a Missing Context Problem

MEDICAL HISTORY   FUTURE OUTCOMES
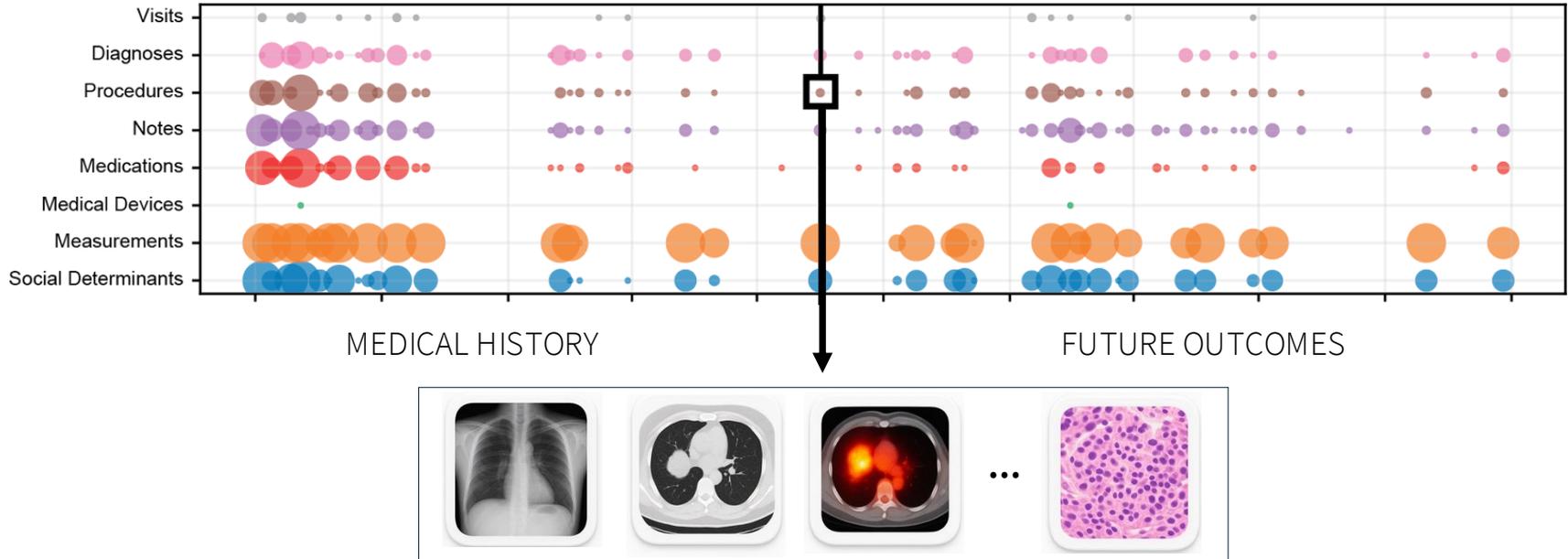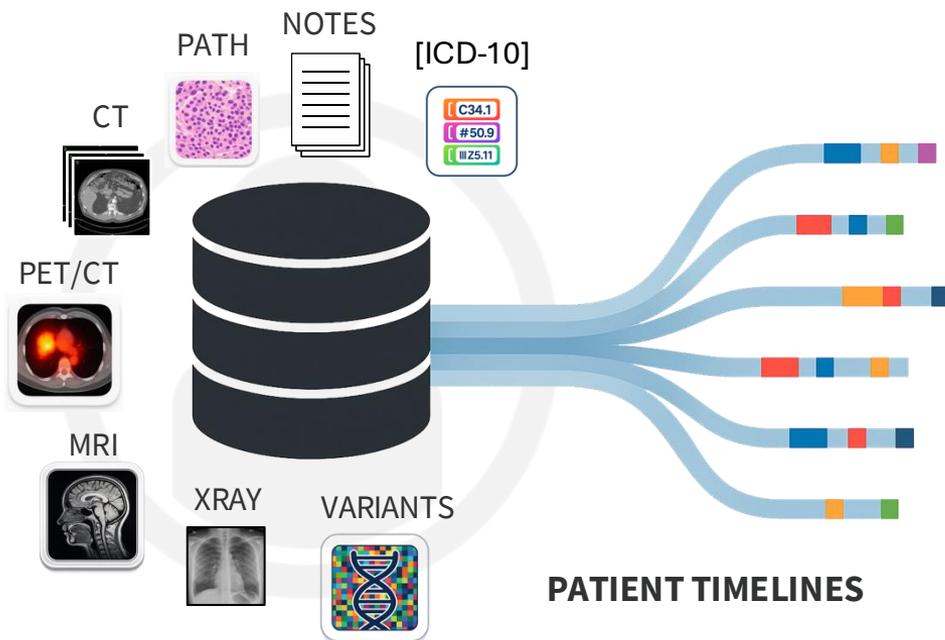
The next generation of healthcare AI will **reason over multimodal patient data across months and years**

The Future: Multimodal, Interactive Healthcare AI

Multimodal timelines as a unifying framework for retrieval, interaction, and feedback.

PATH
NOTES
[ICD-10]
CT
PET/CT
MRI
XRAY
VARIANTS

PATIENT TIMELINES
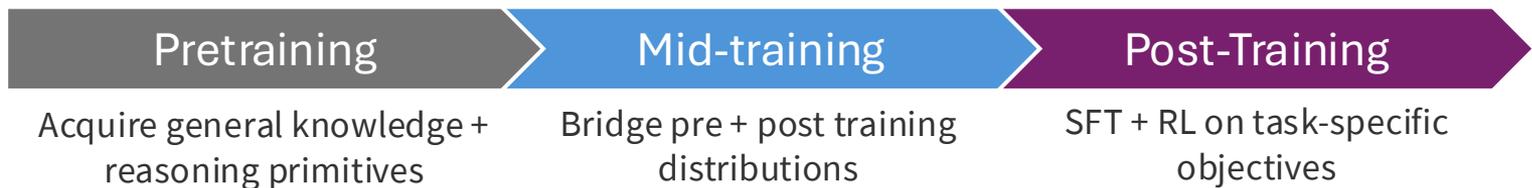
Chat interfaces for EHR

Use similar patients to inform decision making

Feedback loops to capture error signals and **improve data and models**

Data-Centric AI

Training data defines an LLM's capabilities, limitations, and biases

Pretraining	Mid-training	Post-Training
Acquire general knowledge + reasoning primitives	Bridge pre + post training distributions	SFT + RL on task-specific objectives

Published as a conference paper at COLM 2025

LIMO: Less is More for Reasoning

Yixin Ye[1,2]*, Zhen Huang[2,3]*, Yang Xiao[4], Ethan Chern[1,2], Shijie Xia[1,2], Pengfei Liu[1,2]†
[1]Shanghai Jiao Tong University [2]SII-GAIR
[3]Fudan University [4]The Hong Kong Polytechnic University

Math Reasoning

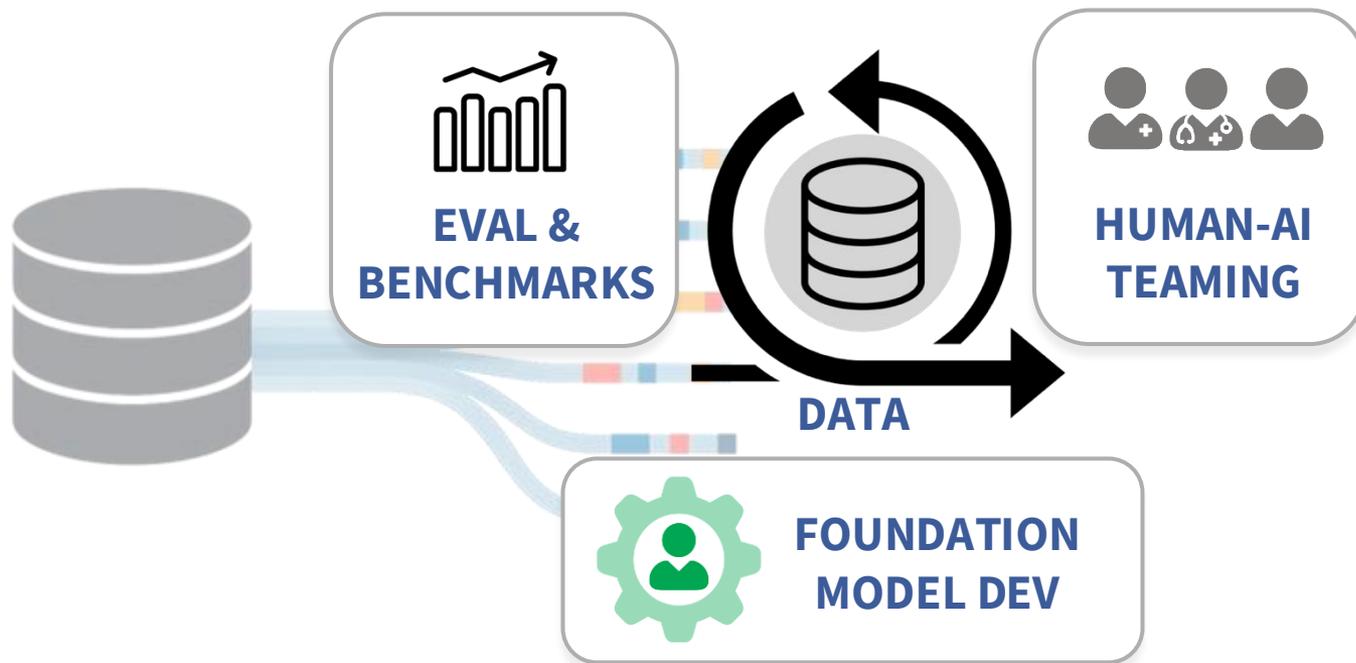**800** carefully selected training examples
1% of training data

AIME24
6.5% → **63.3%**

MATH500
59.2% → **95.6%**

Healthcare AI Requires Data-Centric Systems

Quality > Quantity!

EVAL & BENCHMARKS

DATA

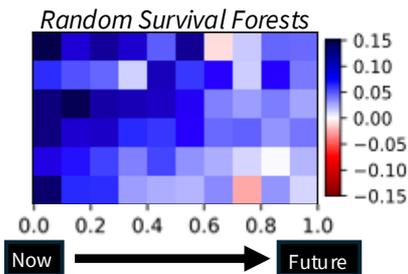HUMAN-AI TEAMING

FOUNDATION MODEL DEV

Foundation Models: Longitudinal EHRs

Time-to-Event Pre/Mid-training

Use EHR data to produce **massive-scale survival supervision** based on patient outcomes. **SOTA on many survival tasks.**
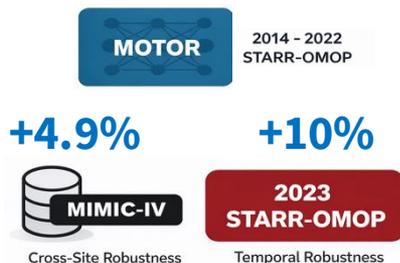
Objective	Celiac	HA	Lupus	NAFLD	Cancer	Stroke
RSF	0.729	0.836	0.787	0.802	0.824	0.840
Next Code	0.774	0.862	0.842	0.860	0.860	0.857
Time-to-Event	**0.802**	**0.887**	**0.863**	**0.864**	**0.865**	**0.875**

Longer Horizons

Random Survival Forests

Now ⟶ Future

Improved Robustness

MOTOR   2014 – 2022 STARR-OMOP

+4.9%   **+10%**

MIMIC-IV   2023 STARR-OMOP

Cross-Site Robustness   Temporal Robustness
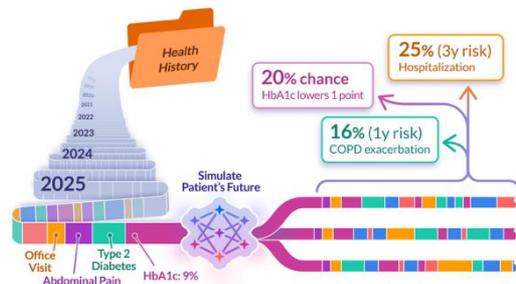
Steinberg et al. 2024, Huo et al. 2025

Zero-Shot EHR Generation

Waxler et al. 2025

Creates **synthetic future timelines** based on patient history…

nature medicine

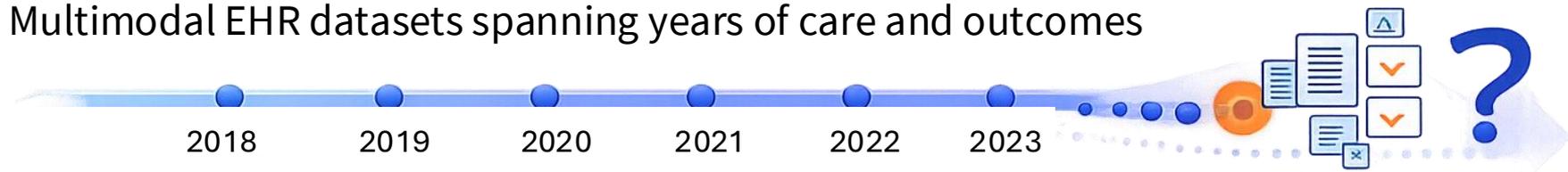Comment | Published: 07 January 2026

How to interpret 'zero-shot' results from generative EHR models

Suhana Bedi ✉, Jason Alan Fries & Nigam H. Shah

Benchmarking: Patient Trajectories

Multimodal EHR datasets spanning years of care and outcomes

2018 2019 2020 2021 2022 2023

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.*	▦		📝

Wornow et al. 2023, Huang et al. 2023, Fleming et al. 2024

26k patients **295M** events **442k** visits

REDIVIS
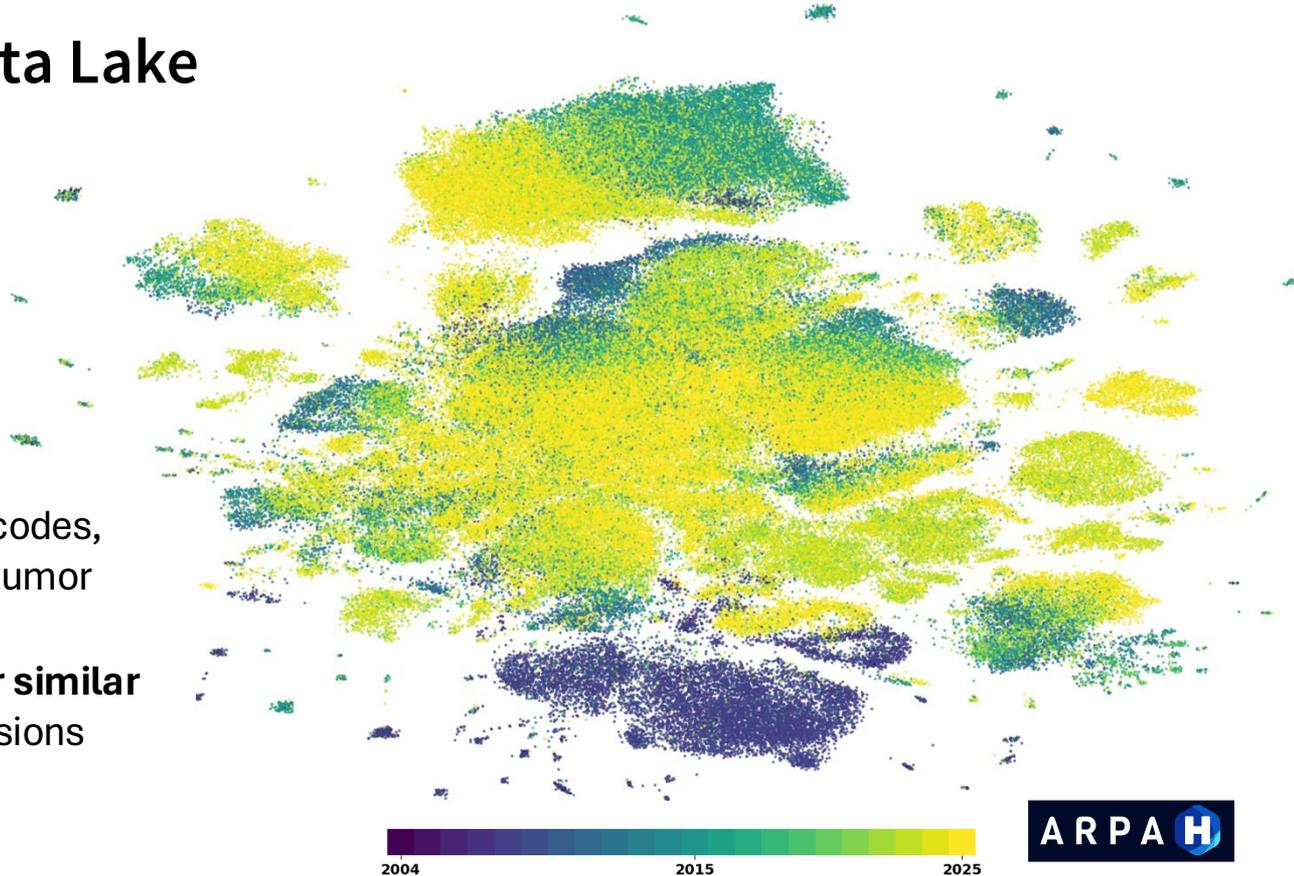
https://redivis.com/ShahLab

Human-AI Teaming: AI-Accelerated Tumor Boards

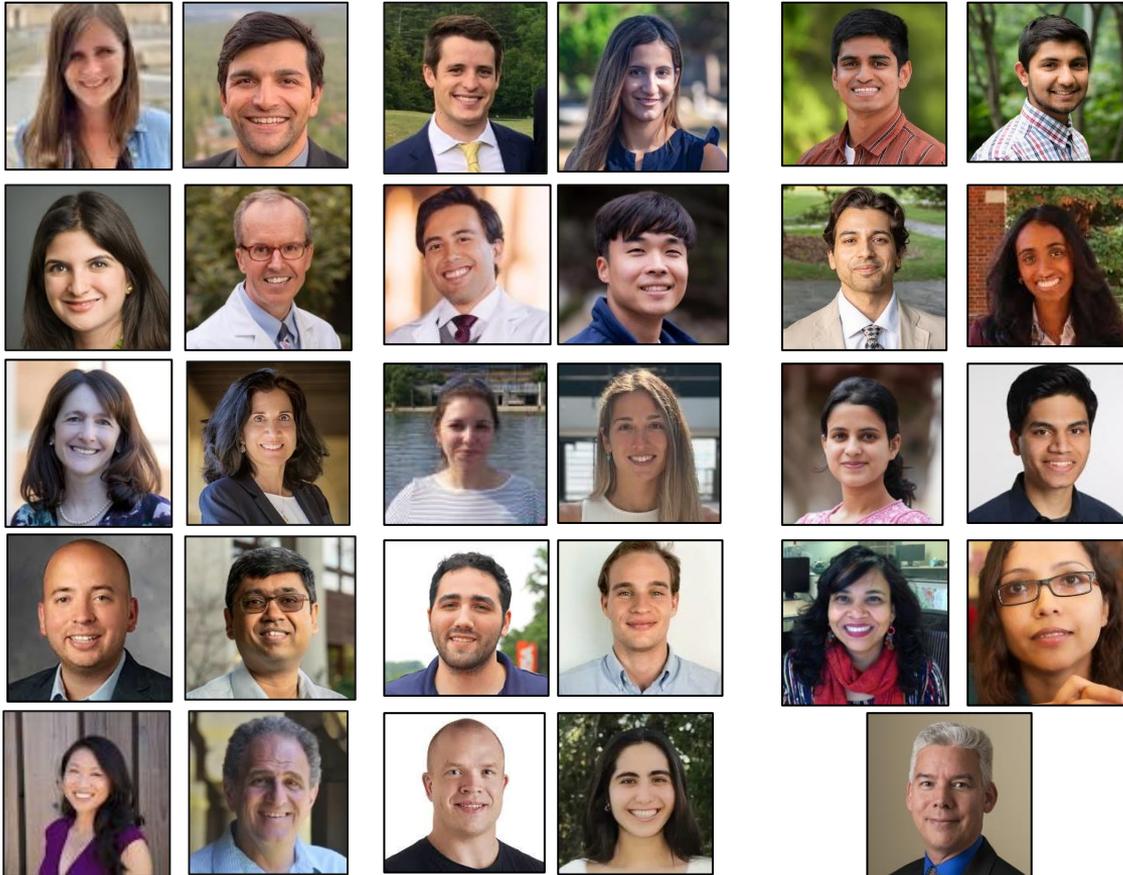## VISTA Oncology Data Lake
211k patients

- **Multimodal cancer data** (codes, notes, imaging, pathology, tumor genetic assays)
- Finding and **reasoning over similar patients** to guide care decisions

Stanford MEDICINE | Department of Biomedical Data Science

ARPA H

2004   2015   2025

Team Science

Thank You!
jfries@stanford.edu

VISTA - AI Tumor Board
Weill Cancer Hub West
Technology & Digital Solutions