TrajOnco: A multi-agent framework for temporal reasoning over longitudinal EHR for multi-cancer early detection

Zero-shot LLM-based framework delivers accurate, interpretable multi-cancer risk prediction from raw longitudinal EHR without model training

Sihang Zeng[1,2,3], Youngwon Kim[1], Wilson Lau[1], Ehsan Alipour[1], Ruth Etzioni[2,3], Meliha Yetisgen[2], Anand Oka[1], Jay Nanduri[1]
[1]Truveta Inc.; [2]University of Washington; [3]Fred Hutch Cancer Center zengsh@uw.edu

Take-away message

TrajOnco, a chain-of-agents framework with long-term memory, achieved strong zero-shot performance across 15 cancers, comparable to trained ML methods, while providing temporally coherent clinical reasoning without any model training or feature engineering.

Background & objective

Background

Early identification of individuals at higher risk for cancer can improve outcomes and help target screening resources. Existing machine learning approaches require substantial data preprocessing, manual feature engineering, and task-specific model training, and their interpretability is often limited.

Objective

We evaluated whether a large language model–based chain-of-agents framework can estimate 1-year cancer risk directly from raw longitudinal EHR data across multiple cancer types.

Study population

Lung cancer (unmatched)

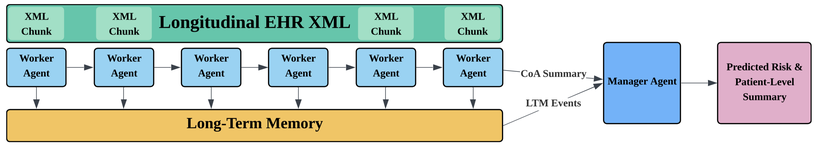
- From 120 million patients from Truveta Data
- 500 lung cancer cases
- 125,000 controls (randomly sampled, approximate cumulative incidence)
- 1-year risk prediction

Multi-cancer evaluation (matched)

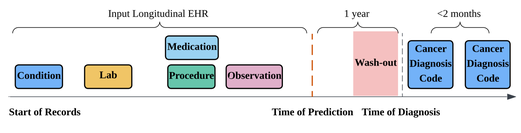
- 15 cancer types
- Each cancer type: 500 cases, 500 matched controls (1:1 matching on sex and 10-year age group)
- 1-year risk prediction

TrajOnco framework

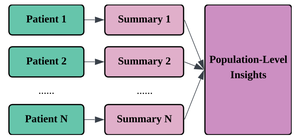
Architecture overview



Study design



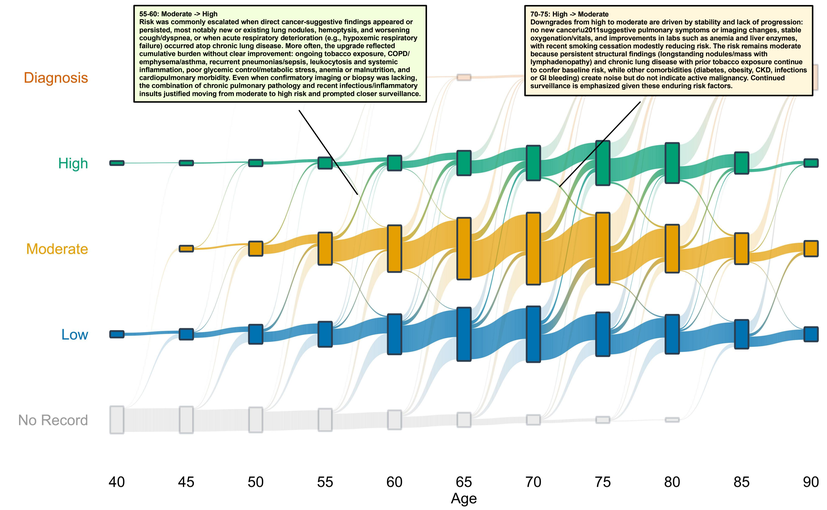
Insights generation



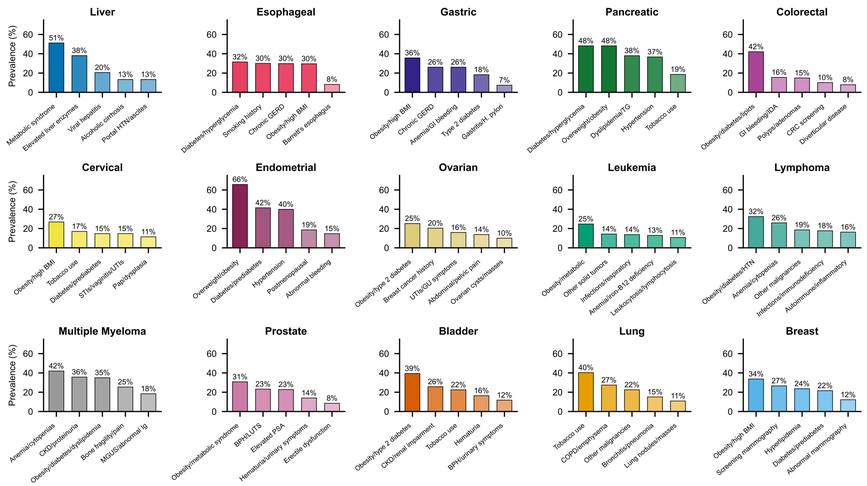
- Zero-shot chain-of-agents (CoA) processes raw longitudinal EHR in XML format with sequential worker agents and a long-term memory
- Manager agent synthesizes patient-level summary, risk score (1-10), and evidence-linked rationale
- Interpretability can be achieved in patient-level and population-level

Interpretation

Risk transition summarization

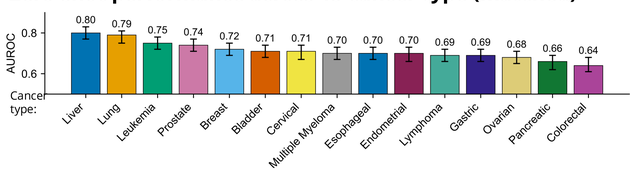


Top themes for risk estimation

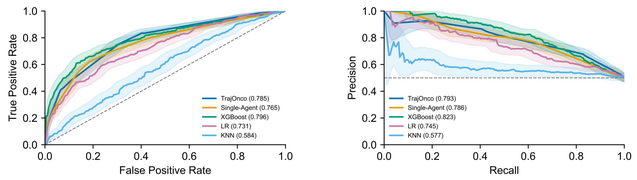


Results

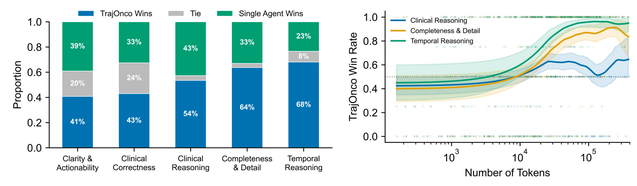
Zero-shot performance across 15 cancer type (matched)



Comparison with trained machine learning methods (matched lung cancer cohort)



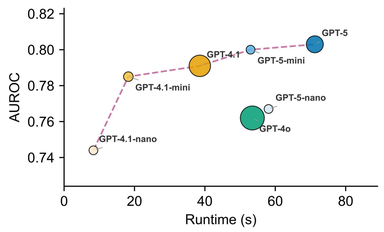
LLM-as-a-judge evaluation (matched lung cancer cohort)



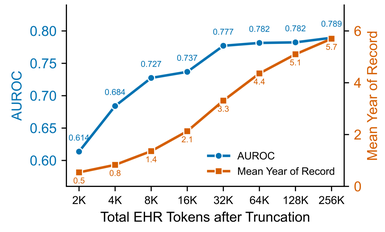
Performance on unmatched lung cancer cohort (unmatched)

	AUROC	AUPRC	NPV	PPV	Sensitivity	Specificity
Logistic Regression	0.903 (0.892, 0.912)	0.055 (0.046, 0.070)	**1.000 (0.999, 1.000)**	0.014 (0.013, 0.016)	**0.907 (0.880, 0.929)**	0.737 (0.734, 0.739)
XGBoost	**0.925 (0.916, 0.934)**	**0.195 (0.162, 0.233)**	0.999 (0.999, 1.000)	0.017 (0.015, 0.019)	0.885 (0.856, 0.912)	0.785 (0.782, 0.787)
KNN	0.790 (0.772, 0.805)	0.014 (0.012, 0.017)	0.999 (0.998, 0.999)	0.010 (0.009, 0.011)	0.794 (0.754, 0.830)	0.657 (0.654, 0.659)
Single-agent (GPT-4.1-mini)	0.866 (0.849, 0.883)	0.068 (0.051, 0.089)	0.999 (0.999, 0.999)	**0.018 (0.012, 0.020)**	0.741 (0.710, 0.859)	**0.836 (0.747, 0.839)**
TrajOnco (GPT-4.1-mini)	**0.871 (0.855, 0.885)**	**0.071 (0.053, 0.092)**	**0.999 (0.999, 0.999)**	**0.017 (0.013, 0.019)**	**0.772 (0.742, 0.861)**	**0.825 (0.741, 0.827)**

Base model selection (matched lung cancer cohort)



Effect of trajectory length (matched lung cancer cohort)



References

- Zhang et al., Chain of Agents: Large Language Models Collaborating on Long-Context Tasks. NeurIPS 2024
- Zeng et al., Traj-CoA: Patient Trajectory Modeling via Chain-of-Agents for Lung Cancer Risk Prediction. NeurIPS 2025 GenAI4Health Workshop
- Zeng et al., TrajOnco: a multi-agent framework for temporal reasoning over longitudinal EHR for multi-cancer early detection. arXiv: 2604.10386